Understanding Masked Image Modeling via Learning Occlusion Invariant Feature

Anonymous authors
Paper under double-blind review

Abstract

Recently, Masked Image Modeling (MIM) achieves great success in self-supervised visual recognition. However, as a reconstruction-based framework, it is still an open question to understand how MIM works, since MIM appears very different from previous well-studied siamese approaches such as contrastive learning. In this paper, we propose a new viewpoint: MIM implicitly learns occlusion-invariant features, which is analogous to other siamese methods while the latter learns other invariance. By relaxing MIM formulation into an equivalent siamese form, MIM methods can be interpreted in a unified framework with conventional methods, among which only a) data transformations, i.e. what invariance to learn, and b) similarity measurements are different. Furthermore, taking MAE (He et al., 2021) as a representative example of MIM, we empirically find the success of MIM models relates a little to the choice of similarity functions, but the learned occlusion invariant feature introduced by masked image – it turns out to be a favored initialization for vision transformers, even though the learned feature could be less semantic. We hope our findings could inspire researchers to develop more powerful self-supervised methods in computer vision community.

1 Introduction

Invariance matters in science (Kosmann-Schwarzbach, 2011). In self-supervised learning, invariance is particularly important: since ground truth labels are not provided, one could expect the favored learned feature to be invariant (or more generally, equivariant (Dangovski et al., 2021)) to a certain group of transformations on the inputs. Recent years, in visual recognition one of the most successful self-supervised frameworks – *contrastive learning* (Oord et al., 2018; Tian et al., 2019; Dosovitskiy et al., 2014a) – benefits a lot from *learning invariance*. The key insight of contrastive learning is, because recognition results are typically insensitive to the deformations (e.g. cropping, resizing, color jittering) on the input images, a good feature should also be invariant to the transformations. Therefore, contrastive learning suggests minimizing the distance between two (or more (Caron et al., 2021)) feature maps from the augmented copies of the same data, which is formulated as follows:

$$\min_{\theta} \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \mathcal{M}\left(z_1, z_2\right), \quad z_1 = f_\theta(\mathcal{T}_1(x)), \quad z_2 = f_\theta(\mathcal{T}_2(x)), \tag{1}$$

where $\mathcal{D}$ is the data distribution; $f_\theta(\cdot)$ means the *encoder network* parameterized by θ; $\mathcal{T}_1(\cdot)$ and $\mathcal{T}_2(\cdot)$ are two transformations on the input data, which defines what invariance to learn; $\mathcal{M}(\cdot,\cdot)$ is the *distance function*[*] (or *similarity measurement*) to measure the similarity between two feature maps z_1 and z_2. Clearly, the choices of $\mathcal{T}$ and $\mathcal{M}$ are essential in contrastive learning algorithms. Researchers have come up with a variety of alternatives. For example, for the transformation $\mathcal{T}$, popular methods include random cropping (Bachman et al., 2019; He et al., 2020; Chen et al., 2020b; Grill et al., 2020), color jittering (Chen et al., 2020b), rotation (Reed et al., 2020; Gidaris et al., 2018a), jigsaw puzzle (Noroozi & Favaro, 2016), colorization (Zhang et al., 2016) and etc. For the similarity measurement $\mathcal{M}$, *InfoMax principle* (Bachman et al., 2019) (which can be implemented with *MINE* (Belghazi et al., 2018) or *InfoNCE loss* (Oord et al., 2018; He et al., 2020; Chen et al., 2020b;c)), feature de-correlation (Zbontar et al., 2021; Bardes et al., 2021), asymmetric teacher (Grill et al., 2020; Chen & He, 2020), *triplet loss* (Li et al., 2021a) and *etc.*, are proposed.

[*] Following the viewpoint in (Chen & He, 2020), we suppose distance functions could contain parameters which are jointly optimized with Eq. 1. For example, weights in *project head* (Chen et al., 2020b) or *predict head* (Grill et al., 2020; Chen & He, 2020) are regarded as a part of distance function $\mathcal{M}(\cdot)$.

Apart from contrastive learning, very recently *Masked Image Modeling* (*MIM*, *e.g.* Bao et al. (2021)) quickly becomes a new trend in visual self-supervised learning. Inspired by *Masked Language Modeling* (Devlin et al., 2018) in *Natural Language Processing*, MIM learns feature via a form of *denoising autoencoder* (Vincent et al., 2008): images which are occluded with random *patch masks* are fed into the encoder, then the decoder predicts the original embeddings of the masked patches:

$$\min_{\theta,\phi} \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \mathcal{M}\left(d_\phi(z), x \odot (1-M)\right), \quad z = f_\theta(x \odot M), \tag{2}$$

where "$\odot$" means element-wise product; M is *patch mask* [†]; $f_\theta(\cdot)$ and $d_\phi(\cdot)$ are *encoder* and *decoder* respectively; z is the learned representation; $\mathcal{M}(\cdot,\cdot)$ is the *similarity measurement*, which varies in different works, *e.g.* $l2$-distance (He et al., 2021), *cross-entropy* (Bao et al., 2021) or *perceptual loss* (Dong et al., 2021) in *codebook space*. Compared with conventional contrastive methods, MIM requires fewer effort on tuning the augmentations, furthermore, achieves outstanding performances especially in combination with *vision transformers* (Dosovitskiy et al., 2020), which is also demonstrated to be scalable into large vision models (He et al., 2021; Li et al., 2022b).

In this paper, we aim to build up a *unified* understanding framework for *MIM* and *contrastive learning*. Our motivation is, even though MIM obtains great success, it is still an open question how it works. Several works try to interpret MIM from different views, for example, He et al. (2021) suggests MIM model learns "rich hidden representation" via reconstruction from masked images; afterwards, Cao et al. (2022) gives a mathematical understanding for *MAE* (He et al., 2021). However, what the model learns is still not obvious. The difficulty lies in that MIM is essentially *reconstructive* (Eq. 2), hence the supervision on the learned feature (z) is *implicit*. In contrast, contrastive learning acts as a *siamese* nature (Eq. 1), which involves *explicit* supervision on the representation. If we manage to formulate MIM into an equivalent siamese form like Eq. 1, MIM can be *explicitly* interpreted as learning a certain *invariance* according to some *distance measurement*. We hope the framework may inspire more powerful self-supervised methods in the community.

