A Survey on Masked Autoencoder for Self-supervised Learning in Vision and Beyond

Chaoning Zhang, Chenshuang Zhang, Junha Song, John Seon Keun Yi, Kang Zhang, In So Kweon

Abstract—Masked autoencoders are scalable vision learners, as the title of MAE [1], which suggests that self-supervised learning (SSL) in vision might undertake a similar trajectory as in NLP. Specifically, generative pretext tasks with the masked prediction (e.g., BERT) have become a de facto standard SSL practice in NLP. By contrast, early attempts at generative methods in vision have been buried by their discriminative counterparts (like contrastive learning); however, the success of mask image modeling has revived the masking autoencoder (often termed denoising autoencoder in the past). As a milestone to bridge the gap with BERT in NLP, masked autoencoder has attracted unprecedented attention for SSL in vision and beyond. This work conducts a comprehensive survey of masked autoencoders to shed insight on a promising direction of SSL. As the first to review SSL with masked autoencoders, this work focuses on its application in vision by discussing its historical developments, recent progress, and implications for diverse applications.

Index Terms—Survey, Masked Autoencoder, Self-supervised Learning, Masked Image Modeling.

1 Introduction

Deep learning [2] has revolutionized artificial intelligence in the past decade. Early developments focused on the architecture design with scalable size like increasing model depth, from AlexNet [3] to VGG [4] and ResNet [5]. In recent years, the attention has gradually shifted from designing better models to solving the data-hungry issue in deep learning. For example, ImageNet [6] with more than one million labeled images has become a typical benchmark dataset for vision models, and vision transformer (ViT) [7] 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. Self-supervised learning (SSL) [8], [9], outperforming its supervised counterpart for visual pre-training, has attracted significant attention.

With the advent of contrastive SSL in 2018, joint-embedding methods have become a dominant visual pre-trainign framework; however, this status has been recently challenged by the success of a generative method termed masked image modeling (MIM) [10]. BEiT [10] adopts a mask-then-predict strategy to train the model with the target visual tokens generated by an off-the-shelf tokenizer. The tokenizer is pretrained by a discrete variational autoencoder (dVAE) [11], and therefore BEiT can be seen as a two-stage training of denoising autoencoder [12]. Furthermore, an end-to-end masked autoencoder in the vision is proposed in MAE [1], which has attracted unprecedented attention.

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 [12] predicted masked pixels from unmasked ones but was referred to as denoising autoencoder [12], [13]. A similar investigation was conducted again in 2016 with the task of image inpainting [14]. Its reviving success in recent MAE [1], outperforming joint-embedding methods, inspires numerous works to understand its success in vision and to apply it in various applications, such as video, point cloud, and graph.

The reason for high popularity of masked autoencoder in visual pretraining is that a similar generative SSL framework termed masked language modeling (like BERT [15]) has been widely used in NLP. In other words, the success of masked autoencoder in vision paves a path that SSL in vision"*may now be embarking on a similar trajectory as in NLP*" [1] by generative pretext task with masked prediction. Moreover, since NLP and computer vision are two dominant branches in modern AI, many researchers believe that masked autoencoder might be the future for SSL.

To this end, this work conducts a comprehensive survey of masked autoencoders in SSL. This survey covers its application with various data types; however, it focuses on understanding its reviving success in vision. Note that autoencoder-based masked prediction started to become a de facto standard practice in language understanding in 2018/2019 [15]; thus, it is less relevant to discuss it in the 2020s. Moreover, it is the success of masked autoencoder in vision that shows visual SSL can embark on the same path as that in language, which somewhat revolutionizes visual SSL and then inspires the investigation of masked autoencoder in a wide range of applications. With masked autoencoder in vision as the focus, this survey mainly contains three parts. (1) Sec. 3 summarizes its historical development and relation with masked language modeling; (2) Sec. 4 discusses the masked modeling principle in vision and the understanding of its success from various perspectives. (3) Sec. 5 summarizes its implications on pre-training in diverse applications beyond natural images. To facilitate discussion without ambiguity, we include a terminology section (*i.e.* Sec. 2) to discuss essential terms in this survey.

Message to the readers. This survey will be updated on

• *The authors are with KAIST.*
E-mail: chaoningzhang1990@gmail.com

a regular basis to reflect the dynamic progress of masked autoencoder in its development. Since masked autoencoder is a fast-evolving field, and we might not be able to grasp all recent development. Therefore, we encourage researchers to contact us to inform us with their new works, either published ones or arXiv ones, on this topic. Those new works will be included and discussed in the revised version.

2 Background and Terminology

Generative SSL *v.s.* discriminative SSL. 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) [16]. Discriminative SSL typically follows its supervised couterpart to design a discriminative loss. Without ground-truth labels, a discrimiantive pretext task can be designed as solving jigsaw puzzles [17] or predicting rotation [18]. Later, the trend of discriminative visual SSL shifts from such geometry-based prediction to joint-embedding emthods [19], [20], [21].

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 [12], [13]. Note that *denoising* 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 [22]. This predicted property can be the original input [22], handcrafted feature [23], or latent representation [24]. Since masked prediction is a form of denoising process and thus masked autoencoder can be seen as a form of general denoising autoencoder. In this work, we use MAE exclusively to refer to the method in [1] *not* as shorthand for masked autoencoder to avoid confusion.

