Patch-Aware Sample Selection for Efficient Masked Image Modeling

Zhengyang Zhuge[1, 2], **Jiaxing Wang**[3], **Yong Li**[3],
Yongjun Bao[3], **Peisong Wang**[1, 2, 4], **Jian Cheng**[1, 2, 4*]
[1] Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] JD.com
[4] AiRiA
zhugezhengyang2020@ia.ac.cn, {wangjiaxing41, liyong5, baoyongjun}@jd.com
{peisong.wang, jcheng}@nlpr.ia.ac.cn

Abstract

Nowadays sample selection is drawing increasing attention. By extracting and training only on the most informative subset, sample selection can effectively reduce the training cost. Although sample selection is effective in conventional supervised learning, applying it to Masked Image Modeling (MIM) still poses challenges due to the gap between sample-level selection and patch-level pre-training. In this paper, we inspect the sample selection in MIM pre-training and find the basic selection suffers from performance degradation. We attribute this degradation primarily to 2 factors: the random mask strategy and the simple averaging function. We then propose Patch-Aware Sample Selection (PASS), including a low-cost Dynamic Trained Mask Predictor (DTMP) and Weighted Selection Score (WSS). DTMP consistently masks the informative patches in samples, ensuring a relatively accurate representation of selection score. WSS enhances the selection score using patch-level disparity. Extensive experiments show the effectiveness of PASS in selecting the most informative subset and accelerating pretraining. PASS exhibits superior performance across various datasets, MIM methods, and downstream tasks. Particularly, PASS improves MAE by 0.7% on ImageNet-1K while utilizing only 37% data budget and achieves ∼1.7× speedup.

Introduction

Self-supervised pre-training has recently gained great attention due to its label-free nature and ability to learn informative representations. Among them, Masked Image Modeling (MIM), motivated by the remarkable achievements of Masked Language Modeling (MLM) (Devlin et al. 2018; Brown et al. 2020) in Natural Language Process (NLP) and the progress of Vision Transformers (ViTs) (Dosovitskiy et al. 2020; Touvron et al. 2021; Liu et al. 2021), has emerged as a propitious pre-training paradigm for computer vision (CV). However, MIM suffers from large computational burden due to its extensive model size and substantial data demand. A natural way of mitigating is to eliminate the redundant data and train only on the most informative ones, namely, sample selection.

Pioneer works (Paul, Ganguli, and Dziugaite 2021; Coleman et al. 2020; Feldman 2020; Mirzasoleiman, Bilmes, and Leskovec 2020; Killamsetty et al. 2021) have explored efficient training with sample selection in Supervised Learning (SL). However, there still lack successful practice in applying sample selection to accelerate the MIM pre-training process. We hypothesize this is due to the discrepancy between MIM pre-training process and the supervised learning. In conventional supervised training, a sample image is fed into the model as a whole, namely, supervised learning operates at the sample level. While in MIM pre-training, samples are first partitioned into patches and masked image modeling operates at the patch level. Taking MAE (He et al. 2022) for example, the image patches are randomly masked and the model reconstructs the masked patches according to the visible ones. This discrepancy raises a significant question: *How can we effectively evaluate the importance of a sample given the image patches?*

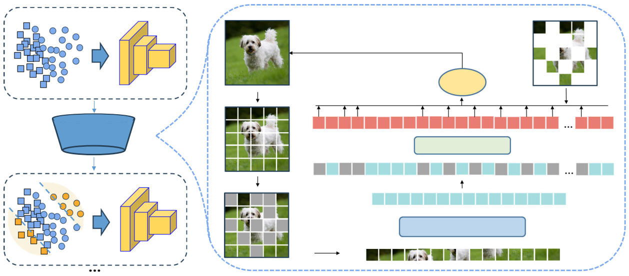

Figure 1: Illustration of the Basic Sample Selection for Masked Image Modeling. It is noteworthy that we divide the original MIM pre-training into two stages: pre-training stage and sample selection stage.

The most natural and straightforward way is to directly aggregate the reconstruction loss of the randomly masked patches. Samples with larger aggregated loss are considered more informative and thus picked. We name this strategy Basic Sample Selection (BSS) as denoted in Figure 1. Meanwhile, following (Killamsetty et al. 2021), we apply a dynamic selection scheme, where the sample selection is conducted periodically during the training process. However, we find this BSS method incurs inevitable performance degradation. we suspect the performance degradation stemming from two aspects: 1. the inconsistency and inaccuracy of the selection score due to random masking strategy. and 2.

[*] Corresponding author.

the ignorance of inter-patch disparity during patch loss aggregation. To address these issues, we propose Patch-Aware Sample Selection (PASS) for MIM, which includes the Dynamic Trained Mask Predictor (DTMP) and the Weighted Selection Score (WSS). We assume that Patches with larger reconstruction loss possess greater informativeness and are more representative for sample importance. Therefore, for consistent and accurate selection score, we propose a mask predictor to consistently identify informative patches and mask them. Furthermore, We empirically find that the predictor can give relatively accurate estimation even after a short training. So we update the predictor just for a few epochs before each sample selection stage. In this way, the predictor can dynamically keep up with the model pre-training process, while avoiding excessive additional training cost. As the predictor is trained to estimate the patch loss and generate the mask, we name this dynamically trained mask predictor (DTMP). For the disparity of patches during patch loss aggregation, we recognize that simple average operation leads to score homogeneity, that is, numerous samples may exhibit similar selection scores, which can undermine the effectiveness of sample importance ranking. To further distinguish different patches with the sample, we propose a weighted sample selection strategy (WSS), which weights patches according to their corresponding predicted loss during loss aggregation. Finally, with this PASS, we are able to evaluate the informativeness of samples in MIM pre-training.

