Bootstrapped Masked Autoencoders
for Vision BERT Pretraining

Xiaoyi Dong[1,⋆], Jianmin Bao[2], Ting Zhang[2], Dongdong Chen[3,†],
Weiming Zhang[1], Lu Yuan[3], Dong Chen[2], Fang Wen[2], Nenghai Yu[1]

[1]University of Science and Technology of China
[2]Microsoft Research Asia [3]Microsoft Cloud + AI
{dlight@mail., zhangwm@, ynh@}.ustc.edu.cn cddlyf@gmail.com
{jianbao, Ting.Zhang, luyuan, doch, fangwen }@microsoft.com

Abstract. We propose bootstrapped masked autoencoders (BootMAE), a new approach for vision BERT pretraining. BootMAE improves the original masked autoencoders (MAE) with two core designs: 1) momentum encoder that provides online feature as extra BERT prediction targets; 2) target-aware decoder that tries to reduce the pressure on the encoder to memorize target-specific information in BERT pretraining. The first design is motivated by the observation that using a pretrained MAE to extract the features as the BERT prediction target for masked tokens can achieve better pretraining performance. Therefore, we add a momentum encoder in parallel with the original MAE encoder, which bootstraps the pretraining performance by using its own representation as the BERT prediction target. In the second design, we introduce target-specific information (e.g., pixel values of unmasked patches) from the encoder directly to the decoder to reduce the pressure on the encoder of memorizing the target-specific information. Thus, the encoder focuses on semantic modeling, which is the goal of BERT pretraining, and does not need to waste its capacity in memorizing the information of unmasked tokens related to the prediction target. Through extensive experiments, our BootMAE achieves 84.2% Top-1 accuracy on ImageNet-1K with ViT-B backbone, outperforming MAE by +0.8% under the same pre-training epochs. BootMAE also gets +1.0 mIoU improvements on semantic segmentation on ADE20K and +1.3 box AP, +1.4 mask AP improvement on object detection and segmentation on COCO dataset. Code is released at https://github.com/LightDXY/BootMAE.

Keywords: Vision Transformer, BERT Pre-training, Bootstrap, Masked Autoencoder

1 Introduction

Self-supervised representation learning [53,43,55,60,25,31,11], aiming to learn transferable representation from unlabeled data, has been a longstanding prob-

---

⋆ Work done during an internship at Microsoft Research Asia. † Dongdong Chen is the corresponding author.

lem in the area of computer vision. Recent progress has demonstrated that large-scale self-supervised representation learning leads to significant improvements over the supervised learning counterpart on challenging datasets. Particularly, masked image modeling (MIM) in self-supervised pre-training for vision transformers has shown remarkably impressive downstream performance in a wide variety of computer vision tasks [22,4], attracting increasing attention.

MIM aims to recover the masked region based on remaining visible patches. Essentially, it learns the transferable representation through modeling the image structure itself by content prediction. A very recent work, masked autoencoder (MAE) [28], introduces an asymmetric encoder-decoder structure where the encoder only operates on visible patches, and the output representation of the encoder along with masked tokens are fed into a lightweight decoder. Shifting the mask tokens into the small decoder results in a large reduction in computation. Besides efficiency, it also achieves competitive accuracy (87.8%), equipped with the ViT-Huge backbone, among methods that only use ImageNet-1K data.

In this paper, we introduce bootstrapped masked autoencoders (BootMAE), a new framework for self-supervised representation learning with two core designs. Firstly, we observe that with the same structure design as MAE, just changing the MIM prediction target from the pixels to the representation of a pretrained MAE encoder boosts the ImageNet classification accuracy from 83.4% to 83.8% using a ViT-Base backbone. Motivated by this observation, we propose to use a momentum encoder to provide an extra prediction target. The momentum encoder is a temporal ensemble of the MAE encoder, *i.e.*, the weights are parameterized by an exponential moving average (EMA) of the MAE encoder parameters [29,27]. For each iteration, we pass the full image to the momentum encoder to provide ground-truth representation for masked patches, and pass the masked image to the encoder followed by a predictor to generate predictions for masked patches. We hypothesize that as the training proceeds, the momentum encoder provides dynamically deeper semantics than fixed targets via bootstrapping. We keep the pixel regression branch in MAE as a good regularization in differentiating images. Moreover, it also provides guidance for the model to learn reasoning about low-level textures. Such multiple supervision helps learn the representation that benefits broader tasks including high-level recognition and accurate pixel-wise prediction that requires low-level information.

Secondly, we propose the target-aware decoder that tries to reduce the pressure on the encoder to memorize target-specific information and encourage the encoder focus on semantic modeling that benefits for pre-training. Recall that MIM aims to recover the missing region given the visible patches. It is based on the fact that natural images, regardless of their diversity, are highly structured (for example, the regular pattern of buildings, the structured shape of cars). The goal of MIM is to enable the model to understand this structure, or so-called semantics, or equivalently the relationship of different patches in the prediction target space (either pixel space or feature space). Afterwards, the predictions are made by two indispensable ingredients: the knowledge of this structure and the target-specific information (*e.g.*, pixel values) of the visible patches. Yet

previous MIM methods couple the two ingredients in a single module, wasting the model capability in "memorizing" the target-specific information of visible patches. In comparison, we try to decouple them so that the encoder exploits its whole model capability for structure learning. More specifically, the target-specific information is explicitly and continuously given to the decoder, just like we humans always see the visible patches when making visual predictions.