Masked autoencoding *v.s.* masked modeling. Masked prediction can be used to both generative and discriminative modeling methods. However, the term masked *X* modeling, namely masked modeling on X-type data, often refers to the generative case, such as masked *language* modeling [15], masked *image* modeling [25], masked *point* modeling [26]. Motivated by its success in generative modeling, a few works [22], [24], [27] have also applied masked prediction in discriminative SSL frameworks, demonstrating competitive performance. In other words, masked modeling is not necessarily masked autoencoding. Take image data, for example, MSN [27] and data2vec [24] can be categorized as masked image modeling but not masked autoencoding since their model architectures are decoder-free. In this work, we still perceive BEiT [10] as a variant of masked autoencoder even though it decouples the pretext task of masked prediction from autoencoder training.

3 Masked Autoencoding: NLP to Vision

NLP and (computer) vision are two dominant research fields for artificial intelligence. Despite the difference in the data types and downstream tasks, vision and language communities have often inspired each other. Towards a unified understanding of language and image, it is interesting to ask whether they can adopt a similar backbone architecture and training strategy. For the backbone architecture, the advent of vision Transformer (ViT) in [7] and its application to various vision tasks demonstrate that Transformers [28] can serve as a unified backbone architecture for both language and vision. Numerous works have further attempted to bridge the gap in their SSL training strategies.

3.1 NLP and vision followed different SSL paths

Generative SSL in NLP. In NLP there exist two leading language models: GPT [29], [30], [31] and BERT [15]. They are both based on the transformer architecture but with notable differences [32]: 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 they can be both perceived to rely on masked prediction as the pretext task.

Discriminative SSL in vision. Before 2018, there was an active investigation of unsupervised visual pretraining with both generative and discriminative modeling. Jigsaw and rotation prediction is designed as the pretext task on the discriminative side, while inpainting and colorization have been actively investigated on the generative side (see Sec.3.2). Since 2018, joint-embedding methods, namely aligning the embedded representations of augmented views of the same image [21], have demonstrated substantial performance boost over prior generative methods. Contrastive learning [33], [34], 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, especially after the advent of MoCo [8] and SimCLR [9]. Negative-free (*i.e.* non-contrastive) joint-embedding methods have also been investigated [19], [35], [36], demonstrating comparable performance of contrastive learning methods. A unified perspective on contrastive learning and negative-free joint-embedding is extensively discussed in [19], [20].

3.2 Is generative SSL suitable for vision?

Very early attempts. Images have spatial and channel dimensions, and therefore we can either predict masked spatial patches from unmasked ones [12], [13], [14] or predict masked channels from unmasked ones [37], [38], [39], [40]. A standard autoencoder takes an image as the input and reconstructs it after the information passes through a low-dimensional bottleneck layer. Without corrupting the input, the encoder focuses on content compression instead of extracting semantically meaningful representations. Denoising autoencoder was proposed in [12], [13] to perform masked autoencoding in the spatial dimension by randomly masking some pixels. To make it a harder task to avoid learning only low-level representation, [14] proposed feature learning by inpainting, *i.e.* to fill in large missing areas of the image and thus prevent hints from nearby pixels. Later, [37],

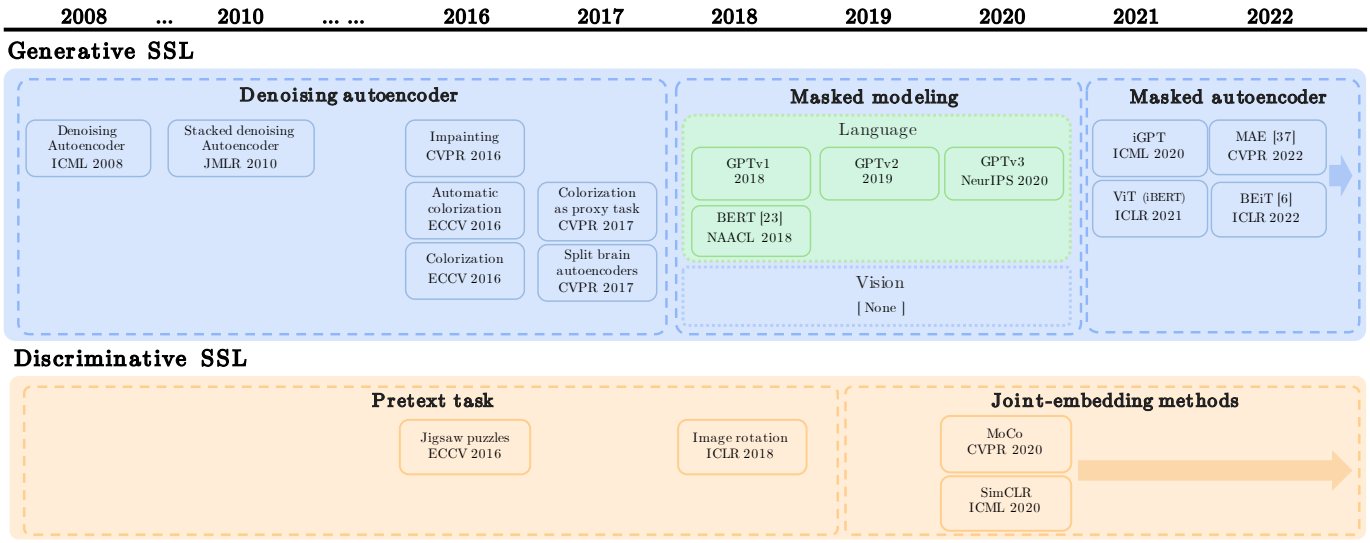

Fig. 1. Timeline of Visual SSL

[39] showed that masked channel prediction yielded superior performance on downstream tasks, especially for dense semantic segmentation, since it keeps the spatial content. They were further improved in [38], [40]. The investigation in this direction has been less active since the emergence of contrastive learning (see discussion in Sec.3.1).

