A Survey on Masked Autoencoder for Visual Self-supervised Learning

Chaoning Zhang[1] , **Chenshuang Zhang**[2] , **Junha Song**[2] , **John Seon Keun Yi**[3] , **In So Kweon**[2]
[1]Kyung Hee University, South Korea
[2]Korea Advanced Institute of Science and Technology, South Korea
[3]Georgia Institute of Technology, United States
chaoningzhang1990@gmail.com, zcs15@kaist.ac.kr, sb020518@kaist.ac.kr
johnsk95@gatech.edu, iskweon77@kaist.ac.kr

Abstract

With the increasing popularity of masked autoencoders, self-supervised learning (SSL) in vision undertakes a similar trajectory as in NLP. Specifically, generative pretext tasks with the masked prediction have become a de facto standard SSL practice in NLP (e.g., BERT). By contrast, early attempts at generative methods in vision have been outperformed by their discriminative counterparts (like contrastive learning). However, the success of masked image modeling has revived the autoencoder-based visual pretraining method. As a milestone to bridge the gap with BERT in NLP, masked autoencoder in vision has attracted unprecedented attention. This work conducts a survey on masked autoencoders for visual SSL.

1 Introduction

In recent years, the mainstream trend of deep learning has gradually shifted from designing better models to solving the data-hungry issue in deep learning. For example, ImageNet with more than one million labeled images has become a typical benchmark dataset for vision models, and vision transformer (ViT) [Khan *et al.*, 2022] is reported to demand hundreds of times more labeled images. A common way to perform satisfactorily with a relatively small labeled dataset is to pre-train the model on another larger dataset, which is widely known as transfer learning. Outperforming its supervised counterpart for pre-training, visual SSL [He *et al.*, 2020; He *et al.*, 2022] has become an active research field.

With the advent of contrastive SSL [He *et al.*, 2020], joint-embedding methods have become a dominant visual pre-training framework; however, this status has been recently challenged by the success of a generative method termed masked image modeling (MIM). A successful attempt [Bao *et al.*, 2022] adopts a mask-then-predict strategy to train the model with the target visual tokens generated by an off-the-shelf tokenizer trained by a discrete variational autoencoder (dVAE). More recently, MAE [He *et al.*, 2022] simplifies this two-stage approach into an end-to-end masked autoencoder method, which has attracted unprecedented attention. Notably, we use MAE to refer to the method in [He *et al.*, 2022] *not* as shorthand for masked autoencoder to avoid confusion.

As the term suggests, a masked autoencoder is an autoencoder with masked prediction, *i.e.* predicting a property of masked input from unmasked input content. It is worth mentioning that masked autoencoder is not something new in unsupervised visual pretraining. Dating back to 2008, an early work [Vincent *et al.*, 2008] predicted masked pixels from unmasked ones but was referred to as denoising autoencoder. The success of MAE [He *et al.*, 2022], outperforming joint-embedding methods, revives this straightforward visual pretraining method. Except for the competitive performance, another reason for the attention on masked autoencoder is that a similar generative SSL framework termed masked language modeling (like BERT [Devlin *et al.*, 2019]) has been widely used in NLP. In other words, the success of masked autoencoder in vision paves a path: SSL in vision"*may now be embarking on a similar trajectory as in NLP*" [He *et al.*, 2022].

To this end, **this work conducts a survey on masked autoencoders for visual SSL, with a longer version available at [Zhang *et al.*, 2022a] discussing beyond vision**. With masked autoencoder in vision as the focus, this survey structure is organized as follows. Sec. 2 introduces the background featured by clear term definitions; Sec. 3 summarizes its historical development and relation with masked language modeling; Sec. 4 summarizes seminal works on masked autoencoders for visual SSL and design principles for improvement. Sec. 5 presents various perspectives on understanding the success of masked autoencoder. Sec. 6 discusses the relationship with joint-embedding methods; Sec. 7 covers the applications beyond pure images.

2 Background and Terminology

SSL: Generative *v.s.* discriminative. In self-supervised learning, modelling methods can be roughly categorized into: discriminative or generative. Generative SSL typically relies on an autoeocnder that consists of encoding (*i.e.* mapping an input to a latent representation with an encoder) and decoding (*i.e.* generating the input from the latent representation with an decoder). Discriminative SSL typically follows its supervised counterpart to design a discriminative loss.

Denoising autoencoder *v.s.* masked autoencoder. As a classical generative SSL method, denoising autoencoder is a class of autoencoders that reconstruct the original clean input from a corrupted input [Vincent *et al.*, 2008]. Note that *de-*

noising in this context (and in this whole survey) refers to reconstruction from general corruption (including but not limited to noise). Since *masked prediction* refers to the practice of predicting a property of masked input from unmasked input, it can be seen as a form of denoising process [Yi *et al.*, 2022]. This predicted property can be the original input [Yi *et al.*, 2022], handcrafted feature [Wei *et al.*, 2022a], or latent representation [Baevski *et al.*, 2022]. Since masked prediction is a form of denoising process and thus masked autoencoder can be seen as a form of general denoising autoencoder.

Masked autoencoding *v.s.* masked modeling. Masked prediction can be used for both generative and discriminative modeling methods. The term masked *X* modeling, namely masked modeling on *X*-type data, often refers to the generative case, such as masked *language* modeling, masked *image* modeling. However, masked modeling is not necessarily masked autoencoding. Take *image* data for example, MSN [Assran *et al.*, 2022] and data2vec [Baevski *et al.*, 2022] can be categorized as masked image modeling but not masked autoencoding since their model architectures are decoder-free.