PASS presents a generic sample selection method, which enables seamless integration with various MIM methods such as MAE and simMIM. PASS significantly accelerates the pre-training process, while simultaneously maintaining or even surpassing the performance of original MIM methods. Extensive experiments are conducted on various datasets (ImageNet-1K, MS-COCO, ADE20K, etc.) and tasks(classification, detection, segmentation) to show its effectiveness. Specifically, our approach outperforms the original MAE and simMIM across multiple datasets with only 37% sample budget and $\leq$59% time consumption.

The main contributions of this work are threefolds:

- We investigate sample selection in MIM pretraining, and identify two factors that contribute to the decline in performance when sample reduction is implemented.
- Then, We propose Patch-Aware Sample Selection (PASS), which utilizes a Dynamic Trained Mask Predictor and Weighted Selection Loss to maintain performance while accelerating MIM pre-training with limited samples.
- Extensive experiments and ablation studies demonstrate the effectiveness of PASS. PASS even outperforms the original MIM across diverse datasets.

Related Works

Sample Selection

The remarkable performance of Deep Neural Networks (Simonyan and Zisserman 2014; He et al. 2016; Dosovitskiy et al. 2020) primarily stems from their extensive training on massive data. Sample selection, which extracts and trains on the most informative subset to achieve efficient training on voluminous data samples, is an effective approach in traditional Supervised Learning (Loshchilov and Hutter 2015; Coleman et al. 2020; Feldman 2020). One of the fundamental components in sample selection is the definition of selection score. Various score formats are proposed to accurately reflect the importance of samples, including "forgetting events" (Toneva et al. 2018), CE-loss (Jiang et al. 2019), uncertainty sampling (Settles 2011; Citovsky et al. 2023) and gradient-based methods (Paul, Ganguli, and Dziugaite 2021; Mirzasoleiman, Bilmes, and Leskovec 2020). While the majority Sample Selection studies primarily concentrate on Supervised Learning, recently a few researchers shifted their attention towards sample selection in self-supervised learning. Among them, (Ju et al. 2022) extends the coreset selection method to self-supervised case with contrastive learning. (Sorscher et al. 2022) leverages k-means clustering to tell the easy/hard sample through a prototypicality metric. However, these methods are designed for the sample-level self-supervised learning (SSL) that operates on individual samples, rendering them unsuitable for the currently popular patch-level SSL paradigm.

Masked Image Modeling

Masked Image Modeling (MIM) (He et al. 2022; Xie et al. 2022a; Bao et al. 2021; Zhou et al. 2021; Wei et al. 2022), as a form of self-supervised pre-training (Chen et al. 2020; Chen and He 2021; He et al. 2020), has garnered significant attention from researchers. With the increasing popularity of vision transformer (Dosovitskiy et al. 2020; Liu et al. 2021; Wang et al. 2021; Liu et al. 2022), MIM pre-training demonstrates remarkable superiority over other SSL methods and emerges as a propitious paradigm for pre-training in CV. However, MIM pre-training suffers from a heavy training budget due to the extensive model size and massive dataset volume. In order to improve the efficiency of pre-training, some studies focus on reducing the model complexity (Huang et al. 2022; Li et al. 2022b; Guo et al. 2022; Wang et al. 2023), while others investigate the module design that facilitates rapid convergence of MIM (Li et al. 2022a; Zhang et al. 2022; Ren et al. 2023). However, few attention is paid to the training acceleration with limited data samples. A few works (El-Nouby et al. 2021; Tong et al. 2022; Xie et al. 2022b) have investigated the performance of MIM under a limited data. Nevertheless, none of them focuses on how to enhance the performance within a limited pre-training data budget.

Approach

In this section, we first outline the preliminary of Masked Image Modeling and Sample Selection. Then we adapt conventional sample selection methods to MIM pre-training in a straightforward way. However, the performance of this simple strategy suffers from a noticeable degradation. We show that it stems from the inconsistency and inaccuracy selection score related to random masking, and ignorance of inter-patch disparity in patch loss aggregation. Based on our findings, we propose patch-aware sample selection (PASS) whose framework is presented in Figure 2.

Preliminary

Here we introduce the concepts of Masked Image Modeling and Sample Selection.

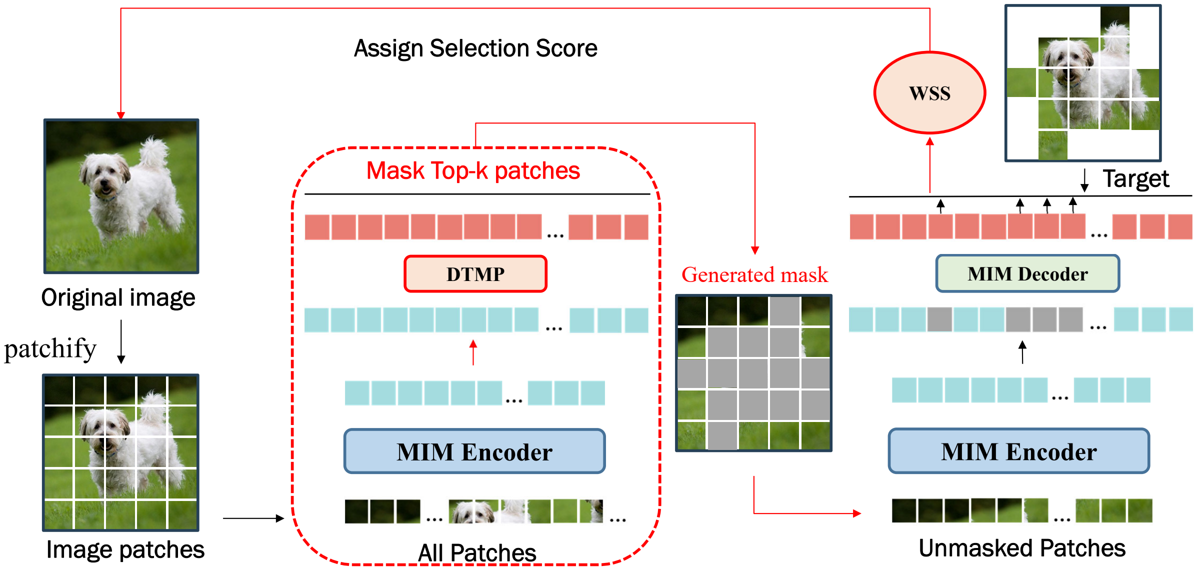

Figure 2: Illustration of the Patch-Aware Sample Selection (PASS) scheme for Sample Selection stage. The proposed Dynamic Trained Mask Predictor (DTMP) and Weighted Selection Score (WSS) are indicated with red solid line box. Note that this DTMP is used to generate mask only in sample selection stage, ensuring accurate and consistent representation of the selection score.