Inspiration from NLP. The above investigation [12], [13], [14], [37], [38], [39], [40] was before 2017. 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 [41] is the first successful attempt in this direction; however, as highlighted in [42], 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. As the first attempt to replace CNN with a transformer in vision, [7] identified that the success of transformer in NLP tasks stems from excellent scalability and self-supervised pre-training. Since the self-supervised pre-training practice in [7] mimicked the masked language modeling task in BERT, we call it iBERT in analogy to iGPT entending GPT from language to vision. iBERT performs a masked patch prediction for visual SSL. However, this preliminary investigation of ViT for SSL also shows inferior performance over joint-embedding methods. This challenge was finally broken by BEiT [10] as well as MAE [1] (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 buried 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.

TABLE 1
Comparison of denosing autoencoder [12] and masked autoencoder [1]

	denoising autoencoder [12]	masked autoencoder [1]
Training dataset	MNIST	ImageNet
Model Architecture	CNN	ViT
Corruption size	pixels	patches
Corruption ratio	maximum 50%	patches

Remark. Early denoising autoencoder [12] and recent masked autoencoder [1] both attempt to reconstruct a clean input from a corrupted one, precisely predicting masked input content from unmasked input content. Despite high similarity regarding pretext task, the masked autoencoder introduced in [1] differs from early denoising autoencoder [12] 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 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.

4.1 BEiT and its improved variants

BEiT. The overview of BEiT is shown in Figure 2. In contrast to iBERT [7] that directly reconstructs the masked patches, BEiT mimicks BERT [15] 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) [11] 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 [43]. Despite its success, it remains unknown whether directly predicting masked image patches as in iBERT [7] might be a simpler alternative.

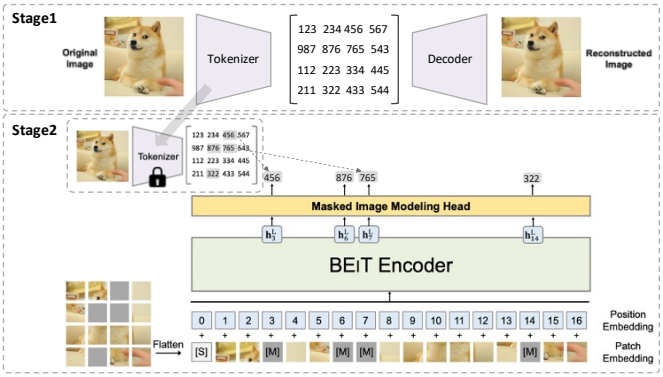

Fig. 2. Overview of BEIT pre-training. The figure is edited from [10].

BEiT [10] consists of two stages: token-based MIM as the main stage and tokenizer training as the preparation stage. Multiple works [44], [45], [46] have followed this two-stage approach by either improving the tokenizer-based MIM process or seeking an alternative tokenizer.

Tokenizer-based MIM. mc-BEiT [45] attempts to effectively utilize the visual tokenizer generated by dVAE. Specifically, it observes that unlike linguistic vocabulary consisting of discrete words, the image tokenizer is continuous. Under visual discretization, visual patches with similar semantics can have different token IDs, and visual patches with different semantics can have the token ID, which is not desired. 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. Following BEiT [10], CAE [46] first trains a image tokenizer via dVAE to generate target visual tokens. BEiT performs the encoding and decoding role implicitly and simultaneously, while CAE 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 without making predictions for masked patches. The CAE encoder exploits the full representation capability by letting the latent contextual regressor handle the prediction pretext task.

Better target tokenizer. PeCo [44] identifies that the visual tokenizer generated by dVAE [11] 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, [23] follows the two-stage approach as BEiT and investigates various target tokenizers. Interestingly, it is found that handcrafted HOG features [47] achieve a competitive performance, suggesting a target tokenizer generated by dVAE might be unnecessary. However, HOG is only compatible with visual data and limits its applications in other data modalities.

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 [22], 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 [1] experiments with end-to-end training of masked autoencoder. We highlight that SimMIM [25] has conducted a very similar investigation. MAE and SimMIM appear on arXiv concurrently (MAE being one week earlier) and are both accepted at CVPR'2022. Here, we summarize the two works and compare their nuanced difference.

MAE. The overview of MAE [1] is shown in Figure 3. MAE revisits the pretext task of predicting masked patches. Specifically, their proposed MAE [1] 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 higher than that in BERT (typically 15%) [15] or prior MIM (20% to 50%) [7], [10], [41]. The ablation findings support such a high masking ratio is beneficial for fine-tuning and linear probing. It is worth mentioning that this also motivates a recent work to experiment with a higher masking rate in masked language modeling for higher effectiveness [48]. To save computation, the encoder of MAE only operates on the unmasked patches. Moreover, MAE designs an asymmetric encoder-decoder architecture with a lightweight decoder. With the above technical tricks, their proposed simple MAE is (3 $\times$ or more) faster than BEiT [10] while achieving superior performance.