In the next sections, we introduce our methodology. Notice that we do not aim to set up a new state-of-the-art MIM method, but to improve the understanding of MIM frameworks. Our findings are concluded as follows:

- We propose *RelaxMIM*, a new *siamese* framework to approximate the original *reconstructive* MIM method. In the view of RelaxMIM, MIM can be interpreted as a special case of contrastive learning: the data *transformation* is random patch masking and the *similarity measurement* relates to the decoder. In other words, **MIM models intrinsically learn occlusion invariant features**.
- Based on RelaxMIM, we replace the similarity measurement with simpler *InfoNCE loss*. Surprisingly, the performance maintains the same as the original model. It suggests that the reconstructive decoder in MIM framework does not matter much; other measurements could also work fine. Instead, **patch masking may be the key to success**.
- To understand why patch masking is important, we perform MIM pretraining on very few images (*e.g.* only **1** image), then finetune the encoder with supervised training on full ImageNet. Though the learned representations lack of semantic information after pretraining, the finetuned model still significantly outperforms those training from scratch. We hypothesize that the encoder learns *data-agnostic* occlusion invariant features during pretraining, which could be a favored initialization for finetuning.

2 MIM INTRINSICALLY LEARNS OCCLUSION INVARIANT FEATURE

In this section, we mainly introduce how to approximate *MIM* formulation (Eq. 2) with a siamese model. For simplicity, we take *MAE* (He et al., 2021) as an representative example of MIM, in which the *similarity measurement* is simply $l2-$distance on the *masked* patches. Other MIM methods can be analyzed in a similar way. Following the notations in Eq. 2, the loss function for MAE training is[‡]:

$$L(x, M) = \|d_\phi(f_\theta(x \odot M)) \odot (1-M) - x \odot (1-M)\|^2. \tag{3}$$

[†] So "$x \odot M$" represents "unmasked patches" and vice versa.

[‡] In original MAE (He et al., 2021), the encoder network only generates tokens of unmasked patches and the decoder only predict the masked patches during training. In our formulations, for simplicity we suppose both networks predict the *whole* feature map; we equivalently extract the desired part via proper *masking* if necessary.

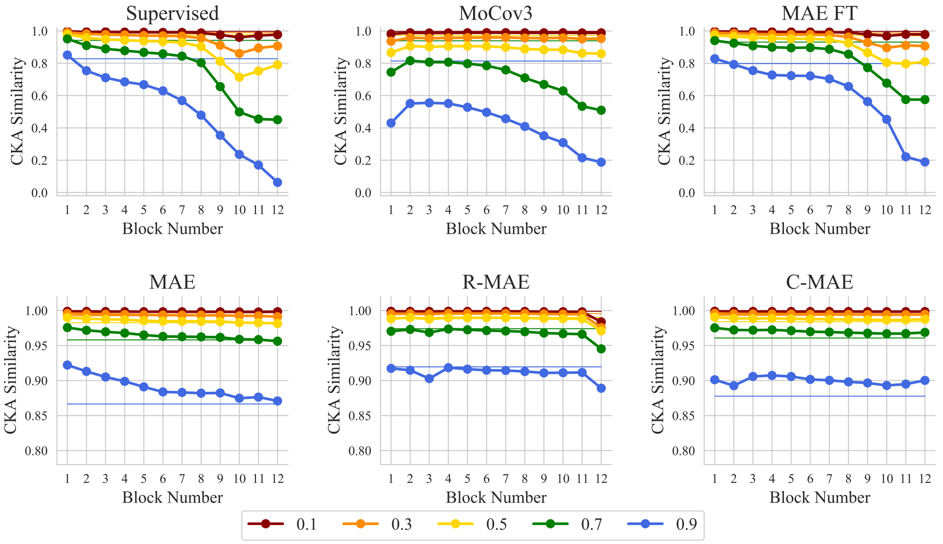


Figure 1: CKA similarity between the representations generated by the masked image and the full image respectively under different mask ratios. *FT* means finetuned model on ImageNet; and other models except "Supervised" are self-supervised pretrained models. (Best view in color.)

Let us focus on the second term. Typically, the dimension of feature embedding is much larger than dimension of input image, thus the encoder (at least) has a chance to be lossless (Li et al., 2022b). That means for the encoder function $f_\theta(\cdot)$, there exists a network $d'_{\phi'}(\cdot)$ parameterized by ϕ' that satisfying $d'_{\phi'}(f_\theta(x \odot (1-M))) \odot (1-M) \approx x \odot (1-M)$. Then, we rewrite Eq. 3 in the following equivalent form:

$$
\begin{aligned}
L(x, M) &= \|d_\phi(f_\theta(x \odot M)) \odot (1-M) - d'_{\phi'}(f_\theta(x \odot (1-M))) \odot (1-M)\|^2 \\
s.t. \quad \phi' &= \arg\min_{\phi'} \mathop{\mathbb{E}}_{x' \sim \mathcal{D}} \|d'_{\phi'}(f_\theta(x' \odot (1-M))) \odot (1-M) - x' \odot (1-M)\|^2
\end{aligned}
\tag{4}
$$

Eq. 4 can be further simplified. Notice that $d'_{\phi'}(\cdot)$ just approximates the "inverse" (if exists) of $f_\theta(\cdot)$, there is no reason to use a different architecture from $d_\phi(\cdot)$. So we let $d' = d$. Then we define a new *similarity measurement*:

$$
\overline{\mathcal{M}_{\phi,\phi'}}(z_1, z_2) \triangleq \|(d_\phi(z_1) - d_{\phi'}(z_2)) \odot (1-M)\|^2, \tag{5}
$$

and transformations:

$$
\mathcal{T}_1(x) = x \odot M, \quad \mathcal{T}_2(x) = x \odot (1-M), \tag{6}
$$

hence Eq. 4 equals to:

$$
\begin{aligned}
L(x, M; \theta, \phi) &= \overline{\mathcal{M}_{\phi,\phi'}}(f_\theta(\mathcal{T}_1(x)), f_\theta(\mathcal{T}_2(x))) \\
s.t. \quad \phi' &= \arg\min_{\phi'} \mathop{\mathbb{E}}_{x' \sim \mathcal{D}} \|(d_{\phi'}(f_\theta(\mathcal{T}_2(x'))) - \mathcal{T}_2(x')) \odot (1-M)\|^2.
\end{aligned}
\tag{7}
$$

We name Eq. 7 **siamese form of MAE**.