3 Masked Autoencoding: NLP to Vision

3.1 NLP and Vision Followed Different SSL Paths

ML-driven AI has two major research fields: NLP and computer vision. Towards a unified understanding of language and image, it is interesting to ask whether they can follow a similar SSL path. For a long time they followed different paths: generative SSL in NLP and discriminative SSL in vision.

Generative SSL in NLP. In NLP there exist two leading language models: GPT and BERT. They are both based on the transformer architecture but with notable differences: GPT works by predicting the next word based on previous words and thus is autoregressive in nature, while BERT uses the entire surrounding context of words all at once. In essence, they both remove a portion of the data and predict the removed content, and therefore, they can be both perceived to rely on masked prediction as the pretext task.

Discriminative SSL in Vision. Joint-embedding methods, namely aligning the embedded representations of augmented views of the same image, have demonstrated substantial performance boost over prior generative methods. After the advent of MoCo [He *et al.*, 2020], contrastive learning, which makes the representations of positive samples close and those of negative samples far from each other, has emerged as a dominant visual SSL method. Negative-free (*i.e.* non-contrastive) joint-embedding methods have also been investigated [Chen and He, 2021], demonstrating comparable performance of contrastive learning methods.

3.2 Is Generative SSL Suitable for Vision?

Very early attempts. Denoising autoencoder was proposed in [Vincent *et al.*, 2008] to perform masked autoencoding by randomly masking some pixels. To make it a harder task to avoid learning only low-level representation, [Pathak *et al.*, 2016] proposed feature learning by inpainting, *i.e.* to fill in large missing areas of the image and thus prevent hints from nearby pixels. Later, [Larsson *et al.*, 2016; Zhang *et al.*, 2016] showed that masked channel prediction yielded superior performance on downstream tasks, especially for dense semantic segmentation.

	denoising	masked
Training dataset	MNIST	ImageNet
Model Architecture	CNN	ViT
Corruption size	pixels	patches
Corruption ratio	maximum 50%	75%

Table 1: Comparison of denosing autoencoder [Vincent *et al.*, 2008] and masked autoencoder [He *et al.*, 2022]

Inspiration from NLP. With GPT and BERT emerging in 2018/2019 to show the success of masked prediction in language understanding, a natural question is: can we transfer the success of masked modeling from language to image? iGPT [Chen *et al.*, 2020a] is the first successful attempt in this direction; however, as highlighted in [Chen *et al.*, 2020b], their work is for proof-of-concept and cannot be used in practice due to two reasons: (1) it takes two orders higher pre-training compute than contrastive methods and (2) it performs worse than contrastive methods based on CNN. [Dosovitskiy *et al.*, 2021] also investigated self-supervised pre-training. Since the self-supervised pre-training practice in [Dosovitskiy *et al.*, 2021], we call it iBERT since it mimicked the masked language modeling task in BERT. iBERT performs a masked patch prediction for visual SSL. However, this preliminary investigation of ViT for SSL also shows inferior performance compared with joint-embedding methods. This challenge was finally broken by BEiT [Bao *et al.*, 2022] as well as MAE [He *et al.*, 2022] (see Sec.4 for their details).

3.3 Summary and Remark

Summary. Figure 1 shows the overall timeline for the development of unsupervised visual pretraining (including GPT and BERT for NLP). Interestingly, unsupervised visual pretraining started with generative SSL in 2008. Its reviving attempt in 2016 and 2017 was then outperformed by discriminative SSL, especially after the advent of joint-embedding methods. However, with the inspiration from NLP, generative SSL with masked prediction comes back again.

Remark. Early denoising autoencoder [Vincent *et al.*, 2008] and recent masked autoencoder [He *et al.*, 2022] both reconstruct a clean input from a corrupted one by predicting masked input content from unmasked input content. Despite high similarity regarding pretext task, the masked autoencoder introduced in [He *et al.*, 2022] differs from early denoising autoencoder [Vincent *et al.*, 2008] in numerous ways, which are summarized in Table 1.

4 Masked Autoencoder for Image Modeling

As discussed in Sec.3, iGPT and iBERT have shown the possibility of transferring the pretext task of masked prediction from language to image data. However, their performance is inferior to joint-embedding methods and thus has

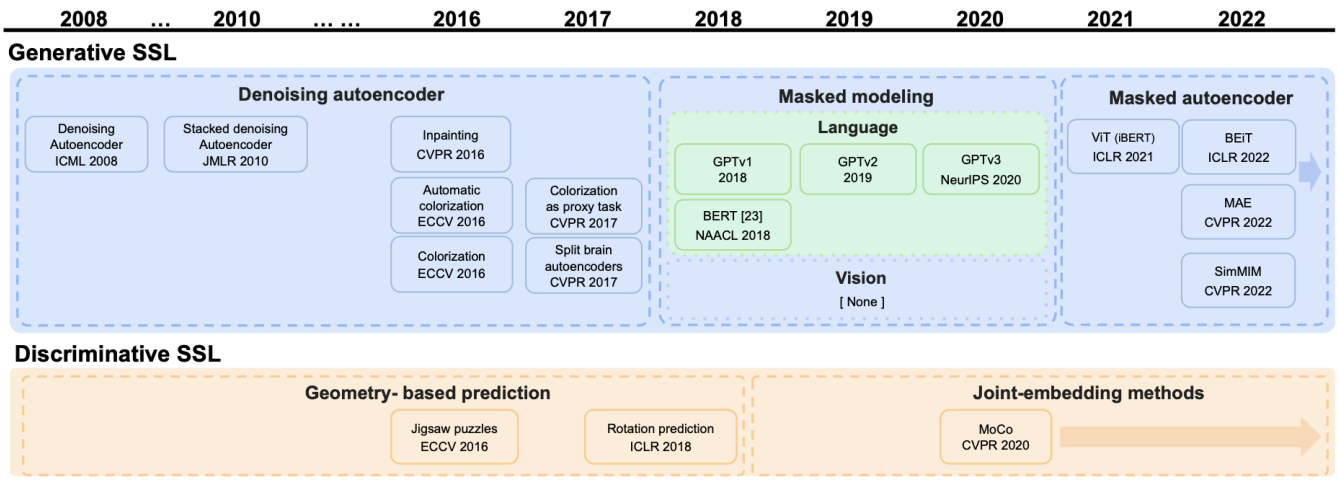

Figure 1: Timeline of Visual SSL

.

