SdAE: Self-distillated Masked Autoencoder

Yabo Chen[1†], Yuchen Liu[2†], Dongsheng Jiang[3†], Xiaopeng Zhang[3] , Wenrui Dai[1], Hongkai Xiong[2], and Qi Tian[3⋆]

[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2]Department of Electronic Engineering, Shanghai Jiao Tong University
{chenyabo, liuyuchen6666, daiwenrui, xionghongkai}@sjtu.edu.cn
[3]Huawei Cloud EI
dongsheng_jiang@outlook.com, zxphistory@gmail.com, tian.qi1@huawei.com

Abstract. With the development of generative-based self-supervised learning (SSL) approaches like BeiT and MAE, how to learn good representations by masking random patches of the input image and reconstructing the missing information has grown in concern. However, BeiT and PeCo need a "pre-pretraining" stage to produce discrete codebooks for masked patches representing. MAE does not require a pre-training codebook process, but setting pixels as reconstruction targets may introduce an optimization gap between pre-training and downstream tasks that good reconstruction quality may not always lead to the high descriptive capability for the model. Considering the above issues, in this paper, we propose a simple Self-distillated masked AutoEncoder network, namely SdAE. SdAE consists of a student branch using an encoder-decoder structure to reconstruct the missing information, and a teacher branch producing latent representation of masked tokens. We also analyze how to build good views for the teacher branch to produce latent representation from the perspective of information bottleneck. After that, we propose a multi-fold masking strategy to provide multiple masked views with balanced information for boosting the performance, which can also reduce the computational complexity. Our approach generalizes well: with only 300 epochs pre-training, a vanilla ViT-Base model achieves an 84.1% fine-tuning accuracy on ImageNet-1k classification, 48.6 mIOU on ADE20K segmentation, and 48.9 mAP on COCO detection with only 300 epochs pre-training, which surpasses other methods by a considerable margin. Code is available at https://github.com/AbrahamYabo/SdAE.

Keywords: Self-supervised Learning, Masked Image Modeling, Vision Transformer

1 Introduction

The masked language modeling task (MLM) [16] has shown great success in self-supervised learning (SSL) for natural language processing. In computer vision, contrastive learning/instance discrimination [11, 21, 23, 28] is a promising

---

⋆ Correspondence to Qi Tian. †Equal contribution.
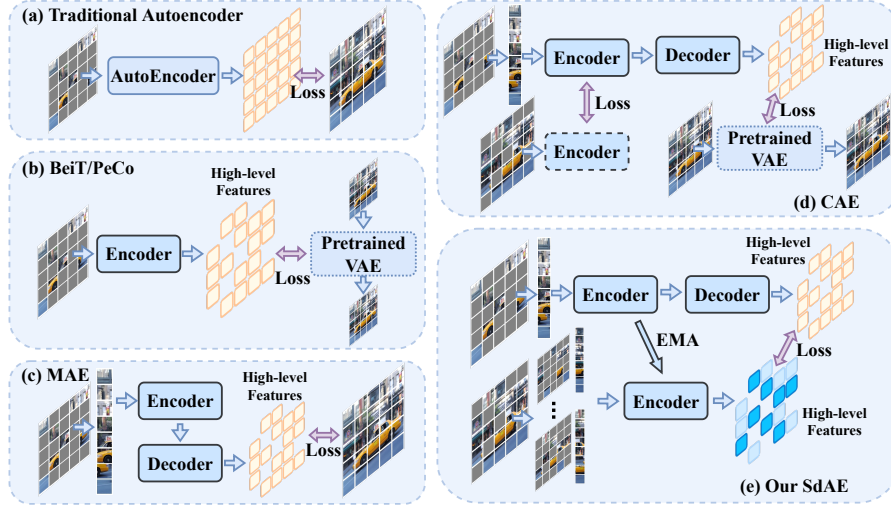
Fig. 1. Comparison of typical generative-based self-supervised learning methods. (a) The inpainting-based methods aim to regress the whole image. (b) BeiT [4] and PeCo [18] model the discrete visual correlation using a pre-trained VAE. They need a pre-trained stage to get the discrete codebooks. (c) MAE [22] mainly models the masked image modeling task as a simply pixel restoration task. (d) Although CAE [13] builds a two-branch structure, it still needs an offline tokenizer. CAE only designs the two-branch to build alignment between masked and unmasked features. (e) Without the need of a pre-trained codebook, our SdAE can build discrete correlations by self-distillating and constructing the reconstruction targets with high-level features, which can eliminate the representation gaps between the actual reconstruction targets. SdAE also proposes a masking strategy to organize the mask tokens for the teacher branch.

direction recently, which regards each instance in the training dataset as a single category. Based on instance discrimination [6, 12, 14], some methods show the effectiveness in many computer vision tasks. With the development of vision transformer [19], inspired by natural language processing (NLP), the generative based self-supervised learning (SSL) methods [3, 4, 13, 18, 22] using masked image modeling (MIM) task have grown in concern. MIM first randomly masks some proportion of image patches, and then recovers the masked patches based on the corrupted image.

BeiT [4] transfers the MIM task into discrete token classification using a pre-trained discrete autoencoder dVAE (DALL-E [30]). PeCO [18] modifies the generating procedure of codebook by enforcing the perceptual similarity during the VAE training. Similarly, these methods rely on a pre-trained feature descriptor to obtain the latent representation of masked tokens. These designs requiring an additional codebook are a kind of "pre-pretraining".