Discussion. Eq. 7 helps us to understand MIM from a *explicit view*. Compared Eq. 7 with Eq. 1, the formulation can be viewed as a special case of *contrastive learning*: the loss aims to minimize the differences between the representations derived from two masking *transformations*. Therefore, we conclude that **MIM pretraining encourage occlusion invariant features**. The decoder joints as a part of the *similarity measurement* (see Eq. 5), which is reasonable: since it is difficult to define a proper distance function directly in the latent space, a feasible solution is to project the representation back into the image space, because similarities like $l2$-distance in image space are usually explainable (analogous to *PSNR*). In addition, the constraint term in Eq. 7 can be viewed as standard *AutoEncoder*

Table 1: Comparisons of self-supervised methods on ImageNet with ViT-B (Dosovitskiy et al., 2020). *Epochs* in the table indicate numbers of pretraining epochs (for random initialization baselines they are total epochs of training from scratch). *PSNR* means the similarity between the generated image (from the masked image) and the original image after pretraining.

Pretrain Methods	Transformation	Framework	Epochs	FT Acc (%)	PSNR (dB)
Random Init	–	–	100	80.9	–
	–	–	300	82.1	–
MoCov3 (Chen et al., 2021)	crop & jitter	siamese	300	83.2	–
DINO (Caron et al., 2021)	crop & jitter	siamese	800	82.8	–
BeiT (Bao et al., 2021)	patch masking	reconstructive	300	82.9	–
MAE (He et al., 2021)	patch masking	reconstructive	1600	83.6	19.3
CAE (Chen et al., 2022)	patch masking	reconst. + siam.	300	83.3	–
MAE (*our impl.*)	patch masking	reconstructive	100	83.1	22.2
R-MAE (*ours*)	patch masking	siamese	100	82.7	23.7

defined on the space of $\mathcal{T}_2(x)$, which guarantees the projection $d_{\phi'}(\cdot)$ to be informative, avoiding collapse of the similarity measurement.

Although Eq. 7 *explicitly* uncovers the invariant properties of MIM in theory, it is a drawback that Eq. 7 involves a nested optimization, which is difficult to compute. We thus propose a *relaxed form* of Eq. 7, named *R-MAE* (or *RelaxMIM* in general):

$$\min_{\theta,\phi,\phi'} \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \overline{\mathcal{M}_{\phi,\phi'}}(f_\theta(\mathcal{T}_1(x)), f_\theta(\mathcal{T}_2(x))) + \lambda\|(d_{\phi'}(f_\theta(\mathcal{T}_2(x))) - \mathcal{T}_2(x)) \odot (1-M)\|^2. \tag{8}$$

Eq. 8 jointly optimizes the distance term and the constraint term in Eq. 7. λ controls the balance of the two terms. In practice, we let $\phi = \phi'$ to save computational cost, as we empirically find the optimization targets of $d_\phi(\cdot)$ and $d_{\phi'}(\cdot)$ in Eq. 8 do not diverge very much.

Empirical evaluation. First, we verify our claim that MIM representation is robust to image occlusion, as suggested by Eq. 7. We compute the CKA similarity (Kornblith et al., 2019) between the learned features from full images and images with different mask ratios respectively, at each block in the encoder. Figure 1 shows the CKA similarities of different models. The numbers (0.1 to 0.9) indicate the mask ratios (i.e. percentages of image patches to be dropped) of the test images respectively. As shown in Figure 1, both original *MAE* and our relaxed *R-MAE* (as well as another variant *C-MAE*, see the next section) obtain high CKA scores, suggesting those methods learn occlusion invariant features. In contrast, other methods such as supervised training or *MoCo v3* (Chen et al., 2021) do not share the property, especially if the drop ratio is large. After finetuning, the CKA similarities drop, but are still larger than those training from scratch.

Next, we verify how well *R-MAE* (Eq. 8) approximates the original MAE. We pretrain the original MAE and R-MAE on ImageNet using the same settings: the mask ratio is 0.75 and training epoch is 100 (λ is set to 1 for ours). Then we finetune the models on labeled ImageNet data for another 100 epochs. Results are shown in Table 1. Our finetuning accuracy is slightly lower than MAE by 0.4%, which may be caused by the relaxation. Nevertheless, R-MAE roughly maintains the benefit of MAE, which is still much better than supervised training from scratch and competitive among other self-supervised methods with longer pretraining. Another interesting observation is that, the reconstruction quality of R-MAE is even better than the original MAE (see *PSNR* column in Table 1), which we think may imply the trade-off by the choice of λ in Eq. 8. We will investigate the topic in the future.

3 Similarity measurement in MIM is replaceable

Eq. 7 bridges *MIM* and *contrastive learning* with a unified siamese framework. Compared with conventional contrastive learning methods (e.g. He et al. (2020); Chen et al. (2020b); Caron et al. (2021); Chen et al. (2021); Grill et al. (2020)), in MIM two things are special: 1) *data transformations* $\mathcal{T}(\cdot)$: previous contrastive learning methods usually employ random crop or other image jittering, while MIM methods adopt *patch masking*; 2) *similarity measurement* $\mathcal{M}(\cdot,\cdot)$, contrastive learning often uses *InfoNCE* or other losses, while MIM implies a relatively complex[§] formulation as Eq. 5. To understand whether the two differences are important, in this section we study how the choice of $\mathcal{M}(\cdot,\cdot)$ affects the performance.

Contrastive MAE (C-MAE). We aim to replace the measurement $\overline{\mathcal{M}_{\phi,\phi'}}(\cdot,\cdot)$ with a much simpler *InfoNCE loss* (Oord et al., 2018). We name the new method *contrastive MAE (C-MAE)*. Inspired by Chen et al. (2021); Grill et al. (2020), we transform the representations with *asymmetric MLPs* before applying the loss. The new distance measurement is defined as follows:

$$\widetilde{\mathcal{M}_{\phi,\phi'}}(z_1, z_2) \triangleq L_{\text{NCE}} = -\log \frac{\exp(s(z_1, z_2)/\tau)}{\sum_j \exp(s(z_1, z'_j)/\tau)}, \tag{9}$$

and

$$s(z, z') = \frac{q_{\phi'}(p_\phi(z)) \cdot p_\phi(z')}{\|q_{\phi'}(p_\phi(z))\| \cdot \|p_\phi(z')\|}, \tag{10}$$

where $p_\phi(\cdot)$ and $q_{\phi'}(\cdot)$ are *project head* and *predict head* respectively following the name in *BYOL* (Grill et al., 2020), which are implemented with *MLPs*; τ is the temperature of the softmax. Readers can refer to (Chen et al., 2021) for details. Hence the objective function of C-MAE is:

$$L(x, M; \theta, \phi, \phi') = \widetilde{\mathcal{M}_{\phi,\phi'}}(f_\theta(\mathcal{T}_1(x)), f_\theta(\mathcal{T}_2(x))). \tag{11}$$

Unlike Eq. 7, C-MAE does not include nested optimization, thus can be directly optimized without relaxing.