Masked Image Modeling. Mask image modeling learns to extract visual representation by masking part of the input image and reconstructing the left visible parts. Taking the Masked Autoencoder (MAE) for example, given an input image $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, MIM methods first split it into a set of patches $\mathbf{x} \in \mathbb{R}^{M\times N\times (P^2C)}$, where C is the number of channels, (H, W) is the resolution of the image. (P, P) is the spatial resolution of each image patch, and $M = H/P$ and $N = W/P$ represent the numbers of rows and columns of the resulting patches. A mask $\mathbf{M} \in \{0, 1\}^{M\times N}$ is then applied on the image patches spatially. 1 indicates the corresponding patch is masked (invisible) for the encoder, while 0 represents the patch is visible. The visible patches $\mathbf{x}_V = \mathbf{x} \odot (1 - \mathbf{M})$ are then fed into the MAE encoder $\mathcal{E}$ and decoder $\mathcal{D}$, and the model tries to reconstruct the masked patches $\mathbf{x}_M = \mathbf{x} \odot \mathbf{M}$ by optimizing the reconstruction loss as follows:

$$\mathcal{L}_{\text{rec}} = \|\mathcal{D}(\mathcal{E}(\mathbf{x}_V)) - \mathbf{x}_M\|_2 \tag{1}$$

Sample Selection. Sample selection accelerates training by eliminating redundant data samples. The problem of constructing the most informative subset can be formulated as:

$$\mathcal{S}^* = \arg\max_{\mathcal{S}\subseteq\mathcal{T}} F(\mathcal{S}), \quad \text{s.t.} \quad \frac{|\mathcal{S}|}{|\mathcal{T}|} \leq \rho \tag{2}$$

that is, find the subset $\mathcal{S}$ of train data $\mathcal{T}$ that maximize some scoring function $F(\cdot)$ under a certain selection budget $\rho|\mathcal{T}|$. The scoring function $F(\cdot)$ evaluates how informative the subset is. Previous works have explored various criteria: by the loss incurred by $\mathcal{S}$ (Feldman 2020), by how frequently the sample is forgotten during training (Toneva et al. 2018), and by how much the subset can approximate the gradients (Feldman 2020; Killamsetty et al. 2021) etc. Here as suggested by (Feldman 2020), we simply use the loss incurred to evaluate the candidate samples.

Gap Between Sample and Masked Patches

Different from supervised learning which takes the image as a whole, MIM pretraining operates at the patch level. Following (Feldman 2020), which uses the loss incurred to indicate the informativeness, the most straightforward way to adapt sample selection to MIM is to aggregate all the reconstruction loss on the predicted patches as the selection score:

$$\mathcal{C}_{\text{sel}} = \frac{1}{\Omega(\mathbf{M})}\mathcal{M}\left(\mathcal{D}(\mathcal{E}(\mathbf{x}_V)), \mathbf{x}_M\right), \tag{3}$$

where $\mathcal{E}$ and $\mathcal{D}$ refer to encoder and decoder in Masked Image Modeling. $\mathcal{M}(\cdot,\cdot)$ represents the similarity measurement (the ℓ_2-distance in this paper), and $\Omega(\mathbf{M})$ is the number of masked patches. As in Equation 3, during the selection stage, we randomly mask patches and simply average the patch-wise loss to measure the informativeness of the sample. Samples with large aggregated loss are then used for subsequent training. The selection strategy is named basic sample selection (BSS), outlined in Figure 1.

However, as shown later in Table 1, this basic sample selection framework for MIM suffers from a noticeable performance degradation. Reviewing the Equation 3, we found two factors are not reasonable to reflect the sample selection score. The first is $\mathbf{x}_M = \mathbf{x} \odot \mathbf{M}$, which indicates the selection score depends on the mask patches.It is evident that different masked areas will affect the final sample importance. Furthermore, we illustrate the inconsistency and inaccuracy problem of selection score with random mask in Figure 3. The second is the simple averaging function that aggregates patch loss for scoring. The disparity between these masked patches can provide additional information for ranking the corresponding samples. However, simple averaging could lead to score homogeneity, that is, numerous samples exhibit

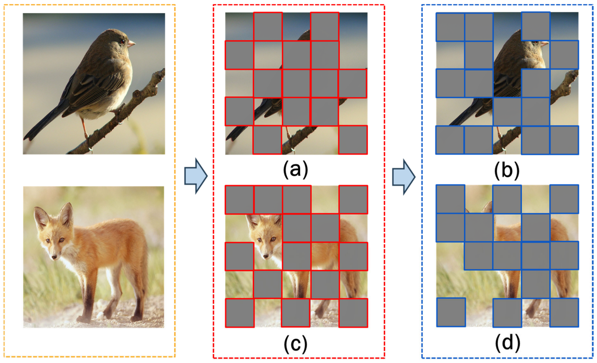

Figure 3: Examples of selection score with random mask in different selection rounds. Grey patch denotes the masked patch. Owing to the random mask, there exists inconsistency of the selection score both between rounds and within a round. For intra-round inconsistency, random mask may make some samples mostly represented by simple patches(e.g. (b)(c)) while the others mostly represented by complex patches(e.g. (a)(d)). This discrepancy leads to inaccurate in the selection scores within the round. Similarly, inter-round inconsistency follows the same logic.

similar selection scores, which will ruin the effectiveness of sample importance ranking. Then, we propose a Patch-Aware Sample Selection (PASS) method for Masked Image Modeling that leverages patch-level information with Dynamic Trained Mask Predictor and Weighted Selection Score.