caught less attention. BEiT is the first to show the success of autoencoder-based masked prediction outperforming DINO, a SOTA joint-embedding method. Therefore, this section starts with introducing BEiT with its improved variants and then discusses the seminal work MAE [He *et al.*, 2022] .

4.1 BEiT and Its Improved Variants

BEiT. In contrast to iBERT [Dosovitskiy *et al.*, 2021] that directly reconstructs the masked patches, BEiT mimicks BERT to reconstruct visual tokens. Since Image patches do not have off-the-shelf tokens as words in the language, BEiT trains an image tokenizer via discrete variational autoencoder (dVAE) before the second-step masked image modeling where the tokenizer is used to guide the learning of BEiT encoder (note that decoder is unused). Specifically, the tokenizer takes the original image, and the BEiT encoder takes a corrupted image, including unmasked patches and masked patches. Then, it outputs the visual tokens of masked patches to match the corresponding visual tokens from the tokenizer (staying fixed in this process). BEiT is the first to show that masked image modeling has downstream task performance superior to SOTA contrastive DINO [Caron *et al.*, 2021]. BEiT [Bao *et al.*, 2022] consists of two stages: token-based MIM as the main stage and tokenizer training as the preparation stage. Multiple works [Dong *et al.*, 2021; Li *et al.*, 2022c; Chen *et al.*, 2022b] have followed this two-stage approach by either improving the tokenizer-based MIM process or seeking an alternative tokenizer.

Tokenizer-based MIM. mc-BEiT [Li *et al.*, 2022c] attempts to effectively utilize the visual tokenizer generated by dVAE. Considering the continous image space and discrete tokenizer, it is not desired that patches with similar semantics can have different token IDs, and patches with different semantics can have the same token ID. Therefore, mc-BEiT recasts the MIM in BEiT from a single-choice classification problem to a multiple-choice one by softening the training objective from a hard-label cross-entropy loss to a soft-label one. BEiT performs the encoding and decoding role implicitly and simultaneously, while CAE [Chen *et al.*, 2022b] performs the two tasks explicitly and separately. A key component realizes this termed *latent contextual regressor* to introduce alignment between the representations of masked patches and unmasked ones. The CAE encoder *exclusively* focuses on feature extraction while the latent contextual regressor handles the prediction pretext task.

Better target tokenizer. PeCo [Dong *et al.*, 2021] identifies that the visual tokenizer generated by dVAE does not consider semantic level. PeCo adds the distance between deep visual features as an extra loss to enforce perceptual similarity between the original image and the reconstructed image to make the target visual tokens more semantically meaningful. For studying masked prediction, [Wei *et al.*, 2022a] follows the two-stage approach as BEiT and investigates various target tokenizers. Interestingly, it is found that handcrafted HOG features achieve a competitive performance, suggesting a target tokenizer generated by dVAE might be unnecessary.

4.2 End-to-End Masked Autoencoder

A drawback of the two-stage methods is that their approach relies on a pretrained dVAE to generate originally continuous but *intentionally discretized* target visual tokens [Yi *et al.*, 2022], and thus is not end-to-end. In essence, BEiT separates masked prediction from autoendoer training, which leaves room for improving effectiveness and efficiency. To this end, MAE [He *et al.*, 2022] experiments with end-to-end training of masked autoencoder. We highlight that SimMIM [Xie *et al.*, 2022b] has conducted a very similar investigation. MAE and SimMIM appear on arXiv concurrently and are both accepted at CVPR'2022. Here, we summarize these two seminal works and compare their nuanced difference.

MAE. The overview of MAE [He *et al.*, 2022] is shown in Figure 2. MAE revisits the pretext task of predicting masked patches. Specifically, their proposed MAE [He *et al.*, 2022] directly predicts masked patches from the unmasked ones with a simple loss of mean squared error (MSE). Moreover, the masking ratio is set to 75%, which is significantly

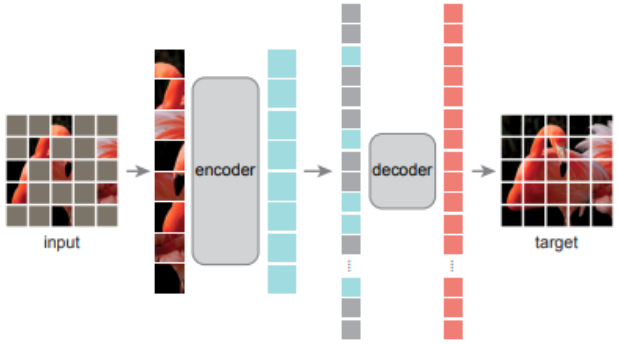

Figure 2: Overview of a masked autoencoder with the figure borrowed from the original work MAE [He *et al.*, 2022].