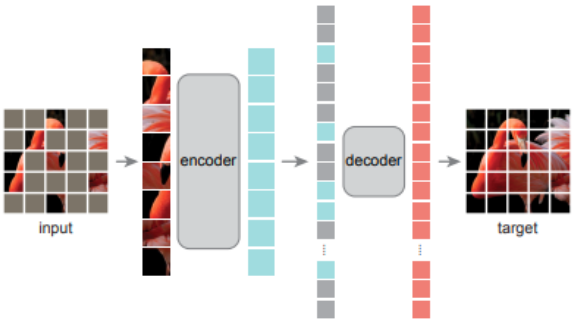

Fig. 3. Overview of a masked autoencoder with the figure borrowed from the original work MAE [1].

SimMIM. Independently and concurrently, a similar architecture termed Simple Masked Image Modeling (SimMIM) is proposed in [25], 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. It is also found that moderately increasing the patch size (32, for instance) is beneficial for a more powerful pretext task. A high masking ratio is also confirmed in MAE to be helpful for performance, especially for a relatively small patch size. Moreover, as shown in Figure 4, 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.

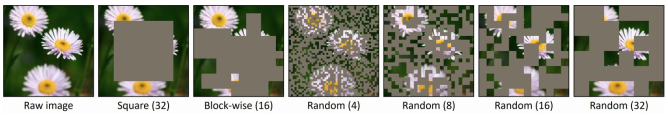

Fig. 4. Various masking strategies in SimMIM with the figure borrowed from the original paper [25].

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 [1] and SimMIM [25], 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 [25] 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 [1] 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 [1], 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 [49], [50]). The reason for its incompatibility and solutions to address them are discussed in the following.

4.3 Towards improving efficiency

Despite the impressive performance, 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 [7] used in MAE has a crucial issue that decreasing the patch size will quadratically increase computing resources, hierarchical ViT (hViT) [49], [50] was introduced. Specifically, Swin and PVT [49], [50] use a shrinking pyramid structure with additional tricks such as shifted windows [50] to learn local feature correlations or spacial reduction attention [49] to reduce computation in the attention layer used to further improve performance. 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.

Several works [51], [52], [53] attempt to boost the power of hViTs while achieving efficiency in MAE. Huang et al. [51] present a unique masking strategy called group window attention that gathers unmasked patches into several equal-sized groups to perform masked attention. Their method, based on a Swin transformer, combines the multi-scale feature learnability of hViT and the efficiency of masked image modeling by making the hierarchical transformer compatible with MAE. Similarly, Uniform Masking MAE (UM-MAE) [52] 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. The first step maintains similar elements across the non-overlapped local windows, while the second step makes the self-supervisory task more challenging by avoiding shortcuts for pixel reconstruction from neighboring low-level features. HiViT [53] proposes a new hViT architecture to substitute window attention layers in Swin [50] with MLP layers which enables masking as in MAE. The above works [51], [52], [53] achieve comparable performance to the baselines (MAE, SimMIM) while requiring less training time as well as less GPU memory.

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 windows [54] or objects in the image [55]. These methods reduce the required computation while achieving comparable or better downstream task performance.

Local masked reconstruction (LoMaR) [54] 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. It excels on high-resolution images since the required compute increase linearly with image size where it is quadratic for MAE. ObjMAE [55] achieves input efficiency by dropping non-object patches and learning object-wise representations. A class activation map (CAM) [56] is used to identify a rough object region, and the object regions are masked and used as input for the MAE. ObjMAE reduces the pre-training compute cost by 72% while achieving comparable performance to MAE, which masks the whole scene. MixMIM [57] 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, [57] is not only able to be suitable for hierarchical ViTs such as Swin [50] but also achieves stronger results efficiently compared to existing MIM works.

4.4 Various perspectives on the success of masked autoencoder in vision

To explain why BEIT [10] 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, [1] 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 [1] "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" [1]. 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 [7] as the default backbone in MAE [1], 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 [58], [59], [60], [61] have attempted to unify ViT and CNN in a compatible masked autoencoder framework. Inspired by the observation that early convolutions help transformers see better [62], ConvMAE [58] 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, [61] 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 [60] 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 [60]. It is found in [60] that the success of masked autoencoder lies in learning middle-level patch interaction, which is agnostic to architecture choices. Early attempts of CNN-based inpainting [14] resembles masked autoencoder but focuses on reconstruction task with low-level interactions, which causes higher feature uncertainty [60].

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, [63] 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 [7]. Interestingly, [7] shows that pretraining masked autoencoder (either BEiT or SplitMakk [63]) on 1% of ImageNet dataset achieves comparable transfer performance to the iNaturalist-2019 dataset as pretraining on full ImageNet dataset. By contrast, prior DINO [43] is much more sensitive to the data size (as well as the data type). More recently, [64] 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 [64]. Beyond size, some works have also investigated domain issues in data and found that they can be alleviated by training the masked autoencoder on images of mixed-style [65] or multi-tasks [66].