In summary, our framework, as illustrated in Figure 1, contains four components: (1) an encoder that aims to capture the structure knowledge; (2) a regressor that takes the structure knowledge from the encoder along with the low-level context information for pixel-level regression; (3) a predictor that takes the structure knowledge from the encoder and the high-level context information for latent representation prediction; (4) feature injection modules in both regressor decoder and predictor decoder, responsible for incorporating each own necessary target-specific information.

In addition, we find that masking strategy is crucial for these two different prediction targets. They favor different masking strategies. Particularly, pixel regression relies on random masking while block-wise masking is better for feature prediction. The reason might be that block-wise masking tends to remove large blocks, which is a difficult task for pixel regression as pixel regression heavily relies on hints from local neighbors for prediction. While for feature prediction not compelled by precise pixel-wise alignment, a large masked patch is more helpful for the model to reason about the semantic structure.

In the experiments we demonstrate the effectiveness of our framework in various downstream tasks including image classification, object detection and semantic segmentation. Our approach achieves superior performance than previous supervised methods as well as self-supervised methods. We also provide extensive ablation studies validating that the two core designs in our model works. We further provide comprehensive comparison with MAE in various epochs and various models and show our framework achieves consistently better performance.

2   Related Works

Computer vision has made tremendous progress on image content understanding in the past decade. The features learned by neural networks trained on ImageNet using over 1 million images associated with labels usually generalize very well across tasks [19,46,34,9]. Another line of image content understanding explores whether such semantically informative features can be learned through raw images alone [23,18,53,43,55,60,25,31,11,29]. Representative methods along this line include autoencoding, clustering based, contrastive learning and masked image modeling.

Autoencoding. Autoencoding (AE) [32,5] is a type of neural networks used to learn a representation (embedding) for unlabeled data. It consists of an encoder that maps data to a low-dimensional latent embedding and a decoder that recovers the data from the latent embedding, with the goal of learning a compact feature representation for the data. AE is commonly used for feature selection

and feature extraction. The denoising autoencoder (DAE) [48] learns the representation robust to noise as the observed data in encoder is an addition of the original data and the noise. The decoder aims to undo the noise and recover the original data. Numerous efforts generalize DAE using different noise modelings, such as masking pixels [49,45,10], removing color channels [56,35], and shuffling image patches [42] and so on.

Clustering based methods. Clustering is a class of unsupervised learning methods that has been widely studied in the computer vision community. Traditional works are mostly designed under the assumption that feature representation is fixed. With the emerging of deep learning area, lots of efforts [55,53] explore adapting clustering to the end-to-end training to jointly learn the feature representation as well as clustering. The representative work DeepCluster [6] utilizes k-means to generate pseudo-labels to alternatively update the weights of the convnets and the clustering assignments of the image descriptors. Recently, several attempts [1,7] aim to maximize the mutual information between the pseudo labels and the input data, scaling up to large datasets.

Contrastive learning. Contrastive learning aims to learn an embedding space where similar data pairs stay close to each other while dissimilar data pairs are far apart. In the self-supervised scenario, it can be interpreted as a special case of clustering where each instance itself forms a class. Thus, the positive pairs are formed by two augmented views of the same image and the negative paris are views from different images. The typical methods include MOCO [29,14,15], Sim-CLR [11,12], BYOL [27] and more [44,36,2]. However, contrastive-based methods rely heavily on the data augmentation strategies that need to introduce non-essential variations while without modifying semantic meanings. Crucial augmentations include random cropping and random color distortion. Meanwhile, large quantities of negative samples are usually required in order to avoid trivial solution in which the model outputs a constant representation for all data.

Masked image modeling. Masked image modeling for self-supervised pre-training has recently grown in popularity motivated by the success of BERT pre-training in NLP [17]. ViT [22] and BEiT [4] are two initiatives along this direction. MIM that predicts masked patches from visible ones in a sense can be viewed as context prediction. Feature representation learned through such within-image context prediction shows surprisingly strong performance in downstream tasks. Recently, lots of works [28,26] exploring MIM have been concurrently developed from different perspectives. The efforts include (i) framework design, such as MAE [28], SplitMask [24], SimMIM [54], CAE [13]; (ii) prediction targets, such as PeCo [21], MaskFeat [51], data2vec [3], iBOT [59]; (iii) video extension BEVT [50]; (iv) integration with vision-language contrastive learning FaRL[57]. Our work belongs to the first group and introduces a novel framework called Bootstrapped MAE. We progressively bootstrap the latent representation in MAE to learn dynamically deeper semantics. Moreover, in comparison with previous methods coupling the context information with semantic modeling in a single model, we separate them by explicitly passing the context information to the decoder so that the encoder leverage the whole model for structure learning.
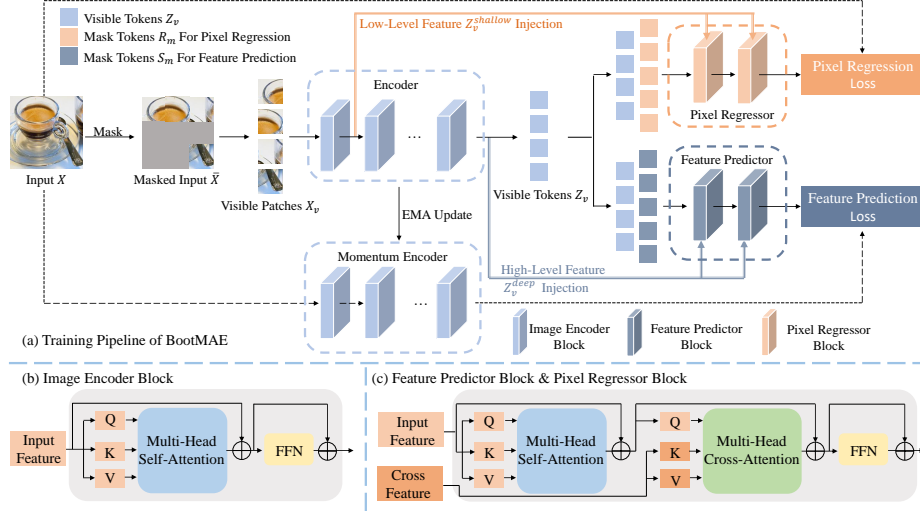
Fig. 1: Illustrating the details of our BootMAE in (a) the overall framework and training pipeline, (b) the image encoder block, (c) the feature predictor block & pixel regressor block.