The design of transformation $\mathcal{T}$. We intend to use the same transformation as we used in *MAE* and *R-MAE* (Eq. 6). However, we find directly using Eq. 6 in C-MAE leads to convergence problem. We conjecture that even though the two transformations derive different patches from the same image, they may share the same color distribution, which may lead to information leakage. Inspired by *SimCLR* (Chen et al., 2020b), we introduce additional color augmentation after the transformation to cancel out the leakage. The detailed color jittering strategy follows *SimSiam* (Chen & He, 2020).

Token-wise vs. instance-wise loss. We mainly evaluate our method on *ViT-B* (Dosovitskiy et al., 2020) model. By default, the model generates a latent representation composed of 14×14 patch tokens and one class token, where each patch relates to one image patch while the class token relates to the whole instance. It is worth discussing how the loss in Eq. 11 applies to the tokens. We come up with four alternatives: apply the loss in Eq. 11 1) only to the class token; 2) on the average of all patch tokens; 3) to each patch token respectively; 4) to each patch token as well as the class token respectively. If multiple tokens are assigned to the loss, we gather all loss terms by averaging them up. Table 5 shows the ablation study results. It is clear that token-wise loss on the patch tokens achieves the best finetuning accuracy on *ImageNet*. In comparison, adding the class token does not lead to improvement, which may imply that class token in self-supervised learning is not as semantic as in supervised learning. Therefore, we use a token-wise-only strategy for C-MAE by default.

Implementation details. Following Chen et al. (2021), we use a siamese network, which contains an online model and a target model whose parameters are EMA updated by the online model. We use 2-layer projector (i.e. $p_\phi(\cdot)$ in Eq. 11) and 2-layer predictor ($q_{\phi'}$), and use GELU as activation layer. To represent the masked patches into the encoder network, we adopt *learnable mask tokens* as Xie et al. (2021b); Bao et al. (2021) does rather than directly discard the tokens within the masked region as the original MAE, because unlike MAE, our C-MAE does not include a heavy transformer-based decoder to predict the embeddings for the masked region.

Result and discussion. Table 2 shows the finetuning results of C-MAE and a few other self-supervised methods. C-MAE achieves comparable results with the counterpart MAE baselines, suggesting that *in MIM framework the reconstructive decoder, or equivalently the measurement in siamese form (Eq. 5), does not matter much*. A simple *InfoNCE loss* works fine. We also notice that our findings agree with recent advances in *siamese MIMs*, *e.g. iBOT* (Zhou et al., 2021), *MSN* (Assran et al., 2022) and *data2vec* (Baevski et al., 2022), whose frameworks involve various distance

[§] Notice that the constraint term in Eq. 7 also belongs to the similarity measurement.

Table 2: Comparisons of C-MAE and other pretraining methods on ImageNet finetuning. All models are based on ViT-B.

Pretrain Methods	Epochs	FT Acc (%)
Random Init	–	80.9
MoCo v3 (Chen et al., 2021)	300	83.2
DINO (Caron et al., 2021)	800	82.8
MAE (He et al., 2021)	1600	83.6
MAE (*our impl.*)	100	83.1
C-MAE (*ours*)	100	82.9
MAE (*our impl.*)	400	83.2
C-MAE (*ours*)	400	83.1

measurements between the siamese branches instead of reconstructing the unmasked parts, however, achieve comparable or even better results than the original reconstruction-based MIMs like He et al. (2021); Bao et al. (2021). In addition to those empirical observations, our work uncovers the underlying reason: both reconstructive and siamese methods target learning occlusion invariant features, thereby it is reasonable to obtain similar performances.

Table 2 also indicates that, as siamese frameworks, C-MAE achieves comparable or even better results than previous counterparts such as *DINO* (Caron et al., 2021), even though the former mainly adopts random patch masking while the latter involves complex strategies in *data transformation*. He et al. (2021) also reports a similar phenomenon that data augmentation is less important in MIM. The observation further supports the viewpoint that *learning occlusion invariant feature is the key to MIM, rather than the loss.* Intuitively, to encourage occlusion invariance, patch masking is a simple but strong approach. For example, compared with random crop strategy, patch masking is more general – cropping can be viewed as a special mask pattern on the whole image, however, according to the experiments in Xie et al. (2021b); He et al. (2021), it is good enough or even better to leave patch masking fully randomized[¶].

Additional ablations. Table 3 presents additional results on *MAE* and *C-MAE*. First, Although C-MAE shows comparable fine-tuning results with MAE, we find under *linear probing* (He et al., 2020; 2021) and *few-shot* (i.e. fine-tuning on 10% ImageNet training data) protocols, C-MAE models lead to inferior results. Further study shows the degradation is mainly caused by the usage of mask tokens in C-MAE, which is absent in the original MAE – if we remove the mask tokens as done in MAE's encoder, linear probing and few-shot accuracy largely recover (however fine-tuning accuracy slightly drops), which we think is because mask tokens enlarge the structural gap between pretraining and linear/few-shot probing, since the network is not fully fine-tuned under those settings.

Second, we further try replacing the *InfoNCE* loss (Eq. 9) with *BYOL* (Grill et al., 2020) loss in *C-MAE*. Following the ablations in Table 5, we still make the BYOL loss in *token-wise* manner. Compared with InfoNCE, BYOL loss does not have explicit negative pairs. Results imply that BYOL loss shows similar trend as InfoNCE loss, which supports our viewpoint "similarity measurement in MIM is replaceable". However, we also find BYOL loss is less stable, resulting in slightly lower accuracy than that of InfoNCE.

Last, since our *C-MAE* involves *color jittering* (Chen et al., 2020b), one may argue that color transformation invariance could be another key factor other than occlusion invariance. We study the original *MAE* with additional color jittering (Table 3). We compare two configurations: a) augmenting the whole image before applying MAE; b) only augmenting the unmasked patches (i.e. the reconstruction targets keep the same). Results show that neither setting boosts MAE further, which implies the invariance of color jittering does not matter much. Moreover, we try the compositions of three different augmentation strategies on MAE and C-MAE. As shown in Table 4, although C-MAE gets more benefit with stronger augmentations, both MAE and C-MAE drop a lot of performance when removing patch masking. That indicates learning occlusion invariance is critical for the models. The details of the experiments are explained in Appendix A.

[¶] Although very recent studies (Shi et al., 2022; Kakogeorgiou et al., 2022; Li et al., 2022a; Wu & Mo, 2022) suggest more sophisticated masking strategies can still help.