Model	Finetuning	Segmentation
MoCO v3 [Chen *et al.*, 2021]	83.2	47.3
DINO [Caron *et al.*, 2021]	82.8	46.8
BEiT [Bao *et al.*, 2022]	83.2	45.6
MAE [He *et al.*, 2022]	**83.6**	**48.1**

Table 2: Performance comparison of SOTA methods: discriminative methods (MoCo V3 and DINO) and generative methods (BEiT and MAE). The backbone architecture is ViT-B. Finetuning classification accuracy is measured on ImageNet-1K, downstream segmentation task is measured with mIoU on ADE20K.

higher than that in BERT (typically 15%) or prior MIM (20% to 50%) [Chen *et al.*, 2020a; Dosovitskiy *et al.*, 2021; Bao *et al.*, 2022]. The ablation findings support such a high masking ratio is beneficial for fine-tuning and linear probing. To save computation, the encoder of MAE only operates on the unmasked patches. Moreover, the encoder-decoder architecture is designed in an asymmetric manner with the decoder being lightweight. With the above technical tricks, their proposed simple MAE is (3× or more) faster than BEiT while achieving superior performance. A performance comparison of the state-of-the-art methods is given in Table 2.

SimMIM. Concurrently, a similar architecture termed Simple Masked Image Modeling (SimMIM) is proposed in [Xie *et al.*, 2022b], where similar findings are reported. Specifically, SimMIM confirms that directly predicting the pixels as in MAE performs no worse than other methods with complex design, such as tokenization, clustering, or discretization. A high masking ratio is also confirmed to be beneficial for performance, especially for a relatively small patch size. Moreover, SimSIM investigates multiple masking strategies, such as square, block-wise, and random. Their best performance is achieved with the random masking strategy, which is the same as that in MAE.

Difference between MAE and SimMIM. One of their non-trivial differences lies in the position of masked patch tokens. Specifically, masked patch tokens are adopted as the input of decoder and decoder in MAE [He *et al.*, 2022] and SimMIM [Xie *et al.*, 2022b], respectively. With the pretext task of masked prediction, the autoencoder in MAE and SimMIM fulfills two roles: representation encoding (for unmasked patches) and pretext prediction (for masked patches). With both masked and unmasked patches as the input, the encoder of SimMIM [Xie *et al.*, 2022b] simultaneously performs representation encoding and pretext prediction, due to which the decoder can be designed as simple as a single layer. By contrast, the encoder in MAE [He *et al.*, 2022] exclusively realizes representation encoding, leaving the role of pretext prediction to the decoder. As a result, MAE still relies on a transformer decoder, as reported in [He *et al.*, 2022], even though it does not need to be as heavy as the encoder. Due to this, MAE achieves significantly higher linear probing accuracy than SimMIM; however, this superiority diminishes with finetuning. For example, with ViT-B as the backbone on ImageNet, SimMIM achieves a finetuning performance of 83.8%, slightly higher than the reported 83.6% for MAE. Another merit of MAE by feeding only the unmasked patches into the encoder is its higher efficiency, especially when the masking ratio is high. Unlike SimMIM with Swin-B as the default backbone, MAE is not compatible with hierarchical ViT (like Swin). The reason for its incompatibility and solutions to address them are discussed in the following.

4.3 Towards Improving Efficiency

A significant bottleneck of masked autoencoder for visual SSL is that it requires large computation. In this section, we introduce multiple works that attempt to improve the efficiency of masked autoencoders from roughly two perspectives: (1) hierarchical structure and (2) input manipulation.

Hierarchical structure. Since ViT [Dosovitskiy *et al.*, 2021] used in MAE has a crucial issue that decreasing the patch size will quadratically increase computing resources, hierarchical ViT (hViT) was introduced by using a shrinking pyramid structure with additional tricks, e.g., Swin and PVT. Specifically, Swin adopts shifted windows to learn local feature correlations, and PVT applies spacial reduction attention to reduce computation in the attention layer. Unfortunately, it is not intuitive to adapt hViT to enable MAE pre-training since the local window attention used in hViT is challenging to handle randomly masked patches as in MAE. Multiple works [Huang *et al.*, 2022; Li *et al.*, 2022b; Zhang *et al.*, 2022b] attempt to improve MAE by boosting hViT, achieving comparable performance to the baselines (MAE, SimMIM) while requiring less training time as well as less GPU memory. Based on Swin transformer, [Huang *et al.*, 2022] proposes a unique masking strategy called group window attention, and combines the multi-scale feature learnability of hViT and the efficiency of masked image modeling by making them compatible. Similarly, Uniform Masking MAE (UM-MAE) [Li *et al.*, 2022b] introduced a two-stage sampling and masking process. The proposed Uniform Masking strategy first uniformly samples a quarter (25%) of patches in each block, then further masks random patches on top of the sampled patches. HiViT [Zhang *et al.*, 2022b] proposes a new hViT architecture to substitute window attention layers in Swin with MLP layers.

Input manipulation. Several methods attempt to improve the efficiency of MAE by changing the input. Specifically, they aim to reduce the input size by attending to small win-

dows [Chen *et al.*, 2022a] or objects in the image [Wu and Mo, 2022]. These methods reduce the required computation while achieving comparable or better downstream task performance. Local masked reconstruction (LoMaR) [Chen *et al.*, 2022a] is inspired from the fact that local information is enough for reconstructing masked patches. Instead of relying on the entire image for mask reconstruction, a number of small windows with 7x7 patches are sampled to restrict attention to local regions. LoMaR achieves higher downstream task performance faster compared with MAE. ObjMAE [Wu and Mo, 2022] achieves input efficiency by dropping non-object patches and learning object-wise representations. ObjMAE reduces the pre-training compute cost by 72% while achieving comparable performance to MAE. MixMIM [Liu *et al.*, 2022] takes a slightly different approach: to replace an image's masked tokens with tokens from another image. The mixed image is then fed into a encoder then the decoder reconstructs the two original images. Because of the absence of uninformative masked tokens, MixMIM [Liu *et al.*, 2022] is not only able to be suitable for hierarchical ViTs such as Swin but also achieves stronger results efficiently compared to existing MIM works.