Denoising perspective: Does masked autoencoder benefit from other corruptions? Given that masked autoencoder is a class of denoising autoencoder, [67] 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, [68] 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 investigated but they yield inferior performance.

Theoretical perspective: can masked autoencoder be explained with rigorous mathematics? Towards a mathematical understanding, [69] 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, [69] further proves that the stability of internal representations and that 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, [70] theoretically shows that it can capture discriminative semantics in the pretraining dataset. Deviating from the focus on attention [69], it provides insight on what features MAE learns and why MAE beats conventional supervised learning (SL). Particularly, MAE encoder captures all discriminative semantics in the pretraining dataset, including samples that have either single or multiple independent discriminative semantics, and therefore provably outperforms SL on downstream tasks.

4.5 Relationship with joint-embedding methods

Before the success of masked autoencoder, visual self-supervised pretraining had been dominated by joint-embedding methods, either contrastive ones ([9], [71]) or negative-free ones [43], [72]. Thus, it is highly relevant to compare masked autoencoder with joint-embedding for visual self-supervised pretraining.

4.5.1 Masked autoencoder and joint embedding: boosting each other

An intriguing observation regarding their difference is as follows: compared with joint-embedding methods [43], [71], masked autoencoders [1], [25] 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 [1], [25]. 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) [73], 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 [63] 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 [63]. What differentiates SplitMask [63] from MAE [1] 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. [74] 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 optional 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 [75] extends the DINO framework by combining it with a masked prediction task. It is worth mentioning that MST [75] came out earlier than BEiT and MAE. More recently, RePre [76] improves MoCO v3 [71] with a reconstruction loss by using a decoder to reconstruct the original image from the multi-hierarchy features in the encoder. [77] shows that their inferior fine-tuning 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.

4.5.2 Masked autoencoder and joint embedding: 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 [22] follows the principle of designing the training objective to be masked patch prediction as in [10]. Specifically, resembling MoCo [8], [71], 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 [22] 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 [10]. Therefore, the embedded representations of masked patches are predicted to match the dynamic tokenizer corresponding to the same position [22]. A similar teacher-student framework is adopted in MSN [27] and data2vec [24]. In contrast to ConMIM [22], MSN [27] adopts a global loss to encourage learning semantic-aware representation. CNN-based MSN has also been investigated in [78]. It has also been demonstrated in data2vec [24] that this simple framework works well in the vision field and can be generalized to other data modalities, including speech and language. MSN [27] 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 [79] 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 [79]. More recently, AttMask [80] shows that iBOT can be further improved by performing an attention-guided masking instead of random masking on the student side. Particularly, the teacher model indicates the attention with full image as the input and masking on the attended areas improves the performance on a variety of downstream tasks.

Collapse issue. A shared issue in the above decoder-free MIM methods [22], [24], [27] is the potential feature collapse, *i.e.* outputting a constant for all inputs. They adopt different approaches to avoid this issue. For example, ConMIM [22] adopts a contrastive loss with those feature representations corresponding to different positions in the same image as negative samples. MSN [27] follows [81] to do cluster assignments, while data2vec [24] achieves this goal by carefully fine-tuning the hyperparameters like momentum coefficient in EMA and learning rate. Note that autoencoder-based MIM methods do not have the collapse issue by default.

5 Other Applications: Vision and Beyond

Inspired by the success of MAE [1], numerous works have applied masked autoencoder to various applications. We categorize them into two classes. The first class is related to vision, for which pure natural images have been extensively covered in the above section. Thus, this section covers its other aspects, including images with medical applications, images with temporal information, images with language. Going beyond vision, the second class focuses on different

types of data, such as point clouds, graph, audio, reinforcement learning, etc.

5.1 Vision related applications

5.1.1 Medical images

Medical images are a class of data for medical analysis with data distribution different from natural images. Multiple works have shown that masked autoencoders can also work well in medical applications by either applying MAE directly to medical data [82], [83] or improving the loss design [84], [85], [86], [87]. With MAE [1] and SimMIM [25] as the architecture, [82] and [83] apply masked autoencoders directly in medical images, showing the effectiveness of masked autoencoder in medical applications, *e.g.* CT images. Specifically, [82] shows that MAE pretrained on medical dataset achieves superior performance to its counterpart pretrained on ImageNet, which can be explained from the perspective domain shift. [83] shows that a moderately large patch size (32) achieves satisfactory performance, which aligns with the finding in [25]. There are also attempts to enhance MAE [1] by improving the loss [84], [85], [86], [87], including global loss and self-distillation loss. It is shown in [84], [85] that an additional global loss on top of a local loss makes the representations more semantically meaningful for medical images, which resembles the principle in iBOT and SplitMask. [86] views the output of MAE encoder as a bag of instances and aggregates the most informative tokens into global representation (slide-level) for further classification. To make full use of visible patches, Self-distillation MAE(SD-MAE) [87] improves MAE by adding a self-distillation loss of visible patches between latent representations after encoding and decoding, which achieves competitive performance compared with contrastive methods.

5.1.2 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 [23], [88], [89] and [90], [91] applying videos to BEiT [10] and MAE [1] respectively.