MAE [22] proposes an asymmetric encoder-decoder architecture that can reconstruct the raw pixels from the latent representation. MaskFeat [34] uses the hand-crafted feature descriptor Histograms of Oriented Gradients (HOG)

to tokenize the image features. Although these methods do not need additional codebooks, they employ restoring low-level representations such as pixels for masked image modeling tasks. Nevertheless, restoring low-level representations such as pixels is redundant for high semantic level tasks. Moreover, directly reconstructing pixels may lead to an optimization gap between pre-training and downstream tasks, *i.e.*, good reconstruction quality may not always lead to the high descriptive capability of the model.

Considering the above issue, we propose a simple yet effective self-distillated masked autoencoder structure called SdAE. In SdAE, we claim that MAE itself can produce good representations in an effective and efficient way, and can eliminate the representation gap when used as codebook appropriately. Without needing a codebook in advance nor modeling a low-level representation, SdAE uses a self-distillated teacher-student network to produce the latent representation as reconstruction targets. The student branch consists of the asymmetric encoder-decoder architecture that feeds unmasked images, and the teacher branch contains an encoder to produce latent representation and updates weights from the student using Exponential Moving Average (EMA).

When introducing the teacher branch into the masked autoencoder structure, the easiest way is to feed the full image into the teacher network directly. However, there is no computational loss on the unmasked tokens. Obviously, due to the spatial redundancy that exists in the image, it is not optimal to put the whole image into the teacher branch. In addition, simply putting all masked tokens is still resource-consuming and faced with performance degradation compared with using the whole image as input. So there grows another concern about how to produce better latent representation using the raw images.

We further discuss this problem from the perspective of information bottleneck and propose a multi-fold masking strategy to produce good views for the self-distillated masked autoencoders as well as reduce the computational complexity. The contributions of this paper are summarized as below:

- We propose a novel self-distillated masked autoencoder structure that can construct a learnable high-level reconstruction target rather than extra pre-trained codebooks or low-level pixels, and find that MAE itself can produce a better codebook.
- We discuss how to produce good views for the teacher branch and propose a multi-fold masking strategy to keep mutual information from the teacher branch relevant to the student one. This strategy can also save computation resources.

2   Methods

Firstly, we elaborate the basic framework of masked image modeling and derive the objective function of our masked feature reconstruction methods in Section 2.1. Fig. 2 depicts our proposed self-distillated masked autoencoder that consists of a two-branch network, *i.e.*, teacher and student branches. After that, we present theoretical discussions on how to produce good views for the teacher
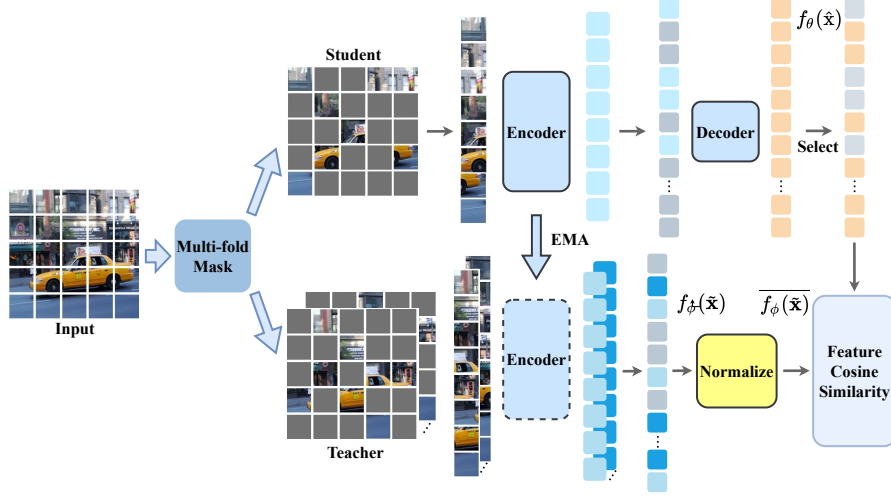
Fig. 2. An overall framework of SdAE. SdAE adopts ViT [19] as the backbone and consists of two branches. The student branch with an asymmetric encoder-decoder reconstructs the missing patches, referring to the teacher branch's latent representation. The teacher branch is simply an exponential moving average (EMA) of the student one. Specifically, all masked tokens are divided into several "small parts" to fit the information bottleneck between two branches for the multi-fold masking strategy.

branch to build latent representation. In addition, we propose a multi-fold masking strategy, which is tailored for balancing the information between the teacher and student branches as in Section 2.2. Finally, in Section 2.3, the distillation strategy for the teacher-student framework is demonstrated.

2.1   Framework

Inspired by Masked Language Modeling (MLM) as the pre-training tasks in BERT [16], Masked Image Modeling (MIM) has been proposed in several recent works [4, 22, 36]. Specifically, an image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ is reshaped and tokenized into a sequence of flattened 2D patches as $\mathbf{x} = \{\mathbf{x}_i\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^{N \times (P^2 \cdot C)}$ and (P, P) is the resolution of each image patch. MIM firstly samples a random binary mask $\mathbf{M} \in \{0, 1\}^N$ to be $\{m^1, m^2, \cdots, m^N\}$, where $m^i \in \{0, 1\}$ and N is the number of tokens. The patch token $\mathbf{x}_i$ with $m^i = 1$ is masked and we denote the mask ratio as $\mathrm{r} = \frac{\sum_{i=1}^{N} m^i}{N} \in (0, 1)$. All the masked tokens $\tilde{\mathbf{x}} \triangleq \{\mathbf{x}_i \mid m^i = 1\}_{i=1}^{N}$ are discarded for efficiency, resulting in a corrupted image $\hat{\mathbf{x}} \triangleq \{\hat{\mathbf{x}}_i \mid (1 - m^i) \, \mathbf{x}_i\}_{i=1}^{N}$. In summarize, we have $\tilde{\mathbf{x}} \in \mathbb{R}^{N \cdot \mathrm{r} \times (P^2 \cdot C)}$ and $\hat{\mathbf{x}} \in \mathbb{R}^{N \cdot (1 - \mathrm{r}) \times (P^2 \cdot C)}$. The objective of MIM is to recover the masked tokens $\tilde{\mathbf{x}}$ by using the corrupted image $\hat{\mathbf{x}}$ as $\log q_\psi(\tilde{\mathbf{x}}|\hat{\mathbf{x}})$. Assuming the reconstruction of each masked token is independent to each other [4, 39], the objective can be