5 Various Perspectives on the Success of Masked Autoencoder in Vision

To explain why BEIT [Bao *et al.*, 2022] helps the finetuning on downstream tasks, its authors analyze the self-attention map and show that BEiT distinguishes semantic regions using self-attention heads without any task-specific supervision. Moreover, [He *et al.*, 2022] shows that an MAE, pretrained with a masking ratio of 75%, infers complex and holistic reconstructions even when 95% of pixels are masked, suggesting it learns various concepts, *i.e.*, semantics. The authors of MAE [He *et al.*, 2022] "hypothesize that this behavior occurs through a rich hidden representation inside the MAE". Given that the masked and reconstructed visual patches are not semantic entities as words in languages, this behavior is somewhat unexpected and is hypothesized to occur "by way of a rich hidden representation" [He *et al.*, 2022]. However, which component in masked autoencoder makes the model learn such a "rich hidden representation" remains unclear. Numerous works have investigated from various perspectives for a better understanding of its success.

Backbone perspective: Is masked autoencoder compatible with CNN? With ViT [Dosovitskiy *et al.*, 2021] as the default backbone in MAE, a natural question is whether masked autoencoder works only with a transformer backbone instead of CNN. Since CNN cannot tackle the masked inputs and positional embedding directly, multiple works [Fang *et al.*, 2022b; Li *et al.*, 2022a; Fang *et al.*, 2022a] have attempted to unify ViT and CNN in a compatible masked autoencoder framework. Inspired by the observation that early convolutions help transformers see better [Xiao *et al.*, 2021], ConvMAE utilizes hybrid convolution-transformer architectures: convolution blocks at early stages and transformer blocks at later stages are in charge of high-resolution token embedding and low-resolution token embedding, respectively. Towards a unified framework of MIM with both transformer and CNN architecture, [Fang *et al.*, 2022a] proposes corrupted image modeling (CIM), which replaces the input images artificially masked in MIM with a corrupted image generated by a trainable generator (BEiT). Therefore, the reconstruction task in MIM can be extended to either generative or discriminative objectives trained by a ViT or CNN enhancer. CIM is the first to unify ViT and CNN in a non-Siamese framework and yields compelling results in vision benchmarks. More recently, it has been highlighted in [Li *et al.*, 2022a] that the success of masked image modeling can be agnostic to the architecture. The proposed Architecture Agnostic Masked Image Modeling framework (A^2MIM) is compatible with ViT and CNN in a unified way [Li *et al.*, 2022a]. It is found in [Li *et al.*, 2022a] that the success of masked autoencoder lies in learning middle-level patch interaction, which is agnostic to architecture choices.

Data perspective: Does masked autoencoder require a very large dataset? A popular belief regarding the benefit of transfer learning comes from pretraining on a much larger dataset than the target dataset. Challenging this belief, [El-Nouby *et al.*, 2021] investigates whether self-supervised pretraining on a smaller dataset can yield the same benefit. The fact that their investigation is performed with ViT-based masked autoencoder makes it more interesting because, compared with its CNN, ViT is found to require much more samples [Dosovitskiy *et al.*, 2021]. Interestingly, [Dosovitskiy *et al.*, 2021] shows that pretraining masked autoencoder (either BEiT or SplitMask [El-Nouby *et al.*, 2021]) on 1% of ImageNet dataset achieves comparable transfer performance to the iNaturalist-2019 dataset as pretraining on full ImageNet dataset. By contrast, prior DINO [Caron *et al.*, 2021] is much more sensitive to the data size (as well as the data type). More recently, [Xie *et al.*, 2022c] performed a comprehensive study on data scaling (from 10% of ImageNet to full ImageNet-22K) on masked autoencoder models of various sizes ranging from 49 million to 1 billion parameters. It shows that MIM is also demanding on larger data, especially for larger models with longer training epochs [Xie *et al.*, 2022c].

Denoising perspective: Does masked autoencoder benefit from other corruptions? Given that masked autoencoder is a class of denoising autoencoder, [Tian *et al.*, 2022] investigates a general question: are there other effective image degradation methods beyond masking for effective visual pretraining? Five methods, namely zoom-in, zoom-out, distortion, blurring, and de-colorizing, have been investigated, and they are found to perform better than None (*i.e.*, no pretraining), suggesting a unified denoising perspective on the success of masked autoencoder. Nonetheless, blurring and de-colorizing perform worse than other degradation methods with spatial transformation because they cause image style shift from the pretext task to the downstream task. Among them, zoom-in performs the best and is complementary with masking to further boost the performance. In contrast to existing spatial masking, [Xie *et al.*, 2022a] also investigates frequency masking by predicting masked high-frequency from the unmasked low-frequency content, or vice versa, demonstrating competitive performance. Moreover, super-resolution, deblur, and denoise have also been inves-

tigated but they yield inferior performance.

Theoretical perspective: Can masked autoencoder be explained with rigorous mathematics? Towards a mathematical understanding, [Cao *et al.*, 2022] was the first to propose a unified theoretical framework for understanding masked autoencoder in vision. Particularly, each image's embedding in MAE can be interpreted not as a 2D pixel grid but as a learned basis function in certain Hilbert spaces. Moreover, under a non-overlapping domain decomposition setting, the patch-based attention in ViT can be understood from the operator theoretic perspective of an integral kernel. With attention as the focus, [Cao *et al.*, 2022] further proves that the stability of internal representations and that of masked latent representations are interpolated globally with an inter-patch topology. To understand why MAE helps in downstream tasks, based on an autoencoder of a two/one-layered CNN, [Pan *et al.*, 2022] theoretically shows that it can capture all discriminative semantics in the pretraining dataset, and therefore provably outperforms supervised pretraining on downstream tasks.