Table 3: Additional comparisons on MAE and C-MAE. All models are pretrained and fine-tuned for 100 epochs respectively.

Pretrain Methods	Lin. Prob Acc (%)	FT Acc (%)	10% FT Acc (%)
MAE	54.5	83.1	67.5
MAE w/ color jitter (whole image)	53.4	83.1	67.3
MAE w/ color jitter (unmasked only)	54.0	83.0	67.3
C-MAE	41.1	82.9	66.4
C-MAE w/o mask token	56.2	82.6	67.5
C-MAE (BYOL loss)	26.9	82.8	65.2
C-MAE w/o mask token (BYOL loss)	55.2	82.5	66.1

Table 4: Ablation study of different augmentation strategies on MAE and C-MAE. The models are trained for 100 epochs with *ViT-S* on *ImageNet-100*. Sup means 200-epoch supervised result. 🟩 means patch masking, 🟦 means cutmix, and 🟥 means color augmentation. (Best view in color.)

Pretrain Methods	Sup		🟩	🟦	🟥	🟩🟦	🟩🟥	🟦🟥	🟩🟦🟥
MAE	81.6	71.9	**87.1**	83.3	81.6	86.1	85.5	83.8	86.4
C-MAE		80.0	86.5	83.5	82.8	86.9	86.8	83.1	**87.1**

4 MIM LEARNS A FAVORED, (ALMOST) DATA-AGNOSTIC INITIALIZATION

As discussed in the above sections, learning occlusion invariant features is the key “philosophy” of *MIM* methods. Hence an interesting question comes up: how do the learned networks model the invariance? One possible hypothesis is that occlusion invariance is represented in an *data-agnostic* way, just analogous to the structure of *max pooling* – the output feature is robust only if the most significant input part is not masked out, thereby the invariance is obtained by design rather than data. Another reasonable hypothesis is, in contrast, the invariance requires knowledge from a lot of data. In this section we investigate the question.

Inspired by Asano et al. (2019), to verify our hypotheses we try to *significantly* reduce the number of images for *MAE* pretraining, i.e. ranging from 1 for 1000 randomly sampled from *ImageNet* training set, hence the semantic information from training data should be very limited in the pretraining phase. Notice that MAE training tends to suffer from over-fitting on very small training set, as the network may easily “remember” the training images. Therefore, we adopt stronger data augmentation and early-stop trick to avoid over-fitting. Table 6 presents the result. Very surprisingly, we find pretraining with only *one* image with 5 epochs already leads to improved finetuning score – much better than 100-epoch training from scratch and on par with training for 300 epochs. The fine-tuning results do not improve when the number of pretrain images increases to 1000. Since it is not likely for only one image to contain much of the semantic information of the whole dataset, the experiment provides strong evidence that MIM can learn a favored initialization, more importantly, which is (almost) data-agnostic.

Moreover, in Table 7 we benchmark various pretraining methods on a 1000-image subset from *ImageNet* training data, which provides more insights on MIM training. We find the linear probing accuracy of *MAE* is very low, which is only slightly better than random feature (first row), suggesting

Table 5: Ablation study on the strategies of C-MAE loss. All models are pretrained for 100 epochs.

Measurement	Class Token	Patch Tokens	FT Acc (%)
instance-wise	✓		82.5
instance-wise		average	82.6
token-wise		✓	82.9
token-wise	✓	✓	82.9

Table 6: Comparisons of MAE pretrained with different numbers of images.

Pretrain Images	Stronger Aug.	Train Epochs	FT Epochs	FT Acc (%)
1	✓	5	100	**82.3**
10		2	100	81.9
	✓	5	100	82.3
100		10	100	82.1
	✓	10	100	82.2
1000		100	100	82.2
	✓	100	100	82.2
Random Init		-	100	80.9
		-	300	82.1

Table 7: Comparisons of different pretraining methods on 1000 images sampled from ImageNet (one image for each class). All methods pretrain for 100 epochs on the sampled dataset (except for random initialized baseline) and then fine-tune for 100 epochs on full/10%-ImageNet accordingly.

Pretrain Methods	Lin. Prob Acc (%)	FT Acc (%)	10% FT Acc (%)
Random Init	6.1	80.9	34.9
Supervised	33.1	81.0	52.6
MoCo v3	37.3	79.2	45.8
MAE	13.8	82.2	57.6
R-MAE	25.9	82.1	58.8
C-MAE	20.1	82.1	61.9

that the feature learned from 1000 images is less semantic; however, the finetuning result as well as few-shot fine-tuning is fine. Our proposed *R-MAE* and *C-MAE* share similar properties as the original MAE – relatively low linear probing scores but high fine-tuning performance. The observation strongly supports our first hypothesis at the beginning of Sec. 4: the occlusion invariance learned by MIM could be data-agnostic, which also serves as a good initialization for the network. In comparison, supervised training and *MoCo v3* (Chen et al., 2021) on 1000 images fail to obtain high fine-tuning scores, even though their linear probing accuracy is higher, which may be because those methods cannot learn occlusion-invariant features from small dataset effectively. In Appendix B, we will discuss more on the topic.

5 Experimental Details

Pretraining. We use ViT-B/16 as the default backbone. For MAE pretraining, we use the same settings as (He et al., 2021), and use the patch normalization when computing loss. We use the mask ratio of 0.75, which is the most effective one in (He et al., 2021). We use AdamW optimizer with cosine decay scheduler and the batch size is set to 1024. We set the base learning rate (learning rate for batch size of 256) as 1.5e-4 with a 20-epoch linear warm-up and scale up the learning rate linearly when batch size increases (Goyal et al., 2017). For R-MAE, we search the learning rate and finally set the base learning rate as 3.0e-4. Other training settings are the same as He et al. (2021). For C-MAE, the momentum to update the teacher model is set to 0.996, and the temperature to compute contrastive loss is set to 0.2. For projector and predictor heads, we set 2048-d for hidden layers. We search the learning rate and finally set the base learning rate as 1.5e-4. Other parameters are the same as C-MAE. We train the model for 100 epochs on the ImageNet (Russakovsky et al., 2015) dataset as default. Due to the computational resource constraints, we report the results of 400 epochs to prove that our method gains better results with longer training.

Finetuning. We follow the training settings in He et al. (2021). We use the average pooling feature of the encoded patch tokens as the input of classifier, and train the model end-to-end. Following He et al. (2021), we reset the parameters of the final normalization layer. We use AdamW optimizer with cosine decay scheduler and set the batch size to 1024. We set the base learning rate as 1.0e-3 with 5-epoch linearly warm-up and train the model for 100 epochs. Note that the supervised trained ViT in our paper uses the same settings as finetuning and the model is trained for 100 epochs.