reformulated as:

$$\log q_\psi(\tilde{\mathbf{x}}|\hat{\mathbf{x}}) \approx \sum_{i=1}^{N} m^i \cdot \log q_\psi(\mathbf{x}_i|\hat{\mathbf{x}}) \tag{1}$$

Considering MIM is essentially a reconstruction task. It is related more to regression tasks than to classification tasks. Thus, we assume the noise in the deviation of reconstructed and ground truth values follow the standard Gaussian distribution as $\eta \sim N(0,1)$, and the objective is to minimize

$$\sum_{i=1}^{N} m^i \cdot \log q_\psi(\mathbf{x}_i|\hat{\mathbf{x}}) = \sum_{i=1}^{N} m^i \cdot \log e^{\frac{-(f_\phi(\mathbf{x}_i)-f_\theta(\hat{\mathbf{x}}))^2}{2}} = \sum_{i=1}^{N} -\frac{1}{2}m^i(f_\phi(\mathbf{x}_i)-f_\theta(\hat{\mathbf{x}}))^2. \tag{2}$$

Specifically, for MAE [22], f_ϕ is the identity function, and f_θ is the masked autoencoder that reconstructs masked patches in the pixel spatial space. However, we argue that (1) The f_ϕ is a fixed identity function without adaptation during pre-training that undermines the effectiveness of the self-supervised training; (2) Memorying each pixel of the images by reconstruction in the low semantic level space is sub-optimal and less efficient for capturing representations for high-level tasks. There exists an optimization direction gap in that the quality of the reconstruction may not always increase the descriptive capability of the model.

Correspondingly, we introduce a self-distillation architecture, *i.e.*, f_θ is the student network trained by gradient descent using $\hat{\mathbf{x}}$ as inputs and f_ϕ is the teacher network updated by exponential moving average (EMA) from the parameters of the student network. The teacher network use $\tilde{\mathbf{x}}$ as inputs and conduct the missing information reconstruction in the high-level latent representation space rather than in the pixel space.

Inspired by MAE [22], we propose a value normalization function upon teacher outputs. Specifically, we compute the mean and standard deviation of feature values within a patch and use them to normalize the teacher outputs as

$$\overline{f_\phi(\mathbf{x}_i)} = \frac{f_\phi(\mathbf{x}_i) - mean(f_\phi(\mathbf{x}_i))}{\sqrt{var(f_\phi(\mathbf{x}_i)) + \epsilon}} \tag{3}$$

where ϵ is a small value to prevent the denominator from being 0. We find that using normalized features as the reconstruction target improves the representation quality.

Then we minimize the normalized teacher features with the output features of the student decoder based on the feature cosine similarity, and Eq. (2) is reformulated as:

$$\log q_\psi(\tilde{\mathbf{x}}|\hat{\mathbf{x}}) \approx \frac{\sum_{i=1}^{n} m^i \overline{f_\phi(\mathbf{x}_i)} f_\theta(\hat{\mathbf{x}})}{\sqrt{\sum_{i=1}^{n} m^i \left(\overline{f_\phi(\mathbf{x}_i)}\right)^2} \sqrt{\sum_{i=1}^{n} m^i \left(f_\theta(\hat{\mathbf{x}})\right)^2}} \tag{4}$$

2.2   Discussions on The Teacher Branch Feeding

As mentioned in MAE [22], languages are human-generated signals that are highly semantic and information-dense. Predicting a few missing words per sen-

tence can induce sophisticated language understanding, but images are natural signals with heavy spatial redundancy. Directly feeding the whole image into the teacher network to produce features may be sub-optimal.

Before discussing the better feeding of the teacher branch, we provide a brief introduction here to show alternatives to the teacher branch feeding. As shown in Fig. 3 (a), MAE [22] and MaskFeat [34] take the whole image as input to produce normalized tokens (normalized raw image pixels) or HOG from the whole images as the reconstruction target. Correspondingly, another simply modeling way is to feed the whole image into the teacher network directly as Fig. 3 (b). However, since we have mentioned that the reconstruction loss only computes for masked tokens, there is no need to feed all the patches, and visible patches may be spatial redundant. Only feeding the masked patches $\hat{\mathbf{x}}$ into the network can relieve the spatial redundancy to some extent. Furthermore, to reduce more spatial redundancy and save the computational resource, we can mask some of the target tokens and feed the remaining tokens into the teacher. We called this operation as teacher crop. In the experiment, we can achieve comparable performance between teacher crop and reconstruct all masked patches. As shown in Fig. 3 (c). We formulate this teacher crop masking as another random binary mask $\mathbf{M^t} \in \{0,1\}^{N \cdot \mathrm{r}}$ to be $\{m_t^1, m_t^2, \cdots, m_t^{N \cdot \mathrm{r}}\}$, where $m_t^i \in \{0,1\}$ and $N \cdot \mathrm{r}$ is the number of total target tokens. Similarly, we formulate the teacher crop mask ratio as $\mathrm{r}_c \in (0,1)$. Based on this binary mask, we can randomly select a group of target tokens as: $\tilde{\mathbf{x}}^* \triangleq \{\tilde{\mathbf{x}}_i \mid m_t^i = 1\}_{i=1}^{N \cdot \mathrm{r}}$, and the latent representation is denoted as $f_\theta(\tilde{\mathbf{x}}^*)$.