Dynamic Trained Mask Predictor

Inspired by some previous work on supervised sample selection (Toneva et al. 2018; Paul, Ganguli, and Dziugaite 2021; Jiang et al. 2019; Wang et al. 2022), which utilize various metric to select more informative samples, we consider that this sample-level phenomenon also exists in patch-level. Specifically, within the same sample, some patches may contain abundant semantic information, while others may not. Therefore, selecting appropriate patches to reflect sample importance is essential in data selection stage.

Here we propose the Dynamic Trained Mask Predictor, a low-cost plug-and-play module for patch-based sample selection in Masked Image Modeling. With this module, we consistently mask the informative patches in different samples and selection rounds, ensuring the ranking order of each samples is relatively accurate and fair.

During selection, as illustrated in Figure 2, the crucial aspect lies in determining which areas to be masked within the sample. The mask prediction function is defined as follows:

$$\mathrm{M}^p(i,j) = \begin{cases} 1, & \mathbf{P}(i,j) \in \text{top-k}(\mathbf{P}), \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

where $\mathbf{P} = \mathcal{P}(\mathcal{E}(\mathbf{x}))$ is the output of mask predictor. $\mathcal{P}$ and $\mathcal{E}$ refer to the predictor and encoder respectively. The value of k in top-k is determined by the mask ratio and patch number of the sample.After that, this generated mask can be utilized in Equation 3 to obtain the corresponding selection score.

During training stage, we employ the Mean Squared Error (MSE) loss function to optimize this mask predictor:

$$\mathcal{L}_{\mathrm{pp}} = \frac{1}{MN} \sum_{i}^{M} \sum_{j}^{N} ((\mathbf{P}(i,j) - \mathbf{L}_{\mathrm{rec}}\,(i,j))^2 \cdot \mathbf{M}(i,j), \tag{5}$$

where $\mathbf{P}$ is the output of mask predictor, $\mathbf{L}_{\mathrm{rec}}$ is the detached reconstruction loss for all patches before reduction corresponding to $\mathcal{L}_{\mathrm{rec}}$, $\mathbf{M}$ is the mask.

Furthermore, we introduce a dynamic training scheme for our mask predictor to reduce the training cost of the additional mask predictor while maintaining model performance. As shown in Figure 4, the predictor doesn't require too much epochs to have a considerable ability of distinguishing the informative patches. Given that we only utilize mask predictor during selection stage, and our predictor does not require a particularly high discriminative ability, we dynamically modify the training loss function:

$$\mathcal{L}_{train} = \begin{cases} \mathcal{L}_{\mathrm{rec}} + \mathcal{L}_{\mathrm{pp}}, & \text{epoch } \in dt\_list \\ \mathcal{L}_{\mathrm{rec}}, & \text{otherwise} \end{cases} \tag{6}$$

where dt_list represents the allowed list of epochs for dynamic mask predictor training. Taking into account the dynamic changes in the subset during each sample selection, we perform γ epochs training on mask predictor branch dynamically before each sample selection stage. In this way, our predictor can keep up with the model pre-training process, while avoiding excessive additional training cost.

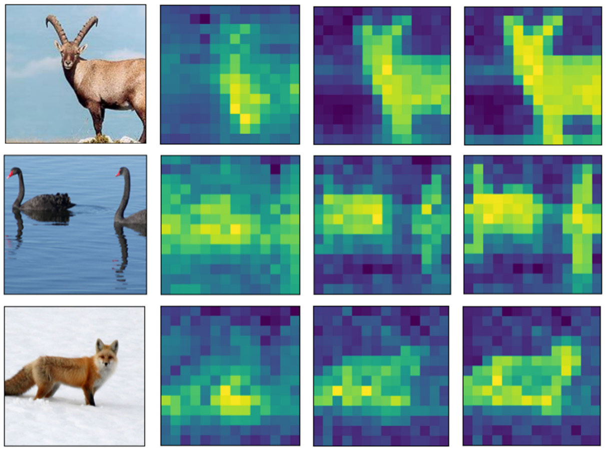

Figure 4: Visualization of the predictions of our mask predictor. For each row, the subgraph displays the original image on the left, followed by predictions from the predictor at different training epochs (10th, 20th, and 100th, respectively).

Weighted Selection Score

In this section, we enhance the simple average in the selection score. As shown in Figure 2, even selected by mask predictor, not all of the masked patches are informative and there still exists disparity among these patches. In this case, using a simple average function on these patches disregards the patch-level information. To further differentiate the masked

Method	Pre-training Data Budget	Pre-training Time	CIFAR-10		STL-10		CIFAR-100		ImageNet-1K	
			Linear	Finetune	Linear	Finetune	Linear	Finetune	Linear	Finetune
MAE	100%	1.0×	84.8	96.3	85.3	95.9	65.0	87.1	50.8	82.2
MAE-BSS	37%	∼0.41×	82.6	95.9	82.1	95.4	63.2	86.4	48.1	81.5
PASS (ours)	37%	∼0.59×	**85.6**	**97.2**	**86.1**	**96.4**	**66.3**	**87.7**	**52.5**	**82.9**
simMIM†	100%	1.0 ×	-	95.0	-	92.3	-	80.3	-	81.5
simMIM-BSS	37%	∼0.39×	-	94.6	-	91.1	-	80.2	-	81.3
PASS (ours)	37%	∼0.45×	-	**97.1**	-	**92.7**	-	**81.7**	-	**82.1**