3   Approach

In this section, we introduce our Bootstrapped MAE framework in details. As illustrated in Figure 1, our framework contains four components: 1) the encoder network focusing on learning the structure knowledge; 2) the pixel regressor decoder network aiming to predict the missing pixels of the masked region given the structure knowledge from the encoder and the context information from the visible patches,i.e., pixel values or low-level features in this case; 3) the feature predictor decoder network attempting to make feature predictions for the masked region given the same structure knowledge from the encoder and the context information of the visible patches, i.e., high-level feature information in this case; and 4) feature injection modules that feed each own context information into the regressor and the predictor explicitly and continuously in each decoder layer. After self-supervised pre-training with our BootMAE, we adopt the encoder network for various downstream tasks.

Formally, suppose an input image is $\boldsymbol{X} \in \mathbb{R}^{H \times W \times C}$, where H and W denote the image height and image width and C denotes the channel number, we first split it into non-overlapping patches. This results in $N = H \times W/P^2$ patches with P denotes the resolution of each patch. In this way, an image is represented by a number of patches $\boldsymbol{X} = \{x^1, x^2, \cdots, x^N\}$ with $x^n \in \mathbb{R}^{P^2C}$ denotes a vector reshaped from the image patch. Thereafter, a large fraction of, say N_m patches are randomly sampled to be masked and leave the remaining N_v patches to be visible, $N = N_m + N_v$. Let $\mathcal{M}$ be the index set of masked patches, $\boldsymbol{X}_v = \{x^k | k \notin \mathcal{M}\}$ denotes the set of visible patches and $\boldsymbol{X}_m = \{x^k | k \in \mathcal{M}\}$ denotes the set

of masked patches, we have $\boldsymbol{X} = \boldsymbol{X}_v \cup \boldsymbol{X}_m$ and $\boldsymbol{X}_v \cap \boldsymbol{X}_m = \emptyset$. Generally, each patch is associated with a positional embedding indicating the location of each patch. Therefore similarly, we have $\boldsymbol{P}_v$, positional embeddings for the visible patches and $\boldsymbol{P}_m$, positional embeddings for the masked patches.

3.1 Encoder

The encoder aims to exploit the whole capability to output a latent representation that models the image structure. Inspired by MAE [28], the encoder only handles the visible patches $\boldsymbol{X}_v$ for training efficiency and outputs the latent representation $\boldsymbol{Z}_v$. Specifically, we first project each visible patch into an embedding and add a positional embedding on each embedding to ensure the awareness of position for each patch. After this, the combined embedding is fed to a ViT [22] composed of a stack of standard vision Transformer blocks based on self-attention. That is,

$$\boldsymbol{Z}_v = \mathrm{Enc}(\boldsymbol{X}_v, \boldsymbol{P}_v). \tag{1}$$

The computation and memory are really efficient even for large scale models as only a small subset (e.g.,25%) of the image patches needs to be handled by the encoder. Moreover, the elimination of the special mask tokens bridges the gap between pre-training and fine-tuning as the fine-tuning stage sees real visible patches without any mask token [28,26]. The mask region (e.g., 75%) is randomly sampled from the image.

We find that the masking strategy is crucial as different prediction targets favor different masking strategies. We study this masking strategy and provide more analysis in the experiments. We provide explanations about which masking strategy is suitable for which prediction target, and reach a conclusion offering guidance about the choice of masking strategy. In our implementation, we adopt the block-wise masking strategy. The output is further normalized to $\hat{\boldsymbol{Z}}_v = \mathrm{norm}(\boldsymbol{Z}_v)$ which captures the image structure and is fed to the subsequent decoders.

3.2 Feature injection module

As mentioned in the introduction, there are two indispensable ingredients for the decoder to make predictions: the structure knowledge and the corresponding context information from the visible patches. Our feature injection module is designed to directly feed the context information into each layer of the decoder. We argue that in this way, the encoder exploits the whole model capability to learn structure knowledge without considering "memorizing" the context information of visible patches that is related to prediction target.