Multi-fold Masking Strategy. Furthermore, we propose a multi-fold masking strategy that groups several teacher crops to sufficiently use masked tokens. As shown in Fig. 3 (d), given a number of fold t, all the masked tokens are divided into t groups of teacher crops without overlap as $\tilde{\mathbf{x}} = \{\tilde{\mathbf{x}}^{*1}, \tilde{\mathbf{x}}^{*2}, ..., \tilde{\mathbf{x}}^{*t}\}$. Each fold containing masked tokens will be fed into a shared teacher network independently. The outputs of the teacher network will be rearranged together to calculate the cosine similarity loss in a parallel way. Specifically, the output of the teacher branch can be reformulated as: $f_\theta(\tilde{\mathbf{x}}) = \{f_\theta(\tilde{\mathbf{x}}^{*1}), f_\theta(\tilde{\mathbf{x}}^{*2}), ..., f_\theta(\tilde{\mathbf{x}}^{*t})\}$. As shown in Table 1, from the view of computational cost, the main difference between multi-fold masking strategy and all masked patches feeding is that in the teacher branch, the self-attention is only calculated in each teacher crop of tokens instead of all masked tokens. So the multi-mask strategy can save complexities by a factor of t. Detailed theoretical complexity can be seen in Table 1.

In practice, directly feeding the whole image into the teacher network will increase almost half of the computation costs compared with MAE, as shown in Table 1. Using all masked patches as input like CAE [13] will also increase 38.1% pre-training time. Using multi-fold masking strategy, we can save the extra pre-training time costs by only 28.1%.

The main difference between our multi-fold masking and previous work multi-crop [6] is discussed as follows. The multi-crop strategy needs to train multiple random views with different sizes without concern of complexity. By contrast, multi-fold masking just rearranges the mask tokens and does not create new
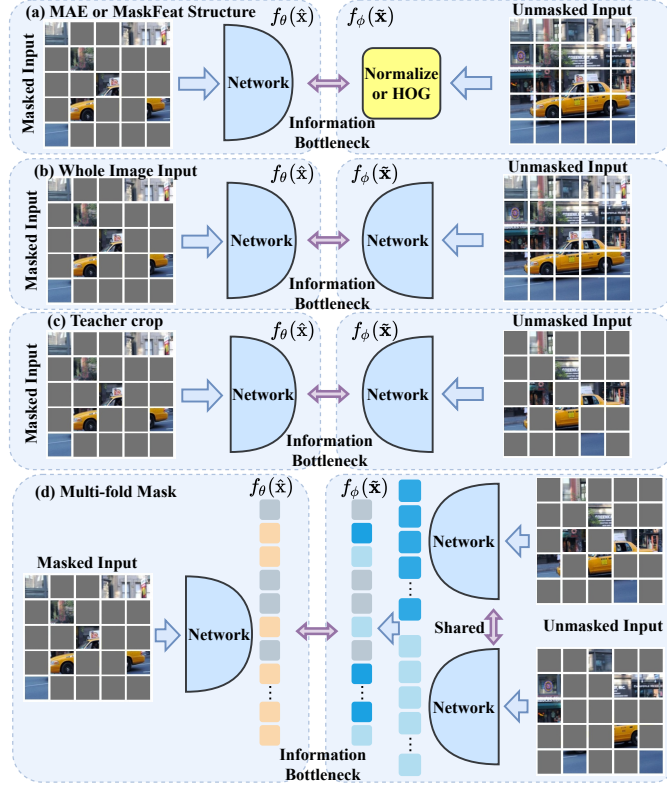
Fig. 3. (a) MAE [22] and MaskFeat [34] reconstruct the low-level representations. (b) Feeding the whole image into the teacher network is the simplest way. (c) We experiment with teacher-crop reconstructing only parts of tokens. (d) Multi-fold masking divides unmasked tokens into independent groups and feeds them into shared networks.

views. Multi-fold masking can also save training complexity by calculating self-attention on only a small group of tokens.

Information Bottleneck and Multi-fold masking In this section, we will follow the assumption from the views of information bottleneck theory in [1][33] to discuss the benefits of multi-fold masking except for saving costs.

Denote $\tilde{\mathbf{x}}$ and $\hat{\mathbf{x}}$ as the input of the teacher and student network, respectively. From the perspective of information theory, maximizing Eq. (2) is equivalent to maximizing the mutual information $I(\tilde{\mathbf{x}}, \hat{\mathbf{x}})$ as

$$I(\tilde{\mathbf{x}}, \hat{\mathbf{x}}) = H(f_\phi(\tilde{\mathbf{x}})) - H(f_\phi(\tilde{\mathbf{x}})|f_\theta(\hat{\mathbf{x}})) = H(f_\phi(\tilde{\mathbf{x}})) + \mathbb{E}_{\tilde{\mathbf{x}},\hat{\mathbf{x}}}[\log p(f_\phi(\tilde{\mathbf{x}})|f_\theta(\hat{\mathbf{x}}))] \quad (5)$$