6 Related Work

Masked Image Modeling. As the ViT models achieve breakthrough results in computer vision, self-supervised pretraining for ViTs becomes an intense scholarly domain. In addition to siamese frameworks such as Chen et al. (2021); Caron et al. (2021), MIM is an efficient and popular way of self-supervised modeling. The model learns rich hidden information by optimizing the reconstruction model (He et al., 2021). Following BERT (Devlin et al., 2018), Chen et al. (2020a) compress the image to a few pixels, and then directly learn the masked pixel color. Bao et al. (2021) maps all image patches to 8192 embeddings by training d-VAE (Ramesh et al., 2021), and then learns the correct embedding correspondence for mask patches. Li et al. (2021b) optimizes the masking process based on BEiT. El-Nouby et al. (2021); Zhou et al. (2021); Li et al. (2021b) combines MIM with siamese frameworks and improves the performance of linear probing. He et al. (2021); Xie et al. (2021b) use a simple method to reconstruct the original image, and also learn rich features effectively. Cao et al. (2022) gives a mathematical understanding of MAE. MSN (Assran et al., 2022), which is a concurrent work of ours, also discusses the invariance to mask.

Siamese approaches in SSL. Self-supervised pretraining achieves great success in classification (Dosovitskiy et al., 2014b; Doersch et al., 2015; Gidaris et al., 2018b; Noroozi & Favaro, 2016; Oord et al., 2018; Wu et al., 2018; He et al., 2020; Chen et al., 2020b; Grill et al., 2020; Zbontar et al., 2021), detection(Liu et al., 2020; Xiong et al., 2020; Lang et al., 2021; Xie et al., 2021a) and segmentation. One of the promising methods is based on siamese frameworks (Tian et al., 2019; He et al., 2020; Chen et al., 2020c; 2021; Misra & van der Maaten, 2020; Chen et al., 2020b;c; Grill et al., 2020; Caron et al., 2020; Chen & He, 2020; Xie et al., 2020; Caron et al., 2021; Zbontar et al., 2021; Bardes et al., 2021), which learns representations by minimizing the distance of positive samples with siamese networks. In practice, Chen et al. (2020b); Caron et al. (2020); Chen & He (2020); Zbontar et al. (2021) uses the same parameters in the online and target model, while He et al. (2020); Chen et al. (2020c; 2021); Grill et al. (2020); Caron et al. (2021) updates online parameters to target using exponential moving average. Only minimizing the distance of positive samples will cause the model to fall into trivial solutions, so a critical problem in SSL is how to prevent such a model from collapsing. Chen et al. (2020b); He et al. (2020) use negative samples from different images, then computes contrastive loss. Grill et al. (2020); Chen & He (2020) add an extra predictor on the top of the online model then stop the gradient of the target model. Instead of optimizing the loss per instance, Zbontar et al. (2021); Bardes et al. (2021) optimize the variance, covariance or cross-covariance on the channel dimension. Caron et al. (2021) optimize the distributions of the two features, and avoid trivial solutions by centering and sharpening.

7 Conclusion

In this paper, we propose a new viewpoint: MIM implicitly learns occlusion-invariant features, and build up a unified understanding framework *RelaxMIM* for MIM and contrastive learning. In the view of RelaxMIM, MIM models intrinsically learn *occlusion invariant features*. Then we verify that the representation of RelaxMIM is robust to image occlusion. Based on RelaxMIM, we replace the similarity measurement with simpler InfoNCE loss and achieve comparable results with the original MIM framework. It suggests that *patch masking* may be the critical component of the framework. To understand why patch masking is important, we perform MIM pretraining on very few images and finetune the encoder with supervised training on full ImageNet. We find that the encoder learns almost *data-agnostic* occlusion invariant features during pretraining, which could be a favored initialization for finetuning. To measure whether the MIM method has learned human recognition patterns, we compare the shape bias of different self-supervised models and conclude that, MIM could improve the recognition ability of ViT to make it closer to human recognition, but the improvement may be limited. We hope the RelaxMIM framework may inspire more powerful self-supervised methods in the community.

REFERENCES

Yuki M Asano, Christian Rupprecht, and Andrea Vedaldi. A critical analysis of self-supervision, or what we can learn from a single image. *arXiv preprint arXiv:1904.13132*, 2019.

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Michael Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. *arXiv preprint arXiv:2204.07141*, 2022.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555*, 2022.

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

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pp. 531–540. PMLR, 2018.

Shuhao Cao, Peng Xu, and David A Clifton. How to understand masked autoencoders. *arXiv preprint arXiv:2202.03670*, 2022.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/70feb62b69f16e0238f741fab228fec2-Abstract.html`.

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

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, pp. 1691–1703. PMLR, 2020a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020b. URL `http://proceedings.mlr.press/v119/chen20j.html`.

Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *arXiv preprint arXiv:2202.03026*, 2022.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020c.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021.

Rumen Dangovski, Li Jing, Charlotte Loh, Seungwook Han, Akash Srivastava, Brian Cheung, Pulkit Agrawal, and Marin Soljačić. Equivariant contrastive learning. *arXiv preprint arXiv:2111.00899*, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 1422–1430. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.167. URL `https://doi.org/10.1109/ICCV.2015.167`.

Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Peco: Perceptual codebook for bert pre-training of vision transformers. *arXiv preprint arXiv:2111.12710*, 2021.

Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27, 2014a.

Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014b. URL `https://proceedings.neurips.cc/paper/2014/file/07563a3fe3bbe7e3ba84431ad9d055af-Paper.pdf`.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Alaaeldin El-Nouby, Gautier Izacard, Hugo Touvron, Ivan Laptev, Hervé Jegou, and Edouard Grave. Are large-scale datasets necessary for self-supervised pre-training? *arXiv preprint arXiv:2112.10740*, 2021.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018a.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018b. URL `https://openreview.net/forum?id=S1v4N2l0-`.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html`.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 9726–9735. IEEE, 2020. doi: 10.1109/CVPR42600.2020.00975. URL `https://doi.org/10.1109/CVPR42600.2020.00975`.

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

Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. What to hide from your students: Attention-guided masked image modeling. *arXiv preprint arXiv:2203.12719*, 2022.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pp. 3519–3529. PMLR, 2019.

Yvette Kosmann-Schwarzbach. The noether theorems. In *The Noether Theorems*, pp. 55–64. Springer, 2011.