In particular, different prediction targets require different context information. Specifically, pixel-level prediction focusing on low-level details probably favors low-level context information of the visible patches while feature-level prediction attempting to predict semantic feature representation probably needs

high-level context information of the visible patches. Therefore, we feed the features from the shallow layer of the encoder to the regressor decoder and the features from the deep layer of the encoder to the predictor decoder. We use $\boldsymbol{Z}_v^{shallow}$ to represent the shallow features and $\boldsymbol{Z}_v^{deep}$ to represent the deep features of the encoder.

Instead of using addition or concatenation, we adopt a very elegant operator cross-attention. To be specific, we leverage the feature from the encoder as keys and values and the features from the regressor/predictor as queries to perform cross-attention. This operator helps leverage the low-level information for better pixel reconstruction and the high-level semantics for feature prediction. We apply this cross attention right after the self-attention in each transformer block of the regressor and predictor.

3.3   Regressor

The regressor aims to recover the missing pixels as in [28]. The pixel-level regression not only helps prevent the model from collapsing but also guide the model to learn reasoning about low-level textures. The input of the regressor includes (1) the normalized latent representation output from the encoder and (2) the shallow features providing context information. We add mask tokens $\boldsymbol{R}_m$ containing N_m learnable vectors representing masked patches to be predicted. To ensure that each mask token is aware of its location in the image, we add the positional embedding to each mask token. We adopt a lightweight architecture for the regressor, consisting of two vision transformer blocks and a fully-connected layer to predict missing pixels. Let the output of the regressor be $\bar{\boldsymbol{X}}$, the formulation can be written as,

$$\bar{\boldsymbol{X}} = \mathrm{Reg}(\hat{\boldsymbol{Z}}_v, \boldsymbol{Z}_v^{shallow}, \boldsymbol{R}_m, \boldsymbol{P}_m). \tag{2}$$

The regressor makes prediction based on the structure knowledge in $\hat{\boldsymbol{Z}}_v$ and the context information in $\boldsymbol{Z}_v^{shallow}$.

3.4   Predictor

The predictor aims to predict the feature representation of the masked patches. This high-level feature prediction target guides the model to learn reasoning about high-level semantics. Moreover, the prediction groundtruth is the representation itself which evolves along with the training, providing richer and deeper semantics than fixed targets. The input of the predictor includes (1) the normalized latent representation same with the regressor and (2) the deep features providing context information different from the regressor. We also add another set of mask tokens $\boldsymbol{S}_m$ representing the masked patches to be predicted and associate them with positional embeddings. The predictor decoder network consists of two transformer blocks with a MLP layer for prediction. Say the output of the predictor is $\bar{\boldsymbol{F}}$, the formulation can be written as,

$$\bar{\boldsymbol{F}} = \mathrm{Pre}(\hat{\boldsymbol{Z}}_v, \boldsymbol{Z}_v^{deep}, \boldsymbol{S}_m, \boldsymbol{P}_m). \tag{3}$$

Similarly, the predictor makes prediction based on the structure knowledge in $\hat{\boldsymbol{Z}}_v$ and the context information in $\boldsymbol{Z}_v^{deep}$.

3.5   Objective function

The regressor and the predictor output all predictions for both visible patches as well as masked patches, but only the predictions for masked patches are involved in the loss calculation. For the regressor, each element in the output is a vector of pixel values representing a patch. We use normalized pixels as the reconstruction target for groundtruth as MAE [28]. The objective function for the regressor is,

$$\mathcal{L}_R = \sum_{k \in \mathcal{M}} \frac{1}{P^2 C} ||g_m^k - \bar{x}_m^k||_2^2, \tag{4}$$

where g_m^k is the normalized patch of x_m^k using the mean and standard deviation computed from all pixels in that patch. $\bar{x}_m^k$ is the reconstructed masked patch in $\bar{\boldsymbol{X}}$

For the predictor, the prediction feature groundtruth is the latent representation itself by passing a full image into the momentum encoder where the weights are parameterized by an exponential moving average of the MAE encoder. Let $\boldsymbol{F} = \text{Enc}_{ema}(\boldsymbol{X}, \boldsymbol{P})$ be the groundtruth, the objective function over the masked patches for predictor is,

$$\mathcal{L}_P = \sum_{k \in \mathcal{M}} \frac{1}{\#\text{dim}} ||f_m^k - \bar{f}_m^k||_2^2, \tag{5}$$

where $\#\text{dim}$ denotes the feature dimension of the token, and $f, \bar{f}$ is one of token in $F, \bar{F}$. The overall loss is a weighted sum,

$$\mathcal{L} = \mathcal{L}_R + \lambda \mathcal{L}_P, \tag{6}$$

where λ is the hyperparameter tuning the loss weight.

4   Experiments

4.1   Implementations

We experiment with the standard ViT[1] base and large architectures, ViT-B (12 transformer blocks with dimension 768) and ViT-L (24 transformer blocks with dimension 1024) for the encoder. The regressor and the predictor consist of 2 transformer blocks as mentioned above. The dimension of the regressor is set to 512 while the the dimension of the predictor is set to the same as the encoder for feature prediction. The input is partitioned 14×14 patches from the image of 224×224, and each patch is of size 16x16. Following the setting in MAE, we only

---

[1] we didn't use some recent techniques like relative position or layer scaling.