Since the true distribution of $p(\tilde{\mathbf{x}}|\hat{\mathbf{x}})$ is unknown, a variational distribution $q(\tilde{\mathbf{x}}|\hat{\mathbf{x}})$ is employed for approximation, and Eq. (5) can be written as:

$$I(\tilde{\mathbf{x}}, \hat{\mathbf{x}}) = H(f_\phi(\tilde{\mathbf{x}})) + \mathbb{E}_{\tilde{\mathbf{x}},\hat{\mathbf{x}}}[\log q(f_\phi(\tilde{\mathbf{x}})|f_\theta(\hat{\mathbf{x}}))] + \mathbb{E}_{\hat{\mathbf{x}}}[KL(p(f_\phi(\tilde{\mathbf{x}})|f_\theta(\hat{\mathbf{x}}))||q(f_\phi(\tilde{\mathbf{x}})|f_\theta(\hat{\mathbf{x}}))]$$
$$(6)$$

Table 1. Alternatives for the input of the teacher branch network. The multi-mask strategy captures content only within each fold at a reduced memory cost compared with all masked patches fed. Comparatively, using a multi-mask strategy reduces complexities by a factor of t. n: input length, d: dimension of channels, r: mask ratio, r_c the teacher crop mask ratio, t: number of fold. The training time is the computational costs for per epoch with NVIDIA V100 GPUs and the memory cost is the maximum usage on per GPU.

Operation	Theoretical complexity	Training Time	GPU Memory
MAE(baseline)	-	551s	14.93G
Full image input	$\Theta(n^2 d)$	826s$(+49.9\%)$	16.24G
Only masked patches input	$\Theta(n^2 r^2 d)$	761s$(+38.1\%)$	16.01G
Teacher crop input	$\Theta(n^2 r_c^2 r^2 d)$	616s$(+11.8\%)$	15.54G
Multi-fold masking strategy	$\Theta((n/t)^2 t r^2 d)$	706s$(+28.1\%)$	16.02G

where $f_\phi(\tilde{\mathbf{x}})$ and $f_\theta(\hat{\mathbf{x}})$ are the representations from the teacher and student networks, and $H(\cdot)$ is the entropy function. Since $KL(\cdot)$ function is non-negative and $H(\tilde{\mathbf{x}})$ is constant w.r.t the parameters to be optimized, maximizing $I(\tilde{\mathbf{x}}, \hat{\mathbf{x}})$ is equivalent to maximizing $\mathbb{E}_{\tilde{\mathbf{x}}, \hat{\mathbf{x}}}[\log q(f_\phi(\tilde{\mathbf{x}})|f_\theta(\hat{\mathbf{x}}))]$ with q modeled by a Gaussian distribution, which is in consistent with Eq. (2). Thus, the objective of MIM can be reinterpreted as $I(\tilde{\mathbf{x}}, \hat{\mathbf{x}})$ in the information theory view.

Definition 1. *(Sufficient Encoder) The encoder f_θ of $\hat{\mathbf{x}}$ is sufficient in the self-supervised learning framework if and only if $|I(\tilde{\mathbf{x}}, \hat{\mathbf{x}}) - I(\tilde{\mathbf{x}}, f_\theta(\hat{\mathbf{x}}))| < \epsilon$.*

Ideally, the encoder f_θ is sufficient if the amount of information in $\hat{\mathbf{x}}$ about $\tilde{\mathbf{x}}$ is lossless during the encoding procedure. In other words, $f_\theta(\hat{\mathbf{x}})$ has kept all the information that the reconstruction objective requires. Symmetrically, f_ϕ is sufficient if $|I(\tilde{\mathbf{x}}, \hat{\mathbf{x}}) - I(\tilde{\mathbf{x}}, f_\theta(\hat{\mathbf{x}}))| < \epsilon$.

Definition 2. *(Minimal Sufficient Encoder) A sufficient encoder f_θ of $\hat{\mathbf{x}}$ is minimal if and only if $I(f_\theta(\hat{\mathbf{x}}), \hat{\mathbf{x}}) \leq I(f(\hat{\mathbf{x}}), \hat{\mathbf{x}})$, $\forall f$ that is sufficient.*

Among these sufficient encoders, the minimal ones only keep relevant information of the reconstruction task and throw away other irrelevant information.

Proposition 1. *Suppose f_θ and f_ϕ are minimal sufficient encoders. Given a downstream task $\mathcal{T}$ with label $\mathbf{y}$, the optimal views created from the data $\mathbf{x}$ are $(\tilde{\mathbf{x}}^*, \hat{\mathbf{x}}^*) = \arg\min_{\tilde{\mathbf{x}}, \hat{\mathbf{x}}} I(\tilde{\mathbf{x}}, \hat{\mathbf{x}})$, subject to $|I(\tilde{\mathbf{x}}, \mathbf{y}) - I(\hat{\mathbf{x}}, \mathbf{y})| < \epsilon$.*

Specifically, based on Proposition 1, we can model the multi-fold masking tokens as multi-views from reconstructed images, which can be formulated as $\{\tilde{\mathbf{x}}^{*1}, \tilde{\mathbf{x}}^{*2}, ..., \tilde{\mathbf{x}}^{*t}\}$ and the unmasked tokens is another view of input images $\hat{\mathbf{x}}$. Then, we draw insights for views that feed into the teacher and student networks: (1) Ours method satisfies with $\arg\min_{\tilde{\mathbf{x}}, \hat{\mathbf{x}}} I(\tilde{\mathbf{x}}, \hat{\mathbf{x}})$, since the feeding of the teacher and student networks do not contains overlap tokens, reducing the shared mutual information. (2) Due to the need to satisfy $|I(\tilde{\mathbf{x}}, \mathbf{y}) - I(\hat{\mathbf{x}}, \mathbf{y})| < \epsilon$, the amount of mutual information between $\tilde{\mathbf{x}}$ and $\hat{\mathbf{x}}$ should be comparable. We propose a multi-fold masking strategy to divide the masked patches into groups to make the number of masked patches feeding into the teacher network equal with that feeding into the student. Thus, their shared mutual information with

the downstream task is comparable. As shown in Fig. 4 (b), we experimentally verify this proposition that when the number of masked patches from each fold is equal with the unmasked patches (25% of total image patches in practice), our SdAE achieves the best performance. (3) Considering the objective function of MIM tasks is $\max I(\tilde{\mathbf{x}}, \mathbf{y})$, we also need to $\max I(\hat{\mathbf{x}}, \mathbf{y})$, so we try to utilize almost all masked tokens to keep as more task related information as possible.