6 Masked Autoencoder and Joint-Embedding

Before the success of masked autoencoder, visual self-supervised pretraining had been dominated by joint-embedding methods, either contrastive ones ([Chen *et al.*, 2021]) or negative-free ones [Caron *et al.*, 2021]. Thus, it is highly relevant to compare masked autoencoder with joint-embedding for visual self-supervised pretraining.

6.1 Boosting Each Other

An intriguing observation regarding their difference is as follows: compared with joint-embedding methods [Chen *et al.*, 2021; Caron *et al.*, 2021], masked autoencoders [He *et al.*, 2022; Xie *et al.*, 2022b] have stronger finetuning performance on the downstream tasks but weaker linear probing accuracy. A popular understanding is that masked autoencoder lacks in learning semantically-meaningful features because it focuses on low-level patch match with a local loss [He *et al.*, 2022; Xie *et al.*, 2022b]. On the other hand, high-level semantic features have the property of being robust to spatial transformation (like random crop) and style change (like color jittering) [Misra and Maaten, 2020], and thus joint embedding approaches adopt a global loss on the features after global average pooling to encourage the learned representation to be augmentation-invariant.

Improving masked autoencoder with global loss. SplitMask [El-Nouby *et al.*, 2021] consists of three steps: split, inpaint, and match. The patches are divided into two disjoint subsets in the split step: $\mathcal{A}$ and . For inpainting, it adopts a similar architecture as MAE in that a lightweight (shallow) ViT decoder is used to recover the masked patches from the representation of unmasked patches [El-Nouby *et al.*, 2021]. What differentiates SplitMask [El-Nouby *et al.*, 2021] from MAE [He *et al.*, 2022] lies in the third match step, which encourages the global prediction of $\mathcal{A}$ and $\mathcal{B}$ subsets of patches to match each other. This global match aligns with the augmentation-invariant goal in joint-embedding approaches, thus making the representation more semantically meaningful. [Tao *et al.*, 2022] improves MAE by combining it with joint-embedding approaches. Specifically, it predicts the masked tokens to match those from another augmented view to encourage semantic learning with an global loss.

Improving joint-embedding methods with local loss. Multiple works in the above analysis show that the global loss in joint-embedding methods can be utilized to improve the semantic meaning of the learned representations. Intuitively, it is possible to improve the joint-embedding techniques by adding a local loss. For example, MST [Li *et al.*, 2021] extends the DINO framework by combining it with a masked prediction task. It is worth mentioning that MST [Li *et al.*, 2021] came out earlier than BEiT and MAE. More recently, RePre [Wang *et al.*, 2022a] improves MoCO v3 [Chen *et al.*, 2021] with a reconstruction loss by using a decoder to reconstruct the original image from the multi-hierarchy features in the encoder. [Wei *et al.*, 2022b] shows that their inferior finetuning performance can be significantly improved by a simple post-processing with feature distillation (FD). After FD, their representations are more suitable for optimization and thus finetuning friendly.

6.2 Bridging Their Gap

Masked autoencoder and joint-embedding perform masked prediction (predicting a property of masked patches from unmasked patches) and augmented alignment (aligning the embedded representation of different augmentations), respectively. From the perspective of the architecture component, the encoder training in masked autoencoder relies on a decoder, while that in joint-embedding uses a Siamese encoder for generating the self-supervision. Motivated by their success, multiple works have attempted masked prediction without a decoder, decoder-free MIM, which bridges the gap between joint-embedding and masked autoencoder for visual pretraining.

Decoder-free MIM. Beyond masked autoencoder, decoder-free MIM can be seen as another line of simplifying BEiT from two stages to single stage. To keep the patch-level visual context, ConMIM [Yi *et al.*, 2022] follows the principle of designing the training objective to be masked patch prediction as in [Bao *et al.*, 2022]. Specifically, resembling MoCo [He *et al.*, 2020; Chen *et al.*, 2021], ConMIM adopts a Siamese encoder, which is updated by the (student) encoder with EMA, as a teacher model to guide the training of the encoder. ConMIM [Yi *et al.*, 2022] feeds an unmasked image and a masked image of the same view into teacher and student encoders, respectively. The teacher encoder can be seen as a dynamic tokenizer as a static one in BEiT [Bao *et al.*, 2022]. Therefore, the embedded representations of masked patches are predicted to match the dynamic tokenizer corresponding to the same position [Yi *et al.*, 2022]. A similar teacher-student framework is adopted in MSN [Assran *et al.*, 2022] and data2vec [Baevski *et al.*, 2022]. In contrast to ConMIM [Yi *et al.*, 2022], MSN [Assran *et al.*, 2022] adopts a global loss to encourage learning semantic-aware representation. CNN-based MSN has also been investigated in [Jing *et al.*, 2022]. It has also been demonstrated in

data2vec [Baevski *et al.*, 2022] that this simple framework works well in the vision field and can be generalized to other data modalities, including speech and language. MSN [Assran *et al.*, 2022] works well for linear probing and few-shot learning but might be inferior to masked autoencoder for the finetuning performance on downstream tasks since patch-level visual context is discarded. To get the merits on both sides, iBOT [Zhou *et al.*, 2022] adopts two losses: a local loss to distill in-view patch tokens and another global loss to distill between cross-view [CLS] tokens, which makes the target patch tokens more semantically-meaningful.