Table 1: Top-1 accuracy (%) of finetuning and linear probing across STL10, CIFAR10/100, ImageNet-1K. BSS denotes Basic Sample Selection. PASS denotes Patch-Aware Sample Selection. † denotes the result referring to (Liu, Gui, and Luo 2023). Our proposed PASS achieves ∼1.7× and ∼2.2× acceleration on MAE and simMIM, respectively. Since simMIM is more computationally intensive than MAE, it results in a higher speedup with the same selection modules.

patches, we introduce a weighted selection score. The weight of masked patch can be determined as follows:

$$w_i = \begin{cases} \frac{\exp(p_i/\tau)}{\Sigma_{j=1}^{\Omega(\mathbf{M})}\exp(p_j/\tau)}, & \forall p_i, p_j \in \mathbb{P} \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where $\mathbb{P} = \{\mathbf{P}(i,j) : \mathbf{M}(i,j) = 1\}$, p_j represents the j-th output of the mask predictor, τ is temperature hyperparameter, $\Omega(\mathbf{M})$ refers to the number of the masked patches, w_i denotes the i-th patch weight. Then we utilize this patch weight to perform a weighted mean of the masked patch loss. Finally we obtain the weighted selection score:

$$\mathcal{C}_{\text{sel}} = \|\mathbf{W} \odot (\sum_{k=1}^{d}(\mathbf{x}_R(:,:,k) - \mathbf{x}_M^p(:,:,k))^2)\|_1 \tag{8}$$

where $\mathbf{W}$ is the weight tensor which stores the weights associated with each masked patch. $\mathbf{x}_R = \mathcal{D}(\mathcal{E}(\mathbf{x}_V^p))$ represents the reconstruction output with the predicted mask. d denotes the embedding dimension. $\mathbf{x}_V^p = \mathbf{x} \odot (1 - \mathbf{M}^p)$ and $\mathbf{x}_M^p = \mathbf{M}^p$, where $\mathbf{M}^p$ refers to the masked patches generated from Dynamic Trained Mask Predictor. Finally, combined with both DTMP and WSS, the pipeline of PASS is illustrated in Algorithm 1 .

Algorithm 1: Patch-Aware Sample Selection

```
Input:
Full dataset $\mathcal{T}$, keeping ratio of $\rho$, encoder $\mathcal{E}$, decoder $\mathcal{D}$, predictor $\mathcal{P}$
for $\mathbf{X} \in \mathcal{T}$ do
    $\mathbf{x} = patchify(\mathbf{X})$
    $\mathbf{P} = \mathcal{P}(\mathcal{E}(\mathbf{x}))$
    generate predicted mask as Equation (4)
        $\mathbf{M}^p(i,j) = \begin{cases} 1, & \mathbf{P}(i,j) \in \text{top-k}(\mathbf{P}), \\ 0, & \text{otherwise,} \end{cases}$
    obtain patch weight according to Equation (7) :
        $w_i = \begin{cases} \frac{\exp(p_i/\tau)}{\Sigma_{j=1}^{\Omega(\mathbf{M})}\exp(p_j/\tau)}, & \forall p_i, p_j \in \mathbb{P} \\ 0, & \text{otherwise} \end{cases}$
    output selection score as Equation (8):
        $\mathcal{C}_{sel} = \|\mathbf{W} \odot \Sigma^d \left(\mathcal{D}(\mathcal{E}(\mathbf{x}_V^p)) - \mathbf{x}_M^p\right)^2 \|_1$
end for
pick up $|\mathcal{S}| = \rho \cdot |\mathcal{T}|$ samples which have the highest $\mathcal{C}_{sel}$
return subset $\mathcal{S}$
```

Experiments

Datasets and Experimental Setups

Our study primarily evaluates performance on linear probing, classification finetune, object detection, instance segmentation, and semantic segmentation tasks in accordance with the setting of MAE (He et al. 2022) and simMIM (Xie et al. 2022a). This section provide a comprehensive overview of the utilized datasets and experimental settings.

Datasets. In this paper, we apply our method to popular MIM methods (MAE, simMIM), and evaluate with the linear probing and finetuning classification task on ImageNet-1K (Deng et al. 2009). Furthermore, we test the transferability of our method on other classification datasets such as CIFAR-10/100 (Krizhevsky 2009) and STL-10 (Coates, Ng, and Lee 2011). Additionally, to evaluate the generalization on semantics tasks, we conduct object detection and instance segmentation experiments on MS-COCO (Lin et al. 2014) and semantic segmentation on ADE20K (Zhou et al. 2019).

Experimental Setup. We pre-train MAE and simMIM on ImageNet-1K for 200 epochs following (He et al. 2022; Xie et al. 2022a). Our method is applicable to various ViT backbones, although the experiments are mainly conducted with ViT-B/16 encoder due to constrained computation resources. For pre-training, we patchify the image of 224×224 into 14×14 patches. We adopt the model with decoder with 8 blocks for MAE, while for simMIM, a linear head is used as the decoder. For fine-tuning, the decoder is omitted, and a fully-connected layer with an n-way output (n =1000 for ImageNet-1k) is appended to the output of the encoder as the classifier. For linear probing, we only train the last linear head while keeping the other layers frozen.

For PASS, we perform sample selection every 20 epochs during pre-training. For the first 20 epochs, we use the full training dataset, while for the following epochs, $\rho|\mathcal{T}|$ subset is selected using PASS for training according to the pre-defined data budget ρ. To keep the overall data budget consistent with the pre-defined data budget, we terminate the training once the pre-computed epoch is reached. We adopt a mask predictor consisting of n_d (by default $n_d = 8$) blocks for the Dynamic Trained Mask Predictor (DTMP). While for dynamic training, we set $\gamma = 3$. For Weighted Selection Score (WSS), we set $\tau = 0.1$. We set the mask ratio to 0.75 for MAE and 0.6 for simMIM, and the mask ratio remains consistent between the selection stage and pre-training stage. All experiments are conducted on 8 RTX-3090 GPUs.