To learn spatial and temporal priors of videos in a decoupled way, BEVT [88] 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. BEVT achieves comparable or superior results to baseline methods on three video datasets. VIMPAC [89] proposes a different single-stage masked video modeling method, which includes a block-wise masking strategy for videos and augmentation-free contrastive learning loss to learn the global features. Experimental results verify the effectiveness and scalability of the proposed VIMPAC. Both BEVT [88] and VIMPAC [89] rely on an external tokenizer which can be limited in compute-intensive video understanding scenarios. Therefore, [23] 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.

Since MAE is found to be more simple yet effective than BEiT, the works in [90], [91] follow the architecture of MAE for simplicity and efficiency. With a similar model architecture to MAE, VideoMAE [90] finds that it learns useful spatio-temporal structures with a very high masking ratio (90% to 95%) in tube masking strategy. Experimental results show that VideoMAE [90] achieves impressive performance on tiny datasets. Similar investigation has also been investigated in [92] but with spacetime-agnostic random masking. Beyond video understanding for existing frames, [93] 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 [91] extends MAE to a unified pre-training of image and video modalities. Trained on images and video with a single ViT encoder, OmniMAE achieves competitive performance on both image and video recognition benchmarks, outperforming models explicitly trained for a single modality.

5.1.3 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) [94] is a pioneering work that propose learning images with language as supervision. By jointly learning an image and text encoder, CLIP takes the pair of image and text as a prediction target during contrastive pre-training and often achieves competitive results compared to fully supervised baselines. Other works extend of CLIP by adding self-supervision [95], data scaling [96] or enabling flexibility to the encoders [97]. Contrastive learning introduces sampling bias due to data augmentations and cannot tackle unpaired samples [98]. To solve these problems, [98] proposes Multimodal Masked Autoencoder (M3AE), which encodes 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. To investigate how to design an effective vision-language model with an end-to-end manner, Multimodal End-to-end TransformER (METER) [99] implements comprehensive experiments and analyses on multiple designs, including encoders, multimodal fusion module, pre-training objectives. However, adding MIM loss does not improve downstream task performance in their settings [99]. [100] also investigates the masking strategies of text data in language-vision tasks, which improves performance on downstream tasks.

Moreover, [101] presents a unified task-agnostic model that can perform various vision and language tasks. It is able to tackle different tasks with a unified model without employing task-specific branches by tokenizing the inputs and outputs of every given task. A standard transformer encoder/decoder is pre-trained with masked language modeling and masked image modeling, then further trained on a large multi-task dataset that encompasses different language/vision tasks. Similarly, [102] proposes VL-BEiT that can tackle both monomodal and multimodal vision-language tasks. A single bidirectional multimodal transformer [97] is pre-trained on mask prediction of monomodal (language, vision) and multimodal (image-text pair) data

to be jointly optimized to different types of data. VL-BEiT achieves strong results on various vision-language benchmarks and image tasks.

5.2 Beyond vision

5.2.1 Point clouds

Motivated by the success of BEiT in vision, [26] extends masked modeling strategy to point cloud with masked point modeling termed Point-BERT. Following BEiT, Point-BERT first trains a discrete VAE to generate discrete point tokens containing meaningful local information and then predicts the tokens of masked point patches from the unmasked point patches. With a pure transformer architecture surpassing carefully designed point cloud models, Point-BERT achieves 93.8% accuracy on ModelNet40 and 83.1% accuracy on the complicated setting of ScanObjectNN, suggesting the BERT-style pre-training technique also works for point cloud. Point-BERT relies on dVAE, which is trained by augmentation-based contrastive learning and thus is sophisticated. Moreover, the masked tokens from their inputs are processed as the input of Transformers, causing high compute and early leakage of location information [103]. MaskPoint [104] alleviates this issue by pre-training with a decoder to contrast masked points and noise. Moreover, Point-MAE [103] follows MAE [1] to adopt a more straightforward approach to directly predict the locations of masked points. Resembling Point-MAE, [105] proposes Point-M2AE, a Multi-scale MAE for point clouds. Different from Point-MAE, Point-M2AE adopts an encoder-decoder with pyramid architectures to capture both fine-grained and high-level semantics in a progressive manner. Accordingly, Point-M2AE also adopts a multi-scale masking strategy to yield visible patches consistent across scales. Moreover, a local spatial self-attention mechanism is also adopted to make the encoder focus on neighboring patterns. [106] has proposed Voxel-MAE to pre-train on large-scale point clouds to improve downstream 3D object detection. The key idea lies in dividing the point clouds into voxel representations and classify whether they contain point clouds.

5.2.2 Graph

MGAE [107] is the first to investigate masked autoencoder for graphs. Prior to its advent, works on learning graph node representations in an unsupervised manner can be categorized into two classes: graph autoencoder and graph self-supervised learning (GSSL). They focus on designing effective encoder networks and advanced pretext tasks, respectively. Even though edge dropping and edge reconstruction are commonly adopted in both lines of investigations, masked autoencoding by recovering the masked edges from randomly masked input graph structure has never been explored until the advent of MGAE [107]. Following MAE [1], MGAE operates only on convolution-based partial network structure (without masked edges). Moreover, the decoder is designed to capture the cross-correlation between an anchor edge's head and tail nodes. MGAE performs better or on par with graph autoencoder and GSSL. GMAE [108] further investigates masked autoencoders for a graph with transformer instead of convolution. Another item that distinguishes them is that MGAE reconstructs masked edges, and GMAE reconstructs the features of masked nodes. [109] also argues that rebuilding the features is more beneficial. Beyond empirical results with experimental trial-and-error, MaskGAE [110] further provides theoretical justifications for the potential benefits of masked graph modeling.