Christopher Lang, Alexander Braun, and Abhinav Valada. Contrastive object detection using knowledge graph embeddings. *arXiv preprint arXiv:2112.11366*, 2021.

Gang Li, Heliang Zheng, Daqing Liu, Bing Su, and Changwen Zheng. Semmae: Semantic-guided masking for learning masked autoencoders. *arXiv preprint arXiv:2206.10207*, 2022a.

Wenbin Li, Xuesong Yang, Meihao Kong, Lei Wang, Jing Huo, Yang Gao, and Jiebo Luo. Triplet is all you need with random mappings for unsupervised visual representation learning. *arXiv preprint arXiv:2107.10419*, 2021a.

Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *arXiv preprint arXiv:2203.16527*, 2022b.

Zhaowen Li, Zhiyang Chen, Fan Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu, Rui Zhao, Ming Tang, et al. Mst: Masked self-supervised transformer for visual representation. *Advances in Neural Information Processing Systems*, 34, 2021b.

Songtao Liu, Zeming Li, and Jian Sun. Self-emd: Self-supervised object detection without imagenet. *CoRR*, abs/2011.13677, 2020. URL `https://arxiv.org/abs/2011.13677`.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 6706–6716. IEEE, 2020. doi: 10.1109/CVPR42600.2020.00674. URL `https://doi.org/10.1109/CVPR42600.2020.00674`.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*. Springer, 2016.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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

Colorado Reed, Sean Metzger, Aravind Srinivas, Trevor Darrell, and Kurt Keutzer. Evaluating self-supervised pretraining without using labels. *CoRR*, abs/2009.07724, 2020. URL `https://arxiv.org/abs/2009.07724`.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Yuge Shi, N Siddharth, Philip Torr, and Adam R Kosiorek. Adversarial masking for self-supervised learning. In *International Conference on Machine Learning*, pp. 20026–20040. PMLR, 2022.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L. Griffiths. Are convolutional neural networks or transformers more like human vision? *CoRR*, abs/2105.07197, 2021. URL `https://arxiv.org/abs/2105.07197`.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103, 2008.

Jiantao Wu and Shentong Mo. Object-wise masked autoencoders for fast pre-training. *arXiv preprint arXiv:2205.14338*, 2022.

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 3733–3742. IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00393. URL `http://openaccess.thecvf.com/content_cvpr_2018/html/Wu_Unsupervised_Feature_Learning_CVPR_2018_paper.html`.

Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8392–8401, 2021a.

Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. *arXiv preprint arXiv:2011.10043*, 2020.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. *arXiv preprint arXiv:2111.09886*, 2021b.

Yuwen Xiong, Mengye Ren, and Raquel Urtasun. Loco: Local contrastive representation learning. *Advances in neural information processing systems*, 33:11142–11153, 2020.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pp. 12310–12320. PMLR, 2021.

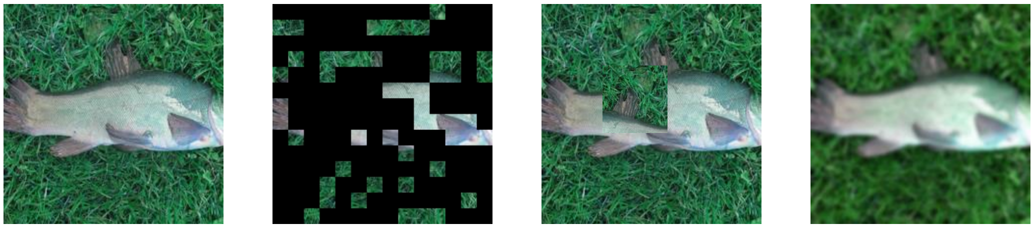
Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pp. 649–666. Springer, 2016.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

A Configurations of Augmentation Experiments

Augmentation strategies. Generally in MAE, the *source image* which is send into encoder and the *target image* which is the target to reconstruct are always the same. Here we try to add additional augmentations on the source image. After the random-resize-cropped and horizontal-flip augmentation, the image is cropped to 224×224. Then we try the compositions of three additional augmentation strategies on the source image. The definitions of augmentation strategies are described as follow:

1. Patch masking (Figure 2(b)): we divide the image into non-overlapping 16×16 patches and randomly mask 75% patches, then the image is occluded by small and neat black blocks.
2. Cutmix (Figure 2(c)): we cropped a patch from the original image then resize the patch and paste it on the image. The image is occluded by an object rather than small blocks.
3. Color augmentation (Figure 2(d)): we use the same color augmentation as SimSiam (Chen & He, 2020). Although there is no occlusion, the entire source image and target image are slightly different in color and texture after color augmentation.



(a) Target image. (b) Patch masking. (c) CutMix. (d) Color augmentation.

Figure 2: Visualization of different augmentation strategies.

Model configurations. We use ViT-S/16 as backbone, and the models are trained and evaluated on ImageNet-100[‖] classification. For C-MAE, we use normalized L2 loss (Grill et al., 2020) as measurement. Other configurations are the same as default.

B More Visualization Experiments

B.1 Occlusion-invariance of Few Images Pretrained MAE

Here we discuss the occlusion invariance of a few images pretrained MAE models. We use **CKA similarities** between the representations generated by the masked image and the full image under different mask ratios as protocol. The numbers (0.1 to 0.9) indicate the mask ratios (i.e. percentages of image patches to be dropped) of the test images respectively. The higher CKA similarity with a large mask ratio means the model learns better occlusion invariance.

Figure 3 shows the CKA similarities of MAE pretrained with different amounts of data. As the figures show, the model learns occlusion invariance even pretrained with one image. Unfortunately, the model does not keep the occlusion invariance after finetuning. When the mask ratio increase to 0.7, the CKA similarities drop significantly below 0.5. In Comparison, full-set pretrained MAE is not so sensitive to the change of mask ratio (after 0.7) after finetuning.

Furthermore, we discuss the relationship between occlusion invariance with overfitting. We train the MAE with different training epochs on 10 images and plot the CKA similarities. Results in Figure 4 show that, overfitting affects the learning of occlusion invariance, and causes the performance to drop. We further explore the way to prevent overfitting, *using stronger data augmentation*, whether beneficial to maintain occlusion invariance. As shown in Figure 4, even the finetuning results increase a little when using stronger augmentations, the occlusion invariance does not been improved.