Table 1: The effect of bootstrapped feature prediction. The performance with pre-trained 300 epochs gets improvement from 83.2% to 83.6%, achieving the same performance with the vanilla MAE with pre-trained 1600 epochs.

Model	Pre-train Epoch	Fine-tuning
MAE	1600	83.6
MAE	800	83.4
MAE	300	83.2
MAE w bootstrapped feature prediction	300	83.6

Table 2: Ablation studies showing the effect of feature injection module in our framework. Providing context for both regressor and predictor achieves the best performance, suggesting that in this target-aware decoder design, the encoder indeed learns stronger semantic modeling.

Model	Context for regressor	Context for predictor	Fine-tuning
BootMAE	×	×	83.6
BootMAE	✓	×	83.9
BootMAE	✓	✓	84.0

use standard random cropping and horizontal flipping for data augmentation. We find that the different prediction tasks favor different masking strategies. We choose the block-wise masking strategy to benefit for feature prediction. The total masking ratio is 75%, same with that in MAE [28]. Both ViT-B and ViT-L model are trained for 800 epochs with batch size set to 4096. We use Adam [33] and a cosine schedule [41] with a single cycle where we warm up the learning rate for 40 epochs to $2.4e^{-3}$. The learning rate is further annealed following the cosine schedule. Our proposed method is pre-trained on ImageNet. The regressor and the predictor are only used during pre-training. After pre-training, only the encoder is used to generate the image representation.

For ImageNet experiments, we average pool the output of the last transformer of the encoder and feed it to a softmax-normalized classifier. We evaluate the pre-trained feature representation using end-to-end fine-tuning along with the backbone model. We fine-tune 100 epochs for ViT-B and 50 epochs for ViT-L. The learning rate are warmed up to 0.005 for 20 epochs for ViT-B and 0.0015 for 5 epochs for ViT-L, after which followed by cosine schedule. The evaluation metric is top-1 validation accuracy of a single 224×224 crop.

4.2   Analysis of BootMAE

Bootstrapped feature prediction. One core design of our framework is the bootstrapped feature prediction that predicts the iteratively evolved latent representation of the image to enable the model to learn from dynamically richer semantic information. Here we investigate the effect of adding this proposed bootstrapped feature prediction branch. The compared models are the vanilla MAE and the MAE with an additional bootstrapped feature prediction without

Table 3: Results comparison of two different masking strategies, random masking and block-wise masking, with two different prediction targets, pixel-level target and feature-level target. This validates our hypothesis that pixel-level target favors random masking while feature-level target favors block-wise masking.

Prediction target	Mask Strategy		Fine-tuning Accuracy
	Random	Block	
Pixel-level	✓		83.2
Pixel-level		✓	82.8
Feature-level	✓		83.1
Feature-level		✓	83.6

Table 4: We study the regressor and predictor design and ablate the performance in terms of (a) the regressor and predictor depth and (b) the feature dimension of the regressor.

(a) The regressor and predictor depth		(b) The feature dimension of regressor	
Depth	Fine-tuning	#Dim	Fine-tuning
1	83.4	256	83.8
2	84.0	384	84.0
4	83.9	512	84.0
8	84.0	768	84.0

the feature injection module. The comparison results are presented in Table 1. We observe that the performance with pre-trained 300 epochs gets improvement from 83.2% to 83.6%, achieving the same performance with the vanilla MAE with pre-trained 1600 epochs. This demonstrates the effectiveness of the proposed bootstrapped feature prediction. Based on this result, we further analysis BootMAE under the 300 epoch pre-training setting in the following.

Feature injection. Another important design in our framework is the feature injection module, which provides different features representing different level of context information for the regressor and the predictor. Specifically, we explicitly feed the feature outputted from the first layer of the encoder to each layer of the regressor to ease the burden of the encoder in "memorizing" the low-level details so that the encoder focuses on structure modeling. Similarly, we directly feed the features from the last layer of the encoder to each layer of the predictor. Here we study the effect of the proposed feature injection and the ablated results are shown in Table 2. We can see that providing both regressor and predictor with each own necessary context achieves the best performance, the encoder indeed learns stronger semantic modeling due to the target-aware decoder design.

Masking strategy. Then we study two widely used masking strategies in masked image modeling: random block-wise masking in [4] and random masking in [28]. The masking ratio is the same and is set to 75%. It has been observed in MAE [28] that block-wise masking degrades at such a large ratio for their model. While in other scenarios, we find that block-wise masking is better than random masking. Here we provide explanations about why this is the case.

Table 5: We study the pre-training efficiency with the momentum encoder and report the performance as well as iteration time cost when feeding different fractions of the image to the momentum encoder.

Model	Fraction of Image for Momentum Encoder	Fine-tuning	Training Iter Time (s)	Speed Up
MAE	—	83.2	0.473	1×
BootMAE	25%	83.8	0.407	1.16×
BootMAE	50%	83.8	0.479	0.98×
BootMAE	75%	83.9	0.588	0.80×
BootMAE	100% (Default)	84.0	0.660	0.72×

We suspect that the reason may come from the prediction target. Pixel-level prediction target pursuing precise pixel-wise alignment requires visible neighboring patches to provide texture information, thus favoring that the masked region should be close to the visible region. While in block-wise masking, there always exists a larger continuous block of the image being masked and more masked patches are near the image center, making it difficult to the pixel-level prediction. As for feature-level prediction which cares less about the textures/details, block-wise masking largely reduces the redundancy and most center patches are masked, forcing the model to learn reasoning about the semantics.