2.3   Distillation Strategy

Considering a teacher network, we hypothesize that it is desirable to build the reconstruction target representations in the (i) online and (ii) consistent way. Now that the student network f_θ is trained by back-propagation to minimize the feature reconstruction loss. The teacher network is updated in a momentum update way using exponential moving average (EMA). Specifically, denoting the parameters of f_ϕ as ϕ and those of f_θ as θ, we update ϕ by:

$$\phi \leftarrow \eta \cdot \phi + (1 - \eta) \cdot \theta \tag{7}$$

Here $\eta \in [0, 1)$ is a momentum coefficient to control the frequency of updates from the student model. The codebook should not update too often. Otherwise, the model may fail to converge.

3   Experiments

This section evaluates our pre-trained feature representation on several unsupervised benchmarks. We first evaluate the classification performance on ImageNet-1k under fine-tuning and linear probing. Then we transfer the pre-trained features to several downstream tasks, *i.e.*, semantic segmentation and object detection. Finally, we conducted an ablation study on the key components of SdAE.

3.1   Fine-tuning on ImageNet-1k

We study the fine-tuning on the ILSVRC-2012 ImageNet dataset [31] with 1k classes and 1.3M images. For a fair comparison, we directly follow most of the hyperparameters of MAE [22] in our fine-tuning experiments. All experiments reported are only fine-tuning for 100 epochs (vs. 300 training from scratch). We compare our SdAE with Vision Transformers trained by random initialization and previous self-supervised learning methods. As shown in Table 2, compared with the models trained by random initialization which only achieves 81.8% top-1 accuracy with ViT-B, our SdAE achieves 84.1%, demonstrating the effectiveness of pre-training with unlabeled data.

Compared with previous self-supervised methods, our proposed SdAE surpasses them on ImageNet fine-tuning by a large margin. For ViT-B, our SdAE outperforms MAE by 1.2% top-1 accuracy with the same number of training epochs, demonstrating that MIM on high-level latent feature space is more effective than low-level pixel space. Besides, our SdAE outperforms BEiT by 1.1%

Table 2. Image classification results on the ILSVRC-2012 ImageNet dataset with top 1 accuracy. "Epochs" refers to the number of pre-training epochs. MoCo v3 and DINO adopt multi-crop augmentation for pre-training. MoCo v3: 2 global crops of 224×224. DINO: 2 global crops of 224×224 and 10 local crops of 96×96.

Method	Epochs	Crops	Finetune	Linear
Methods using ViT-B:				
Train from Scratch	300	–	81.8	–
MoCo v3	300	2	83.2	76.2
DINO	400	12	83.3	**77.3**
BEiT	300	1	83.0	49.4
MAE	100	1	82.1	54.8
MAE	300	1	82.9	61.5
MAE	1600	1	83.6	67.8
CAE	300	1	83.3	64.2
SdAE	100	1	83.5	60.3
SdAE	300	1	**84.1**	64.9

top-1 accuracy. Moreover, compared to the recently proposed CAE, our SdAE achieves 0.8% top-1 accuracy gain, demonstrating the effectiveness of our self-distillated design and multi-fold masking strategy. In addition, with only 100 epochs pre-training, SdAE can achieve comparable performance with MAE using 1600 epochs pre-training and surpass 300 epochs pre-trained CAE. Our proposed SdAE also surpasses above methods on ImageNet linear probing with the same training epochs. As a MIM based method, SdAE can also surpasses MIM based methods with the same pre-training epochs. The phenomenon that contrastive based methods surpass the MIM based ones on linear probing is also discussed in MAE [22] and CAE [13]. In terms of linear probing, contrastive learning mainly cares about the 1000 classes and MIM methods may care about the classes beyond the 1000 classes. So fine-tuning measurement may better validate the effectiveness of MIM based methods.

3.2   Semantic Segmentation

We evaluate the learned representation of our SdAE on the ADE20K benchmark [38] with 25K images and 150 semantic categories. The mean Intersection of Union (mIoU) averaged over all semantic categories is reported as the evaluation metric. Table 3 shows that our SdAE achieves the state-of-the-art performance with 48.6 mIoU for 300 pre-training epochs. Our SdAE outperforms BEiT, MAE, and CAE by 3.1, 2.8, and 1.9 mIoU, respectively. Besides, our SdAE with 300 pre-training epochs even outperforms MAE with 1600 pre-training epochs.

3.3   Object Detection

Following CAE [13], we fine-tune Mask R-CNN [25] in an end-to-end manner on COCO [27]. The ViT backbone is adapted for use with FPN [26]. The box

Table 3. Semantic segmentation on ADE20K. All methods use ViT-B backbone based on the same implementation. "Epochs" refers to the number of pre-training epochs.