7 Applications Beyond Pure Images

Inspired by the success of masked autoencoder for visual pretraining on images, numerous works have extended it to other data modalities, such as graph, audio, time series data, 3D medical image, point clouds, reinforcement learning, etc. However, due to space constraints, this work limits the scope of application to the image-related domain. Since masked autoencoder in pure image has been extensively covered in Sec. 4 and Sec. 6, we focus on its two advanced applications by combining image with extra information: video pretraining that combines temporal information and vision-language pretraining that combines language.

7.1 Video

Numerous works have applied SSL frameworks built on images to videos since videos are essentially a clip of sequential images. This trend is also observed after the success of masked autoencoders, with works in [Wang *et al.*, 2022b; Wei *et al.*, 2022a] and [Tong *et al.*, 2022; Girdhar *et al.*, 2022] applying videos to BEiT [Bao *et al.*, 2022] and MAE [He *et al.*, 2022] respectively.

BEiT-based development. To learn spatial and temporal priors of videos in a decoupled way, BEVT [Wang *et al.*, 2022b] proposes a two-stage solution that learns spatial representations with masked image modeling, then learns temporal representations with jointly masked image modeling and masked video modeling. VIMPAC proposes a different single-stage method, which includes a block-wise masking strategy for videos and augmentation-free contrastive learning loss to learn the global features. Both BEVT and VIMPAC rely on an external tokenizer which can be limited in compute-intensive video understanding scenarios. To avoid an external tokenizer, [Wei *et al.*, 2022a] proposes to replace the tokens with features and investigates five types of features, among which hand-crafted HOG is found to work effectively and efficiently.

MAE-based development. Multiple works [Tong *et al.*, 2022; Girdhar *et al.*, 2022] follow the architecture of MAE for simplicity and efficiency. With a similar model architecture to MAE, VideoMAE [Tong *et al.*, 2022] finds that it learns useful spatio-temporal structures with a very high masking ratio (90% to 95%) in tube masking strategy. Beyond video understanding for existing frames, [Gupta *et al.*, 2022] investigates masked visual modeling for future frame prediction. The gap between masked prediction for partial existing frames and full future frames is addressed by a variable masking ratio. OmniMAE [Girdhar *et al.*, 2022] extends MAE to a unified pre-training of image and video modalities with a single model, achieving competitive performance on both image and video recognition benchmarks.

7.2 Vision and Language

Prior to masked autoencoder, contrastive learning is a popular approach to learn language and vision representations jointly. Contrastive Language-Image Pre-training (CLIP) is a pioneering work that propose learning images with language as supervision in a contrastive manner and achieves competitive results compared to fully supervised baselines. To solve the sampling bias and requirement of paired image-text samples by contrastive learning, [Geng *et al.*, 2022] follows MAE and proposes to encode a flexible mixture of inputs, including image-text pairs and image-only inputs. Experimental results show that M3AE learns generalizable vision representations and unified information from images and languages. Moreover, [Lu *et al.*, 2022] presents a unified task-agnostic model that can perform various vision and language tasks without task-specific branches.

8 Open Issues

- Training masked autoencoders can be computationally expensive, which increases the demand on computation resources and is not Eco-friendly.
- Beyond vision, the performance of masked autoencoders needs to be more extensively investigated in diverse fields (see [Zhang *et al.*, 2022a]).

9 Conclusion

This survey is the first to review the progress of masked autoencoder for visual SSL. We summarize the early attempts of masked autoencoder in vision and its relation with masked language modeling. With a focus on the reviving success of masked autoencoder in unsupervised visual pretraining, we summarize and compare the seminal methods as well as those follow-up works to improve their efficiency. We provide insight into the success of masked autoencoder in vision from various perspectives, including backbone, data, denosing and theoretical perspectives. Moreover, we discuss the relationship between masked autoencoders and joint-embedding methods. Finally, we cover its two advanced applications: video pretraining and vision-language pretraining.

Acknowledgements

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea Government [Ministry of Science and ICT (MSIT)] (Artificial Intelligence Innovation Hub) under Grant 2021-0-02068 and in part by [No.RS-2022-00155911, Artificial Intelligence Convergence Innovation Human Resources Development (Kyung Hee University)] and in part by the National Research Foundation of Korea(NRF) grant funded by MSIT (No. RS-2023-00207816), and in part by a grant from Kyung Hee University (KHU-20222221).

References

[Assran *et al.*, 2022] 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.

[Baevski *et al.*, 2022] 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.

[Bao *et al.*, 2022] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *ICLR*, 2022.

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

[Caron *et al.*, 2021] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021.

[Chen and He, 2021] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021.

[Chen *et al.*, 2020a] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, 2020.

[Chen *et al.*, 2020b] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *OpenAI blog*, 2020.

[Chen *et al.*, 2021] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *ICCV*, 2021.

[Chen *et al.*, 2022a] Jun Chen, Ming Hu, Boyang Li, and Mohamed Elhoseiny. Efficient self-supervised vision pretraining with local masked reconstruction. *arXiv preprint arXiv:2206.00790*, 2022.

[Chen *et al.*, 2022b] 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.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL*, 2019.

[Dong *et al.*, 2021] 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.

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

[El-Nouby *et al.*, 2021] 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.

[Fang *et al.*, 2022a] Yuxin Fang, Li Dong, Hangbo Bao, Xinggang Wang, and Furu Wei. Corrupted image modeling for self-supervised visual pre-training. *arXiv preprint arXiv:2202.03382*, 2022.