We experiment the two masking strategies when using two different prediction targets and the results are given in Table 3. Here we train MAE for 300 epochs with different prediction target: pixel (MAE default setting) or output feature of a 800 epoch pretrained MAE model. The comparison validates our hypothesis analyzed above that pixel-level (feature-level) target favors random masking (block-wise masking). We adopt block-wise masking as bootstrapped feature prediction is key in our framework.

Regressor and predictor design. Our regressor and predictor are pretty lightweight consisting of two transformer layers. In this section, we vary the network depth (number of Transformer blocks) and experiment the performance when setting the depth to 1, 2, 4 and 8. The results are reported in Table 4 (a). We can see that using depth 2 or 8 achieves the best fine-tuning performance while depth 2 enjoys more efficiency. In addition we also study the feature dimension in regressor. Note that the feature dimension in predictor is set as the same with the encoder width. As shown in Table 4 (b). The fine-tuning accuracy with different dimensions is similar, except dim= 256 which is too small.

Pre-training efficiency with the momentum encoder. In our framework, we feed the full image to the momentum encoder to provide the feature prediction ground-truth. We observe that this extra inference incurs additional computation cost compared with MAE. Here we present specific training iteration time in Table 5. The validation is conducted with A100 GPU and batch size 256 per GPU for all models. We further study several variants that only a subset of the masked patches are fed into the momentum encoder and the prediction loss is only evaluated on this subset of masked patches. As the masking ratio is 75%, we

Table 6: Image classification accuracy (%) comparison on ImageNet-1K of different methods using various backbones. -B, -L stands for using ViT-B, ViT-L model, respectively. We report the fine-tuning and linear probing accuracy and our method BootMAE outperforms previous self-supervised methods.

Methods	Pre-train dataset	Pre-train epochs	ViT-B Fine-tuning	Linear	ViT-L Fine-tuning	Linear
Training from scratch (i.e., random initialization)						
ViT$_{384}$ [22]	-	-	77.9	-	76.5	-
DeiT [47]	-	-	81.8	-	–	-
ViT [28]	-	-	82.3	-	82.6	-
Self-Supervised Pre-Training on ImageNet-1K						
DINO [8]	IN-1K	300	82.8	78.2	–	–
MoCo v3 [15]	IN-1K	300	83.2	76.7	84.1	77.6
BEiT [4]	IN-1K + DALLE	800	83.2	56.7	85.2	73.5
MAE [28]	IN-1K	800	83.4	64.4	85.4	73.9
MAE* [28]	IN-1K	1600	83.6	68.0	85.9	76.6
BootMAE	IN-1K	300	84.0	64.1	85.4	74.8
BootMAE	IN-1K	800	<u>84.2</u>	66.1	**85.9**	77.1

study three fractions: 75% (all the masked patches), 50% (sampled from masked patches), 25% (also sampled from masked patches). We report the iteration time as well as the performance in Table 5. We can see that as with a smaller fraction of patches to the momentum encoder, the iteration time cost gets fewer while the performance gets lower due to the model only learns from a fraction of the masked tokens. It is worth noting that our method when feeding 25% image patches to the momentum encoder achieves better performance than MAE while is more efficient. This is because MAE adopt 8 layers for the decoder while our regressor and predictor only consist of 2 layers.

4.3 ImageNet Classification Comparison

We compare our methods with previous state-of-the-art works on ImageNet-1K classification task. We report the top-1 validation accuracy for both fine-tuning and linear probing results in Table 6. Compared to the supervised models trained from scratch, self-supervised pre-training methods achieve significant improvement, suggesting the effectiveness of pre-training.

We further compare our framework with prior self-supervised pre-training models. We can see that the proposed BootMAE achieves the best fine-tuning performance either based on ViT Base network or based on ViT Large network. For example, compared with the recent work MAE [28], our bootMAE with ViT-B achieves 84.2% top-1 accuracy with 0.8% gain, and with ViT-L achieves 85.9% with 0.5% improvement. We also report the linear probing accuracy. Our approach performs better than MIM based self-supervised methods, but not as good as the contrastive-based methods. We suspect that contrastive learning
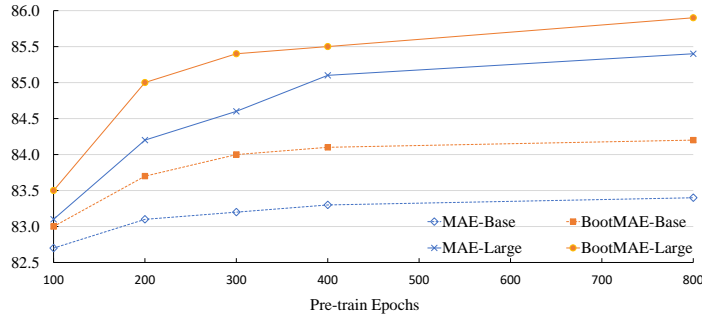
Fig. 2: Performance comparison with MAE in different pre-training epochs using ViT-B and ViT-L, showing that our BootMAE consistently outperforms MAE.

methods pursue linear features by comparing across images while MIM based methods exploit within image structure.