Method	Epochs	Crops	Supervised	Self-supervised	mIoU
DeiT	300	—	✓	✗	47.0
MoCo v3	300	2	✗	✓	47.2
DINO	400	12	✗	✓	47.2
BEiT	300	1	✗	✓	45.5
BEiT	800	1	✗	✓	46.5
MAE	300	1	✗	✓	45.8
MAE	1600	1	✗	✓	48.1
CAE	300	1	✗	✓	47.7
SdAE	300	1	✗	✓	**48.6**

Table 4. Object detection and instance segmentation on COCO. All the results are based on the same implementations. Mask R-CNN is adopted and trained with the $1\times$ schedule. "Epochs" refers to the number of pre-training epochs on ImageNet-1K. "Sup" and "Self-sup" refer to the methods that are supervised or self-supervised.

Methods	Epochs	Sup	Self-sup	Object Detection			Instance Segmentation		
				AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
Methods using ViT-B:									
DeiT	300	✓	✗	46.9	68.9	51.0	41.5	65.5	44.4
MoCo v3	300	✗	✓	45.5	67.1	49.4	40.5	63.7	43.4
DINO	400	✗	✓	46.8	68.6	50.9	41.5	65.3	44.5
BEiT	300	✗	✓	39.5	60.6	43.0	35.9	57.7	38.5
BEiT	800	✗	✓	42.1	63.3	46.0	37.8	60.1	40.6
MAE	300	✗	✓	45.4	66.4	49.6	40.6	63.4	43.7
MAE	1600	✗	✓	48.4	69.4	53.1	42.6	66.1	45.9
CAE	300	✗	✓	48.0	68.7	52.7	42.3	65.6	45.4
SdAE	300	✗	✓	**48.9**	69.6	53.3	**43.0**	66.2	46.2

AP for object detection and the mask AP for instance segmentation is reported in Table 4. Our method (300 epochs, ViT-B) is consistently superior to all the other models. Our SdAE performs better than the recent published CAE (48.9 *vs.* 48.0 AP^b). Besides, it is worth mentioning that our SdAE (300 epochs) even outperforms MAE (1600 epochs) by 0.5 AP^b. As an effective framework for self-supervised learning, we achieve better performance with fewer training epochs.

4   Ablation Studies

In this section, we present ablation studies to better evaluate the contributions of each component and hyperparameter settings in our proposed SdAE. Unless specified, all results are compared with models pre-trained for 100 epochs for efficiency, and we report the top-1 accuracy after fine-tuning for 100 epochs.

Table 5. Ablation studies on each component. "Full image" refers to the teacher network taking the whole image as input. "Only masked" refers to only the masked patches that are fed into the teacher network. "Multi fold" indicates our multi-fold masking strategy. Teacher normalization refers to whether adding feature values normalization within patches or not.

Full image	Only masked	Multi fold	Teacher normalize	Accuracy
✓	✗	✗	✗	83.7
✓	✗	✗	✓	84.0
✗	✓	✗	✓	83.5
✗	✗	✓	✓	**84.1**

### 4.1   Ablation Studies on Each Component

In this subsection, we present ablation studies on each component. Table 5 shows that our proposed teacher normalization achieves 0.3% top-1 accuracy gain. Only inputting the masked tokens into the teacher network suffers from 0.5% performance degradation due to insufficient information exploration. While using our proposed multi-fold masking strategy, we can achieve 0.6% improvement compared with only masked token inputs. Besides, multi-fold masking even outperforms taking full image as inputs more efficiently.

4.2   The EMA Strategy

This experiment is conducted without the multi-fold masking strategy to evaluate the raw performance of the EMA strategy. Specifically, we have two settings for the EMA strategy: (1) update the parameters of the teacher branch with EMA for each training iteration. (2) update the parameters of the teacher branch with EMA for each training epoch. As shown in Fig. 4 (a), conducting the EMA strategy to update the teacher branch each training iteration is extremely sensitive to the value of the momentum coefficient. Specifically, only changing the value by 0.001 results in the sharply degraded performance. In contrast, the EMA strategy to update the teacher branch per batch is more robust to the momentum coefficient. Besides, conducting the EMA strategy per epoch can better benefit from long training epochs.

4.3   The Multi-fold Masking Strategy

We further conduct an ablation study on the multi-fold masking strategy. Firstly, we study experiments on the teacher crop that mask some of the target tokens and feed the remaining tokens into the teacher. The whole image is divided into 196 image patches with the size of 16×16, and we randomly sample a different number of total patches, *e.g.*, 36, 49, 79, 122, 147 and 196 as whole image input into the teacher network. As shown in Fig. 4 (b), for the case that 36 image
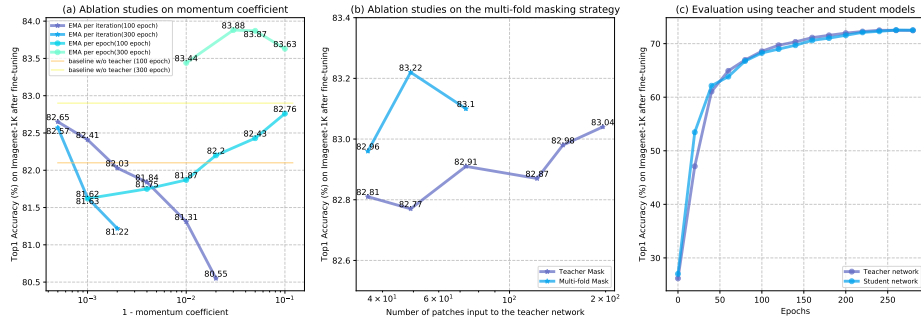
Fig. 4. (a) Ablation studies on the EMA update strategy for the coefficient of the momentum teacher. (b)Ablation studies on the multi-fold masking strategy. (c) Ablation studies on the evaluation of teacher and student models.

patches (roughly 18% of the whole image) are input to the teacher network, our method can achieve 82.81% top-1 accuracy. 83.04% top-1 accuracy is achieved when we take the whole image as input. Comparably, only 0.23% performance gain is achieved when five times as many image patches are input. The problem of spatial redundancy is also common in the input views of the teacher network.