[Fang *et al.*, 2022b] Yuxin Fang, Shusheng Yang, Shijie Wang, Yixiao Ge, Ying Shan, and Xinggang Wang. Unleashing vanilla vision transformer with masked image modeling for object detection. *arXiv preprint arXiv:2204.02964*, 2022.

[Geng *et al.*, 2022] Xinyang Geng, Hao Liu, Lisa Lee, Dale Schuurams, Sergey Levine, and Pieter Abbeel. Multimodal masked autoencoders learn transferable representations. *arXiv preprint arXiv:2205.14204*, 2022.

[Girdhar *et al.*, 2022] Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Omnimae: Single model masked pretraining on images and videos. *arXiv preprint arXiv:2206.08356*, 2022.

[Gupta *et al.*, 2022] Agrim Gupta, Stephen Tian, Yunzhi Zhang, Jiajun Wu, Roberto Martín-Martín, and Li Fei-Fei. Maskvit: Masked visual pre-training for video prediction. *arXiv preprint arXiv:2206.11894*, 2022.

[He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[He *et al.*, 2022] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

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

[Jing *et al.*, 2022] Li Jing, Jiachen Zhu, and Yann LeCun. Masked siamese convnets. *arXiv preprint arXiv:2206.07700*, 2022.

[Khan *et al.*, 2022] Salman Khan, Muzammal Naseer, et al. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 2022.

[Larsson *et al.*, 2016] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *ECCV*, 2016.

[Li *et al.*, 2021] 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. *NeurIPS*, 2021.

[Li *et al.*, 2022a] Siyuan Li, Di Wu, Fang Wu, Zelin Zang, Kai Wang, Lei Shang, Baigui Sun, Hao Li, Stan Li, et al. Architecture-agnostic masked image modeling–from vit back to cnn. *arXiv preprint arXiv:2205.13943*, 2022.

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

[Li *et al.*, 2022c] Xiaotong Li, Yixiao Ge, Kun Yi, Zixuan Hu, Ying Shan, and Ling-Yu Duan. mc-beit: Multi-choice discretization for image bert pre-training. *arXiv preprint arXiv:2203.15371*, 2022.

[Liu *et al.*, 2022] Jihao Liu, Xin Huang, Yu Liu, and Hongsheng Li. Mixmim: Mixed and masked image modeling for efficient visual representation learning. *arXiv preprint arXiv:2205.13137*, 2022.

[Lu *et al.*, 2022] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022.

[Misra and Maaten, 2020] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, 2020.

[Pan *et al.*, 2022] Jiachun Pan, Pan Zhou, and Shuicheng Yan. Towards understanding why mask-reconstruction pretraining helps in downstream tasks. *arXiv preprint arXiv:2206.03826*, 2022.

[Pathak *et al.*, 2016] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016.

[Tao *et al.*, 2022] Chenxin Tao, Xizhou Zhu, Gao Huang, Yu Qiao, Xiaogang Wang, and Jifeng Dai. Siamese image modeling for self-supervised vision representation learning. *arXiv preprint arXiv:2206.01204*, 2022.

[Tian *et al.*, 2022] Yunjie Tian, Lingxi Xie, Jiemin Fang, Mengnan Shi, Junran Peng, Xiaopeng Zhang, Jianbin Jiao, Qi Tian, and Qixiang Ye. Beyond masking: Demystifying token-based pre-training for vision transformers. *arXiv preprint arXiv:2203.14313*, 2022.

[Tong *et al.*, 2022] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022.

[Vincent *et al.*, 2008] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, 2008.

[Wang *et al.*, 2022a] Luya Wang, Feng Liang, Yangguang Li, Wanli Ouyang, Honggang Zhang, and Jing Shao. Repre: Improving self-supervised vision transformer with reconstructive pre-training. *arXiv preprint arXiv:2201.06857*, 2022.

[Wang *et al.*, 2022b] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. Bevt: Bert pretraining of video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14733–14743, 2022.

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

[Wei *et al.*, 2022b] Yixuan Wei, Han Hu, Zhenda Xie, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *arXiv preprint arXiv:2205.14141*, 2022.

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

[Xiao *et al.*, 2021] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. *NeurIPS*, 2021.

[Xie *et al.*, 2022a] Jiahao Xie, Wei Li, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Masked frequency modeling for self-supervised visual pre-training. *arXiv preprint arXiv:2206.07706*, 2022.

[Xie *et al.*, 2022b] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, 2022.

[Xie *et al.*, 2022c] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Yixuan Wei, Qi Dai, and Han Hu. On data scaling in masked image modeling. *arXiv preprint arXiv:2206.04664*, 2022.

[Yi *et al.*, 2022] Kun Yi, Yixiao Ge, Xiaotong Li, Shusheng Yang, Dian Li, Jianping Wu, Ying Shan, and Xiaohu Qie. Masked image modeling with denoising contrast. *arXiv preprint arXiv:2205.09616*, 2022.

[Zhang *et al.*, 2016] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016.

[Zhang *et al.*, 2022a] Chaoning Zhang, Chenshuang Zhang, Junha Song, John Seon Keun Yi, Kang Zhang, and In So Kweon. A survey on masked autoencoder for self-supervised learning in vision and beyond. *arXiv preprint arXiv:2208.00173*, 2022.

[Zhang *et al.*, 2022b] Xiaosong Zhang, Yunjie Tian, Wei Huang, Qixiang Ye, Qi Dai, Lingxi Xie, and Qi Tian. Hivit: Hierarchical vision transformer meets masked image modeling. *arXiv preprint arXiv:2205.14949*, 2022.

[Zhou *et al.*, 2022] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *ICLR*, 2022.