In addition, we present comprehensive comparison with MAE under different pre-training epochs for both ViT-B and ViT-L. We plot the results in Figure 2. We can see that our approach consistently performs better than MAE. It is worth mentioning that the proposed bootMAE at 200 epochs achieves 83.7% accuracy, which is alredy better than MAE pre-trained at 800 epochs. This demonstrate that our approach is more efficient to achieve similar performance, though with the extra inference of the momentum encoder. To be specific, under the same setting that using 16 V100 GPUs, MAE takes 51 hours for 800 epochs to get an 83.4% accuracy, while our BootMAE only takes 18 hours for 200 epochs to get a better result 83.7%.

4.4   Downstream Tasks

To further validate the learned visual representation of our BootMAE, we present transfer learning experiments on two popular downstream tasks.

Semantic segmentation. We compare our model on the widely used semantic segmentation dataset ADE20K [58]. We adopt UperNet framework [52] in the experiments. We train Upernet 160K iterations with batch size set as 16. We report the results in Table 7 (a). The evaluation metric is mean Intersection of Union (mIoU) averaged over all semantic categories and we report single-scale test results here. We compare our method with supervised pre-training on ImageNet-1K as well as state-of-the-art self-supervised models. We can see that the proposed BootMAE gets superior performance than all the other baselines, further validating the effectiveness of our framework.

Object detection and segmentation. We perform fine-tuning on the COCO object detection and segmentation [39]. We choose the Mask R-CNN [30] framework. Concretely, we adopt FPNs [38] to scale the feature map into different size as introduced in [37]. The fine-tuning is conducted with "1x" (12 training epochs) schedule and single-scale input on the COCO training set. The perfor-

Table 7: (a) Semantic segmentation mIoU (%) comparison on ADE20K. (b) Object detection and instance segmentation comparison in terms of box AP (AP^{bb}) and mask AP (AP^{mk}) on COCO. The same ViT-B backbone is used.

(a) Semantic segmentation			(b)Object detection and instance segmentation			
Models	Pre-train epochs	ADE-20K mIoU	Models	Pre-train epochs	COCO AP^{bb}	AP^{mk}
Supervised	300	47.4	Supervised	300	44.1	39.8
MoCo [16]	300	47.3	MoCo [16]	300	44.9	40.4
BEiT [4]	800	47.1	BEiT [4]	800	46.3	41.1
MAE [28]	800	47.6	MAE [28]	800	46.8	41.9
MAE* [28]	1600	48.1	MAE* [28]	1600	47.2	42.0
BootMAE	800	**49.1**	BootMAE	800	**48.5**	**43.4**

mance is tested on COCO validation set, following the strategy used in previous works [40,20]. The results are reported in Table 7 (b) in terms of box AP for detection and mask AP for segmentation. We observe that our model achieves 48.5% for object detection and 43.4% for segmentation, surpassing MAE by 1.3% and 1.4% respectively.

5 Discussion and Conclusion

In this paper, we introduce a new framework BootMAE with two core designs. (1) We propose to bootstrap the latent feature representation in MAE for better performance since the prediction target evolves with training, providing progressively richer information. (2) We propose to decouple the target-specific context from the encoder so that the encoder focuses on modeling the image structure. We present extensive experiments on various downstream tasks and comprehensive ablation studies to validate the effectiveness of the proposed framework. In addition, we find that different prediction targets likely favor different masking strategies.

Previous MIM methods couple the target-specific information with structure learning in a single model. We argue that as the goal of MIM through inpainting is essentially modeling the within image structure, it is advantageous to enable the whole encoder to focus on semantic modeling which we empirically demonstrate its advantage in the experiments. In the future, we would like to seek theoretical connection between semantic modeling and representation learning.

References

1. Asano, Y.M., Rupprecht, C., Vedaldi, A.: Self-labelling via simultaneous clustering and representation learning. arXiv preprint arXiv:1911.05371 (2019) 4
2. Bachman, P., Hjelm, R.D., Buchwalter, W.: Learning representations by maximizing mutual information across views. arXiv preprint arXiv:1906.00910 (2019) 4
3. Baevski, A., Hsu, W.N., Xu, Q., Babu, A., Gu, J., Auli, M.: Data2vec: A general framework for self-supervised learning in speech, vision and language. arXiv preprint arXiv:2202.03555 (2022) 4
4. Bao, H., Dong, L., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021) 2, 4, 10, 12, 14
5. Bengio, Y.: Learning deep architectures for AI. Now Publishers Inc (2009) 3
6. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: Proceedings of the European conference on computer vision (ECCV). pp. 132–149 (2018) 4
7. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in Neural Information Processing Systems **33**, 9912–9924 (2020) 4
8. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. arXiv preprint arXiv:2104.14294 (2021) 12
9. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017) 3
10. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I.: Generative pretraining from pixels. In: International Conference on Machine Learning. pp. 1691–1703. PMLR (2020) 4
11. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020) 1, 3, 4
12. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.: Big self-supervised models are strong semi-supervised learners. arXiv preprint arXiv:2006.10029 (2020) 4
13. Chen, X., Ding, M., Wang, X., Xin, Y., Mo, S., Wang, Y., Han, S., Luo, P., Zeng, G., Wang, J.: Context autoencoder for self-supervised representation learning. arXiv preprint arXiv:2202.03026 (2022) 4
14. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020) 4
15. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. arXiv preprint arXiv:2104.02057 (2021) 4, 12
16. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9640–9649 (2021) 14
17. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018) 4
18. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: Proceedings of the IEEE international conference on computer vision. pp. 1422–1430 (2015) 3

19. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: International conference on machine learning. pp. 647–655. PMLR (2014) 3

20. Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., Chen, D., Guo, B.: Cswin transformer: A general vision transformer backbone with cross-shaped windows. arXiv preprint arXiv:2107.00652 (2021) 14

21. Dong, X., Bao, J., Zhang, T., Chen, D., Zhang, W., Yuan, L., Chen, D., Wen, F., Yu, N.: Peco: Perceptual codebook for bert pre-training of vision transformers. arXiv preprint arXiv:2111.12710 (2021) 4

22. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 2, 4, 6, 12

23. Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with convolutional neural networks. Advances in neural information processing systems **27**, 766–774 (2014) 3

24. El-Nouby, A., Izacard, G., Touvron, H., Laptev, I., Jegou, H., Grave, E.: Are large-scale datasets necessary for self-supervised pre-training? arXiv preprint arXiv:2112.10740 (2021) 4

25. Ermolov, A., Siarohin, A., Sangineto, E., Sebe, N.: Whitening for self-supervised representation learning. In: International Conference on Machine Learning. pp. 3015–3024. PMLR (2021) 1, 3

26. Fang, Y., Dong, L., Bao, H., Wang, X., Wei, F.: Corrupted image modeling for self-supervised visual pre-training. arXiv preprint arXiv:2202.03382 (2022) 4, 6

27. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., et al.: Bootstrap your own latent: A new approach to self-supervised learning. arXiv preprint arXiv:2006.07733 (2020) 2, 4

28. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. arXiv preprint arXiv:2111.06377 (2021) 2, 4, 6, 7, 8, 9, 10, 12, 14

29. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020) 2, 3, 4

30. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017) 13

31. Henaff, O.: Data-efficient image recognition with contrastive predictive coding. In: International Conference on Machine Learning. pp. 4182–4192. PMLR (2020) 1, 3

32. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. science **313**(5786), 504–507 (2006) 3

33. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 9

34. Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., Houlsby, N.: Big transfer (bit): General visual representation learning. In: European conference on computer vision. pp. 491–507. Springer (2020) 3

35. Larsson, G., Maire, M., Shakhnarovich, G.: Learning representations for automatic colorization. In: European conference on computer vision. pp. 577–593. Springer (2016) 4

36. Li, S., Chen, D., Chen, Y., Yuan, L., Zhang, L., Chu, Q., Liu, B., Yu, N.: Improve unsupervised pretraining for few-label transfer. In: ICCV (2021) 4

37. Li, Y., Xie, S., Chen, X., Dollar, P., He, K., Girshick, R.: Benchmarking detection transfer learning with vision transformers. arXiv preprint arXiv:2111.11429 (2021) 13

38. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017) 13

39. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014) 13

40. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021) 14

41. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016) 9

42. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: European conference on computer vision. pp. 69–84. Springer (2016) 4

43. Van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv e-prints pp. arXiv–1807 (2018) 1, 3

44. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018) 4

45. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2536–2544 (2016) 4

46. Sun, C., Shrivastava, A., Singh, S., Gupta, A.: Revisiting unreasonable effectiveness of data in deep learning era. In: Proceedings of the IEEE international conference on computer vision. pp. 843–852 (2017) 3

47. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. pp. 10347–10357. PMLR (2021) 12

48. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on Machine learning. pp. 1096–1103 (2008) 4

49. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A., Bottou, L.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of machine learning research **11**(12) (2010) 4

50. Wang, R., Chen, D., Wu, Z., Chen, Y., Dai, X., Liu, M., Jiang, Y.G., Zhou, L., Yuan, L.: Bevt: Bert pretraining of video transformers. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2022) (2022) 4

51. Wei, C., Fan, H., Xie, S., Wu, C.Y., Yuille, A., Feichtenhofer, C.: Masked feature prediction for self-supervised visual pre-training. arXiv preprint arXiv:2112.09133 (2021) 4

52. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 418–434 (2018) 13

53. Xie, J., Girshick, R., Farhadi, A.: Unsupervised deep embedding for clustering analysis. In: International conference on machine learning. pp. 478–487. PMLR (2016) 1, 3, 4

54. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. arXiv preprint arXiv:2111.09886 (2021) 4
55. Yang, J., Parikh, D., Batra, D.: Joint unsupervised learning of deep representations and image clusters. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5147–5156 (2016) 1, 3, 4
56. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: European conference on computer vision. pp. 649–666. Springer (2016) 4
57. Zheng, Y., Yang, H., Zhang, T., Bao, J., Chen, D., Huang, Y., Yuan, L., Chen, D., Zeng, M., Wen, F.: General facial representation learning in a visual-linguistic manner. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2022) (2022) 4
58. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017) 13
59. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832 (2021) 4
60. Zhuang, C., Zhai, A.L., Yamins, D.: Local aggregation for unsupervised learning of visual embeddings. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6002–6012 (2019) 1, 3