Correspondingly, we take the multi-fold masking strategy. Apart from the 49 patches that are input to the student network, the left 147 image patches are divided into 2 fold as 2×79 patches, 3 fold as 3×49 patches or 4 fold as 4×36 patches. As shown in Fig. 4 (b), a multi-fold masking strategy can consistently improve the performance. For 3 fold with 49 patches, our method achieves the best performance. This experiment also proves our information bottleneck discussion, that with comparable mutual information between each fold and the student input, we can get the best performance.

4.4 Evaluation of Teacher and Student Models

As shown in Fig. 4 (c), for every 20 epochs of our training, we finetune our pretrained teacher and student models for 5 epochs on ImageNet-1K. The student model performs better than the teacher model in the initial epochs when the learning rate is relatively low due to warm up and then is very close to the teacher when the learning rate increases. We demonstrate that the teacher model and student model trained by our SdAE achieve similar performance. That proves the teacher branch can really learn high semantic level related representations. The teacher model can be seen as an ensemble over many student models. We follow most previous works to use the student model as the final model.

5 Related Work

Previous self-supervised methods utilize different priors of images to design clever pretext tasks, such as predicting the patch positions [17], inpainting [29], col-

orization [37], and rotation prediction [20]. Recent progress in self-supervised learning focuses on contrastive learning and masked image modeling.

Masked Image Modeling. Motivated by BERT [16] for masked language modeling(MLM), masked image modeling(MIM) is proposed to learn representations from images corrupted by masking. Recent works leverage the powerful Vision Transformer, which matches the masked image modeling task. iGPT [9] operates on sequences of pixels and predicts unknown pixels. Vision transformer [19] studies masked average value prediction for self-supervised learning. BEiT [4] proposes to predict discrete tokens based on a pre-trained image tokenizer while iBOT [39] proposes an online tokenizer. MAE [22] proposes a masked autoencoder for reconstructing the image pixels. CAE [13] provides a two-branch network for MIM, and the masked features are modeled as a regularization to align the mask and unmasked features.

Contrastive Learning. Contrastive learning is proposed based on the Info-Max principle, which aims at maximizing the mutual information [35] across different augmentations of the same image [10, 32]. This augmentation invariance is achieved by enforcing the similarity over different views of the same image while avoiding model collapse. Model collapse can be avoided by introducing negative samples for noise-contrastive estimation [12, 10, 24, 35]. The models typically regard various data augmentations as different views of an image and then make the representations of positive pairs similar while pushing negative pairs away. Large memory banks or large batch size is leveraged to obtain more informative negative samples. BYOL [21] and SimSiam [15] employ an asymmetric network and eliminate the requirement of negative samples. Other methods use clustering [2, 5, 7, 8] to organize image examples.

6   Conclusion

In this paper, we propose a novel Self-distillated Masked Autoencoder for masked feature reconstruction, namely SdAE. We first formulate the framework of the masked image modeling task, based on which we analyze that existing methods are sub-optimal due to the need of pre-trained codebooks or just reconstructing the low-level pixels. We propose a self-distillated framework for reconstruction in the high-level feature space. Besides, we analyze how to build good views for the teacher branch to produce latent representation from the perspective of information bottleneck theory and propose a multi-fold masking strategy that can also relieve the spatial redundancy. Experimentally, based on a vanilla ViT-Base model, our SdAE achieves a new state-of-the-art of 84.1% top-1 accuracy with only 300 epochs pre-training.

References

1. Alemi, A.A., Fischer, I., Dillon, J.V., Murphy, K.: Deep variational information bottleneck. arXiv: Learning (2016)
2. Asano, Y.M., Rupprecht, C., Vedaldi, A.: Self-labelling via simultaneous clustering and representation learning. arXiv preprint arXiv:1911.05371 (2019)
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. Bao, H., Dong, L., Wei, F.: Beit: Bert pre-training of image transformers. arXiv: Computer Vision and Pattern Recognition (2021)
5. 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)
6. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: Neural Information Processing Systems (2020)
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)
8. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9650–9660 (2021)
9. 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)
10. 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)
11. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E.: A simple framework for contrastive learning of visual representations. arXiv: Learning (2020)
12. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. arXiv: Learning (2020)
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 (2022)
14. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv: Computer Vision and Pattern Recognition (2020)
15. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15750–15758 (2021)
16. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: North American Chapter of the Association for Computational Linguistics (2018)
17. 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)
18. 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 (2022)

19. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
20. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728 (2018)
21. 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., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent: A new approach to self-supervised learning. arXiv: Learning (2020)
22. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. arXiv preprint arXiv:2111.06377 (2021)
23. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Computer Vision and Pattern Recognition (2020)
24. 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)
25. 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)
26. 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)
27. 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)
28. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv: Learning (2018)
29. 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)
30. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. arXiv: Computer Vision and Pattern Recognition (2021)
31. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
32. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. In: European conference on computer vision. pp. 776–794. Springer (2020)
33. Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., Isola, P.: What makes for good views for contrastive learning? Advances in Neural Information Processing Systems **33**, 6827–6839 (2020)
34. 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)
35. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3733–3742 (2018)
36. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

37. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: European conference on computer vision. pp. 649–666. Springer (2016)
38. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. International Journal of Computer Vision **127**(3), 302–321 (2019)
39. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer (2022)