[‖] https://www.kaggle.com/datasets/ambityga/imagenet100

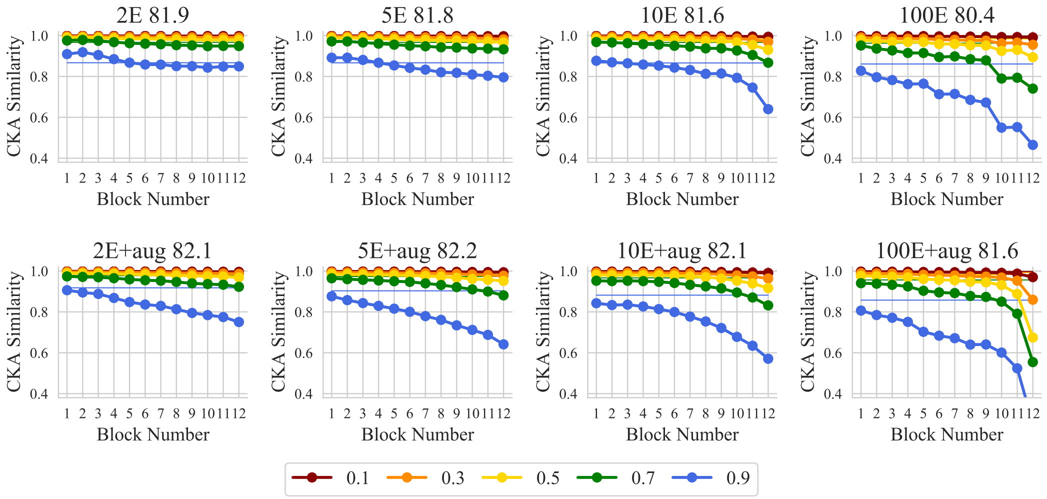


Figure 4: CKA similarity between the representations generated by the masked image and the full image respectively under different mask ratios. The number on the subtitle is the finetuning result of the model. All models are pretrained on 10 images for different epochs and finetuned on full ImageNet training set for 100 epochs. N**E** means the model pretrained for N epochs. **+aug** means adding stronger augmentations.

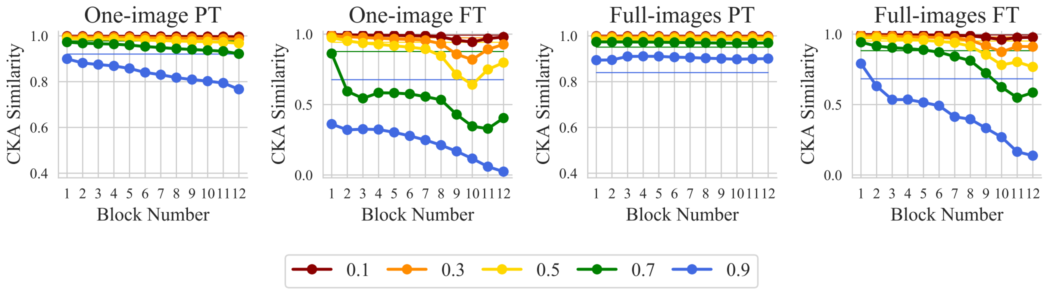


Figure 3: CKA similarity between the representations generated by the masked image and the full image respectively under different mask ratios. **One-image PT** indicates the model that pretrained on one image for 5 epochs, and **Full-images PT** indicates the model pretrained on ImageNet training set for 100 epochs. The finetuning models (**FT**) are all trained on ImageNet training set for 100 epochs.

B.2 Comparison with Human Recognition

Tuli et al. (2021) shows that, ViT behaves more like humans in classification, and we wonder whether our proposed siamese framework learns more high-level perception. Following the method in Tuli et al. (2021), we plot the shape bias of MIM models in Figure 5.

Figure 5 shows the shape bias of MAE, MoCov3, R-MAE and C-MAE. As shown in the figure, the grey line represents the supervised trained model, which has the lowest shape bias. That means fully supervised learning prefers to learn texture information rather than self-supervised pretrained models. Both MAE (blue line) and R-MAE (green line) learn less shape bias than MoCo (yellow line) and C-MAE (orange line). We speculate that it is because the target of the pretext task of MIM is closer to the original images (or exactly the origin images), which makes the model learn more texture features. Additionally, C-MAE learns a similar shape-bias compared with MoCo v3. The results indicate that instance-wise learning is not necessary for models to learn as human does, learning occlusion

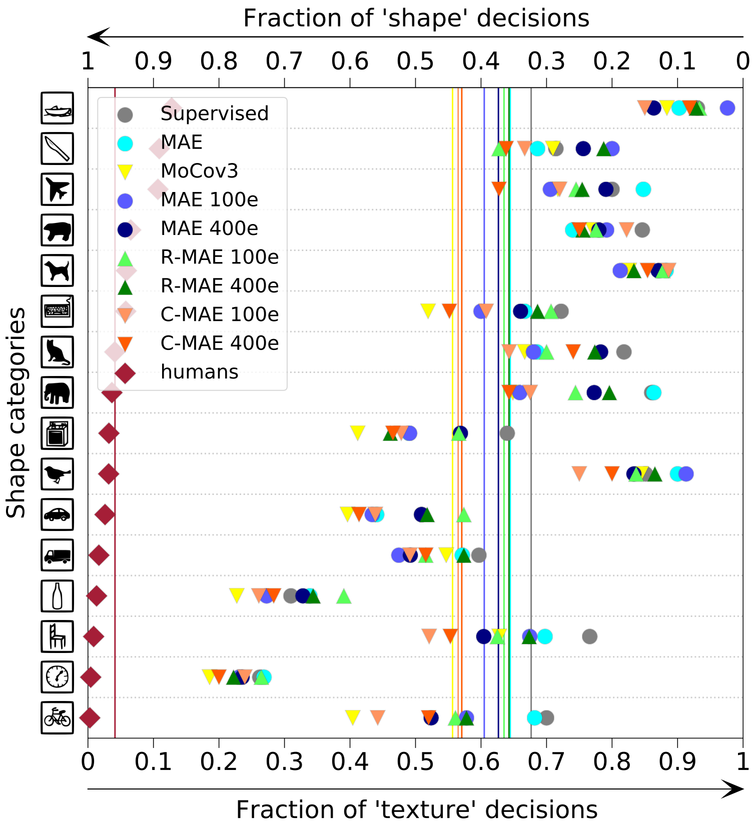


Figure 5: Shape bias of MAE, MoCov3, R-MAE and C-MAE pretrained on ImageNet. The vertical line is the average shape bias of 16 classes.

invariance could also improve the ability of the model to learn shape-bias. When training longer, all masked-based models are biased to learn texture features. We conclude that the masked-based models could learn the ability to complete object shape quickly in a few epochs, and then learn to reconstruct the texture of images.