Method	Detection			Segmentation		
	AP^b	AP$^b_{50}$	AP$^b_{75}$	AP^m	AP$^m_{50}$	AP$^m_{75}$
Random Init	28.13	46.11	29.78	26.17	43.54	27.61
MAE	41.54	61.70	45.61	37.63	58.75	40.31
MAE-BSS	38.46	58.35	42.1	35.16	55.67	37.64
PASS (ours)	42.25	61.93	46.32	38.03	59.28	40.78

Table 2: Results of object detection and instance segmentation on MS-COCO using Mask R-CNN. We adopt Mask R-CNN with FPN, and report the bounding box AP^b and mask AP^m on MS-COCO val2017.

ImageNet-1K Classification

We conduct pre-training of MAE and simMIM on ImageNet-1K for 200 epochs, followed by finetuning and linear probing. Additionally, we evaluate the transferability to other classification datasets such as STL10, CIFAR10 and CIFAR100 through finetuning and linear probing.

Finetuning Results. The results presented in Table 1 demonstrate a significant improvement in accuracy achieved by our PASS method compared to Basic Sample Selection (BSS) across all the datasets, thereby highlighting the superiority of our approach over selection with random mask. Even compared with the original MAE or simMIM methods that utilize 100% data budget for pre-training, our methods achieve comparable or even superior results across all the datasets with only 37% data budget. For example, our method improves the finetune accuracy on ImageNet-1K by 0.7% and 0.6% for MAE and simMIM, respectively. Moreover, benefiting from the small data budget, our method significantly reduces the pre-training time by only requiring 59% and 45% of the original pre-training time for these two MIM methods.

Linear Probing Results. The linear probing performance of MAE is evaluated, As shown in Table 1. Utilizing only 37% pre-training data budget, our PASS improves the linear probing accuracy on ImageNet-1K by 1.7% and 4.4% compared with the original MAE and MAE-BSS, respectively.

Method	DTMP	WSS	ADE20K		
			mIoU	aAcc	mAcc
MAE	/	/	40.64	79.84	51.12
MAE-BSS	/	/	39.52	79.51	49.94
PASS (ours)	✓	/	41.83	80.87	52.20
PASS (ours)	✓	✓	42.28	80.88	52.54

Table 3: Results of semantic segmentation on ADE20K using UperNet. The effectiveness of our DTMP and WSS modules is validated using mIoU, aAcc, mAcc as the metrics.

Object Detection and Instance Segmentation on MS-COCO

To further validate the learned visual representation of PASS, we test our MAE pre-trained model on the MS-COCO (Lin et al. 2014) object detection and instance segmentation. We take the Mask R-CNN (He et al. 2017) framework with FPNs (Lin et al. 2017) as the object detector, and apply our ViT-B/16 to this detection framework according to ViTDet (Li et al. 2022c) based on detectron2 (Wu et al. 2019). All pre-trained models are finetuned on the MS-COCO `train2017` for 1 × (12 epochs) with a resolution of 1024 × 1024 and batch size 16. Then we evaluate on the MS-COCO `val2017`

As shown in Table 2, we report AP^b for object detection and AP^m for instance segmentation. We observe that our method achieves the best result. Equipped with our method, it not only outperforms the basic sample selection baseline significantly, but also surpasses the original MAE. Our PASS outperforms +0.71 AP^b and +0.4 AP^m on MS-COCO.

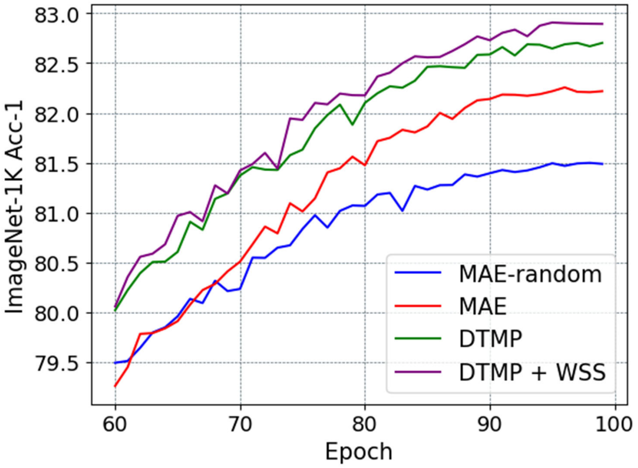

Figure 5: Ablation study of our proposed DTMP and WSS on ImageNet-1K finetuning.

Semantic Segmentation on ADE20K

We also evaluate our PASS on another dense prediction task, semantic segmentation on the ADE20K (Zhou et al. 2019) dataset. We utilize UperNet (Xiao et al. 2018) as the segmentation model and conduct finetuning for 80k iterations with a resolution of 512 × 512. To evaluate the performance, we consider mean Intersection of Union (mIoU), all pixel accuracy (aAcc), and mean accuracy of each class (mAcc) as the evaluation metrics. As illustrated in Table 3, the ablation study on ADE20K demonstrates the effectiveness of our PASS method. In comparison to the original MAE, our method achieves superior results with an improvement of +1.64 mIoU, +1.04 aAcc, and +1.42 mAcc, respectively.

Ablation Studies

Here we conduct several ablation studies to verify the effectiveness of our method.

Effectiveness of DTMP and WSS. As illustrated in Figure 5, both DTMP and WSS effectively enhance the performance in the proposed PASS, consistently improving performance compared to the original MAE.