5.2.3 Reinforcement learning

The auxiliary tasks and RL updates [111] are jointly trained in [112], where the performances of ViT models is compared to that of a CNN-based RL method [113]. The results indicate that CNN-based RAD [113] performs better on most image-based deep RL tasks, but reconstruction-based ViT models [1], [24] outperform RAD on some tasks. With ViT architecture, Xiao et. al [114] adopted pre-trained visual representations to train various motor control tasks. First, MAE [1] is used to learn visual representations from real-world images. Then, the encoder is freezed and the feature vector is used alongside propropceptive robot information to train task-specific motor cotrolling policies with model-free reinforcement learning [115]. The authors demonstrate that a single encoder can be used to learn various tasks without task-specific fine-tuning, and achieves superior performance compared to supervised baselines. Seo et. al [116] demonstrate a similar approach, but show that convolutional *feature* masking is more effective than pixel patch masking since it learns fine-grained features within patches.

5.2.4 Audio

Recent works learning audio representations create different input views by temporal relationships or data augmentations, which cannot provide information from the intact input. Inspired by the success of MAE in vision, [117], [118] apply masked autoencoders successfully on the masked audio spectrogram to learn audio representations from both time and frequency axes. Moreover, [118] is trained jointly on discriminative and generative loss for instance-wise classification and reconstruction, respectively.

5.2.5 More diverse applications

Masked autoencoder has also been attempted in more diverse applications. For example, [119] experiments with masked autoencoder with extrapolator (ExtraMAE) to recover complex original time series signals from masked observations. Recognizing contrastive tabular-SSL does not sufficiently capture the underlying manifold due to the ad-hoc fashion of its augmentation design, [120] proposes Masked Encoding for Tabular data (MET). With the MAE [1] in vision as the baseline, MET adopts individual representation for each coordinate with an additional adversarial loss by considering the property of tabular data. Some works [61], [121] have also adopted a pretrained masked autoencoder as an augmentation generator. For example, the reconstructed views from MAE are found to outperform hand-crafted augmentations (like scale, flip, and color jitter) in both supervised and semi-supervised setups [121].

6 Conclusion

This survey is the first to review the progress of masked autoencoder for SSL. We summarize the early attempts of

masked autoencoder in vision and its relation with masked language modeling. With the 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 them. We also provide insight on the success of masked autoencoder in vision from various perspectives, including backbone perspective, data perspective, denosing perspective and theoretical perspective. Finally, we summarize its application in vision and beyond. Preliminary summarization of the works in table format is provided in the appendix.

REFERENCES

[1] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *CVPR*, 2022. 1, 2, 3, 4, 5, 6, 7, 8, 9, 13

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, 2015. 1

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *NeurIPS*, 2012. 1

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015. 1

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016. 1

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *CVPR*, 2009. 1

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021. 1, 2, 3, 4, 5, 6

[8] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020. 1, 2, 7

[9] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020. 1, 2, 6

[10] H. Bao, L. Dong, and F. Wei, "Beit: Bert pre-training of image transformers," *ICLR*, 2022. 1, 2, 3, 4, 5, 7, 8, 13

[11] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *ICML*, 2021. 1, 3, 4

[12] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *ICML*, 2008. 1, 2, 3

[13] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, and L. Bottou, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion." *Journal of machine learning research*, 2010. 1, 2, 3

[14] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *CVPR*, 2016. 1, 2, 3, 6

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *NAACL*, 2019. 1, 2, 3, 4

[16] A. Ng *et al.*, "Sparse autoencoder," *CS294A Lecture notes*, 2011. 2

[17] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *ECCV*, 2016. 2

[18] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *ICLR*, 2018. 2

[19] C. Zhang, K. Zhang, C. Zhang, T. X. Pham, C. D. Yoo, and I. S. Kweon, "How does simsiam avoid collapse without negative samples? a unified understanding with self-supervised contrastive learning," in *ICLR*, 2022. 2

[20] C. Zhang, K. Zhang, T. X. Pham, C. Yoo, and I.-S. Kweon, "Dual temperature helps contrastive learning without many negative samples: Towards understanding and simplifying moco," in *CVPR*, 2022. 2

[21] L. Jing, P. Vincent, Y. LeCun, and Y. Tian, "Understanding dimensional collapse in contrastive self-supervised learning," *arXiv preprint arXiv:2110.09348*, 2021. 2

[22] K. Yi, Y. Ge, X. Li, S. Yang, D. Li, J. Wu, Y. Shan, and X. Qie, "Masked image modeling with denoising contrast," *arXiv preprint arXiv:2205.09616*, 2022. 2, 4, 7, 13

[23] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," in *CVPR*, 2022. 2, 4, 8, 13

[24] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," *arXiv preprint arXiv:2202.03555*, 2022. 2, 7, 9

[25] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "Simmim: A simple framework for masked image modeling," in *CVPR*, 2022. 2, 4, 5, 7, 8, 13