MAE	Tr. Ep. of $\mathcal{P}$	Pre-training time	Finetune Top-1 Acc(%)
original	0	1x	82.22
+static-20	20	∼0.61×	82.49
+static-100	100	∼0.66×	82.54
+static-200	200	∼0.72×	**82.75**
+DTMP	30	∼**0.59**×	82.71

Table 4: Ablation study on dynamic training scheme. Each model is pre-trained on ImageNet-1K for 200 epochs. Static-N denotes that the mask predictor $\mathcal{P}$ is trained only in the first N epochs, whereas the training of our DTMP mask predictor keeps up with the selection stages

As shown in Table 4, DTMP achieves comparable accuracy with the lowest pre-training cost compared to the static training examples. Note that the time of our 30 epochs dynamic training of $\mathcal{P}$ remains lower than the 20 epochs static training due to the reduced data volume in the later epochs.

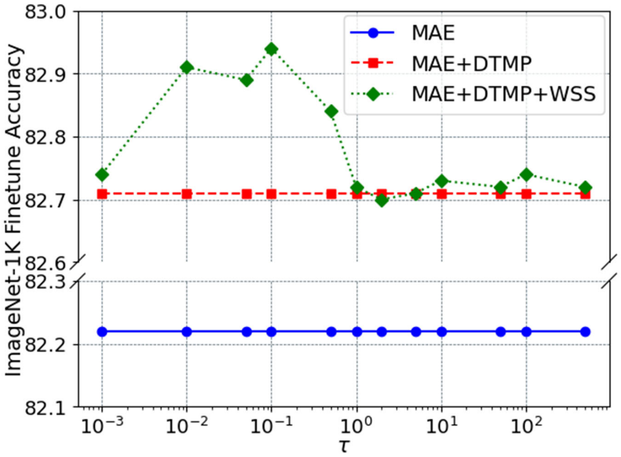

Figure 6: Impact of different τ. We evaluate each pre-trained model with Top-1 finetune accuracy on ImageNet-1K.

Ablation Study of τ and mask ratio. For different hyper-parameters, we maintain a fixed number of 200 pre-training epochs for MAE, then report the top-1 finetuning accuracy on ImageNet-1K. As depicted in Figure 6, we find that our WSS gains a consistent improvement when $\tau < 1$. While $\tau \geq 1$, the performance of WSS gradually approaches that of MAE with DTMP. This observation confirms our conjecture of strengthening the importance of more informative patches, since $\tau < 1$, the more informative patches will get more attention. As for the mask ratio, Table 5 shows that our PASS gains more speedup compared to the corresponding MAE pre-training with the mask ratio decreasing.

Mask Ratio(%)	40	60	75	90
Speedup Times	1.92×	1.83×	1.69×	1.51×
Top-1 Finetune Acc	82.72	82.89	**82.95**	82.25

Table 5: Impact of different mask ratios on ImageNet-1K finetuning for our PASS method.

Different Data Budget. As shown in Figure 7, we perform PASS on MAE with different data budget. Then we evaluate these pre-trained model on 4 classification datasets with linear probing. We observe a phenomenon of diminishing marginal effect of data budget. When the data budget is extra low, increasing data amount can achieve a significant improvement, but as the data budget increases, the performance gains become increasingly limited.

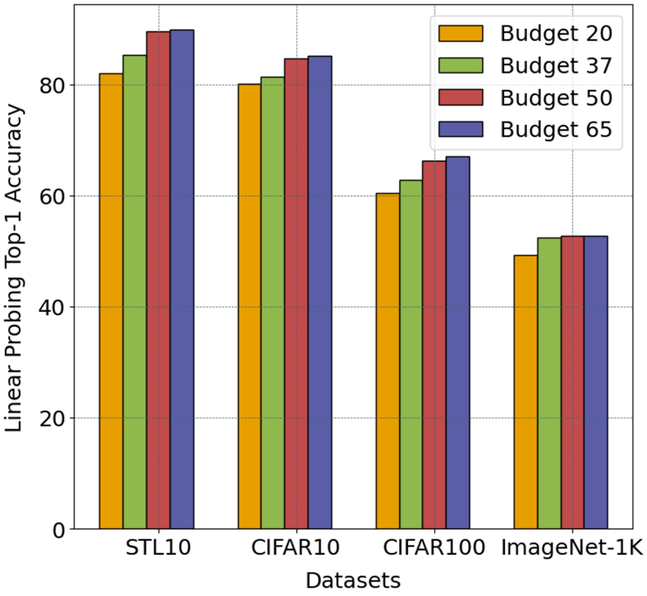

Figure 7: Ablation of different data budget across STL10, CIFAR10, CIFAR100 and ImageNet-1K

Conclusion

In this paper, we pioneer to accelerate MIM pre-training by Patch-Aware Sample Selection. PASS is a versatile method that can be seamlessly integrated with various MIM techniques, including MAE and simMIM. By leveraging the low-cost Dynamic Trained Mask Predictor and Weighted Selection Loss, our approach achieves remarkable acceleration in the pre-training phase without compromising performance. In fact, in many cases PASS surpasses the performance of the original MIM methods.

Acknowledgements

This work was supported in part by the National Key R&D Program of China (No.2022ZD0160304), the NSFC-General Technology Collaborative Fund for Basic Research

(No.U1936204), Jiangsu Key Research and Development Plan (No.BE2021012-2) and the CCF-Baidu Open Fund.

References

Bao, H.; Dong, L.; Piao, S.; and Wei, F. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15750–15758.

Citovsky, G.; DeSalvo, G.; Kumar, S.; Ramalingam, S.; Rostamizadeh, A.; and Wang, Y. 2023. Leveraging Importance Weights in Subset Selection. *arXiv preprint arXiv:2301.12052*.

Coates, A.; Ng, A.; and Lee, H. 2011. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 215–223. JMLR Workshop and Conference Proceedings.

Coleman, C.; Yeh, C.; Mussmann, S.; Mirzasoleiman, B.; Bailis, P.; Liang, P.; Leskovec, J.; and Zaharia, M. 2020. Selection via Proxy: Efficient Data Selection for Deep Learning. In *International Conference on Learning Representations*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

El-Nouby, A.; Izacard, G.; Touvron, H.; Laptev, I.; Jegou, H.; and Grave, E. 2021. Are large-scale datasets necessary for self-supervised pre-training? *arXiv preprint arXiv:2112.10740*.

Feldman, D. 2020. Introduction to Core-sets: an Updated Survey. *arXiv preprint*, arXiv:2011.09384.

Guo, J.; Han, K.; Wu, H.; Tang, Y.; Wang, Y.; and Xu, C. 2022. Fastmim: Expediting masked image modeling pre-training for vision. *arXiv preprint arXiv:2212.06593*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, L.; You, S.; Zheng, M.; Wang, F.; Qian, C.; and Yamasaki, T. 2022. Green hierarchical vision transformer for masked image modeling. *Advances in Neural Information Processing Systems*, 35: 19997–20010.

Jiang, A. H.; Wong, D. L.-K.; Zhou, G.; Andersen, D. G.; Dean, J.; Ganger, G. R.; Joshi, G.; Kaminksy, M.; Kozuch, M.; Lipton, Z. C.; et al. 2019. Accelerating deep learning by focusing on the biggest losers. *arXiv preprint arXiv:1910.00762*.

Ju, J.; Jung, H.; Oh, Y.; and Kim, J. 2022. Extending contrastive learning to unsupervised coreset selection. *IEEE Access*, 10: 7704–7715.

Killamsetty, K.; Sivasubramanian, D.; Ramakrishnan, G.; De, A.; and Iyer, R. 2021. GRAD-MATCH: Gradient Matching based Data Subset Selection for Efficient Deep Model Training. In *International Conference on Machine Learning*.

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images.

Li, G.; Zheng, H.; Liu, D.; Wang, C.; Su, B.; and Zheng, C. 2022a. Semmae: Semantic-guided masking for learning masked autoencoders. *Advances in Neural Information Processing Systems*, 35: 14290–14302.

Li, X.; Wang, W.; Yang, L.; and Yang, J. 2022b. Uniform masking: Enabling mae pre-training for pyramid-based vision transformers with locality. *arXiv preprint arXiv:2205.10063*.

Li, Y.; Mao, H.; Girshick, R.; and He, K. 2022c. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, 280–296. Springer.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, Z.; Gui, J.; and Luo, H. 2023. Good helper is around you: Attention-driven masked image modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1799–1807.

Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. 2022. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12009–12019.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Loshchilov, I.; and Hutter, F. 2015. Online batch selection for faster training of neural networks. *arXiv preprint arXiv:1511.06343*.

Mirzasoleiman, B.; Bilmes, J. A.; and Leskovec, J. 2020. Coresets for Data-efficient Training of Machine Learning Models. In *International Conference on Machine Learning, ICML*, volume 119 of *Proceedings of Machine Learning Research*, 6950–6960. PMLR.

Paul, M.; Ganguli, S.; and Dziugaite, G. K. 2021. Deep Learning on a Data Diet: Finding Important Examples Early in Training. *Advances in Neural Information Processing Systems*, 34: 20596–20607.

Ren, S.; Wei, F.; Albanie, S.; Zhang, Z.; and Hu, H. 2023. DeepMIM: Deep Supervision for Masked Image Modeling. *arXiv preprint arXiv:2303.08817*.

Settles, B. 2011. From theories to queries: Active learning in practice. In *Active learning and experimental design workshop in conjunction with AISTATS 2010*, 1–18. JMLR Workshop and Conference Proceedings.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sorscher, B.; Geirhos, R.; Shekhar, S.; Ganguli, S.; and Morcos, A. 2022. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35: 19523–19536.

Toneva, M.; Sordoni, A.; des Combes, R. T.; Trischler, A.; Bengio, Y.; and Gordon, G. J. 2018. An Empirical Study of Example Forgetting during Deep Neural Network Learning. In *International Conference on Learning Representations*.

Tong, Z.; Song, Y.; Wang, J.; and Wang, L. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357. PMLR.

Wang, H.; Tang, Y.; Wang, Y.; Guo, J.; Deng, Z.-H.; and Han, K. 2023. Masked Image Modeling with Local Multi-Scale Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2122–2131.

Wang, J.; Li, Y.; Zhuo, J.; Shi, X.; Zhang, W.; Gong, L.; Tao, T.; Liu, P.; Bao, Y.; and Yan, W. 2022. DynaMS: Dynamic Margin Selection for Efficient Deep Learning. In *The Eleventh International Conference on Learning Representations*.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, 568–578.

Wei, C.; Fan, H.; Xie, S.; Wu, C.-Y.; Yuille, A.; and Feichtenhofer, C. 2022. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14668–14678.

Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.-Y.; and Girshick, R. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; and Sun, J. 2018. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, 418–434.

Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; and Hu, H. 2022a. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9653–9663.

Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Wei, Y.; Dai, Q.; and Hu, H. 2022b. On Data Scaling in Masked Image Modeling. *arXiv preprint arXiv:2206.04664*.

Zhang, S.; Zhu, F.; Zhao, R.; and Yan, J. 2022. Contextual Image Masking Modeling via Synergized Contrasting without View Augmentation for Faster and Better Visual Pretraining. In *The Eleventh International Conference on Learning Representations*.

Zhou, B.; Zhao, H.; Puig, X.; Xiao, T.; Fidler, S.; Barriuso, A.; and Torralba, A. 2019. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127: 302–321.

Zhou, J.; Wei, C.; Wang, H.; Shen, W.; Xie, C.; Yuille, A.; and Kong, T. 2021. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*.