[26] X. Yu, L. Tang, Y. Rao, T. Huang, J. Zhou, and J. Lu, "Point-bert: Pre-training 3d point cloud transformers with masked point modeling," in *CVPR*, 2022. 2, 9

[27] M. Assran, M. Caron, I. Misra, P. Bojanowski, F. Bordes, P. Vincent, A. Joulin, M. Rabbat, and N. Ballas, "Masked siamese networks for label-efficient learning," *arXiv preprint arXiv:2204.07141*, 2022. 2, 7

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017. 2, 13

[29] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018. 2

[30] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, 2019. 2

[31] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, 2020. 2

[32] A. Bilogur, "Notes on gpt-2 and bert models," *Kaggle blog*, 2019. 2

[33] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *CVPR*, 2018. 2

[34] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018. 2

[35] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," *ICML*, 2021. 2

[36] X. Chen and K. He, "Exploring simple siamese representation learning," in *CVPR*, 2021. 2

[37] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *ECCV*, 2016. 2, 3

[38] ——, "Colorization as a proxy task for visual understanding," in *CVPR*, 2017. 2, 3

[39] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *ECCV*, 2016. 2, 3

[40] ——, "Split-brain autoencoders: Unsupervised learning by cross-channel prediction," in *CVPR*, 2017. 2, 3

[41] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever, "Generative pretraining from pixels," in *ICML*, 2020. 3, 4

[42] ——, "Generative pretraining from pixels," in *OpenAI blog*, 2020. 3

[43] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," *arXiv preprint arXiv:2104.14294*, 2021. 4, 6, 7

[44] X. Dong, J. Bao, T. Zhang, D. Chen, W. Zhang, L. Yuan, D. Chen, F. Wen, and N. Yu, "Peco: Perceptual codebook for bert pre-training of vision transformers," *arXiv preprint arXiv:2111.12710*, 2021. 4, 13

[45] X. Li, Y. Ge, K. Yi, Z. Hu, Y. Shan, and L.-Y. Duan, "mc-beit: Multi-choice discretization for image bert pre-training," *arXiv preprint arXiv:2203.15371*, 2022. 4

[46] X. Chen, M. Ding, X. Wang, Y. Xin, S. Mo, Y. Wang, S. Han, P. Luo, G. Zeng, and J. Wang, "Context autoencoder for self-supervised representation learning," *arXiv preprint arXiv:2202.03026*, 2022. 4, 13

[47] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005. 4

[48] A. Wettig, T. Gao, Z. Zhong, and D. Chen, "Should you mask 15% in masked language modeling?" *arXiv preprint arXiv:2202.08005*, 2022. 4

[49] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," 2021. 5

[50] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," 2021. 5

[51] L. Huang, S. You, M. Zheng, F. Wang, C. Qian, and T. Yamasaki, "Green hierarchical vision transformer for masked image modeling," *arXiv preprint arXiv:2205.13515*, 2022. 5, 13

[52] X. Li, W. Wang, L. Yang, and J. Yang, "Uniform masking: Enabling mae pre-training for pyramid-based vision transformers with locality," *arXiv preprint arXiv:2205.10063*, 2022. 5, 13

[53] X. Zhang, Y. Tian, W. Huang, Q. Ye, Q. Dai, L. Xie, and Q. Tian, "Hivit: Hierarchical vision transformer meets masked image modeling," *arXiv preprint arXiv:2205.14949*, 2022. 5, 13

[54] J. Chen, M. Hu, B. Li, and M. Elhoseiny, "Efficient self-supervised vision pretraining with local masked reconstruction," *arXiv preprint arXiv:2206.00790*, 2022. 5, 13

[55] J. Wu and S. Mo, "Object-wise masked autoencoders for fast pre-training," *arXiv preprint arXiv:2205.14338*, 2022. 5, 13

[56] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *CVPR*, 2016. 5

[57] J. Liu, X. Huang, Y. Liu, and H. Li, "Mixmim: Mixed and masked image modeling for efficient visual representation learning," *arXiv preprint arXiv:2205.13137*, 2022. 5, 13

[58] P. Gao, T. Ma, H. Li, J. Dai, and Y. Qiao, "Convmae: Masked convolution meets masked autoencoders," *arXiv preprint arXiv:2205.03892*, 2022. 6

[59] Y. Fang, S. Yang, S. Wang, Y. Ge, Y. Shan, and X. Wang, "Unleashing vanilla vision transformer with masked image modeling for object detection," *arXiv preprint arXiv:2204.02964*, 2022. 6

[60] S. Li, D. Wu, F. Wu, Z. Zang, K. Wang, L. Shang, B. Sun, H. Li, S. Li *et al.*, "Architecture-agnostic masked image modeling–from vit back to cnn," *arXiv preprint arXiv:2205.13943*, 2022. 6

[61] Y. Fang, L. Dong, H. Bao, X. Wang, and F. Wei, "Corrupted image modeling for self-supervised visual pre-training," *arXiv preprint arXiv:2202.03382*, 2022. 6, 9, 13

[62] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, "Early convolutions help transformers see better," *NeurIPS*, 2021. 6

[63] A. El-Nouby, G. Izacard, H. Touvron, I. Laptev, H. Jegou, and E. Grave, "Are large-scale datasets necessary for self-supervised pre-training?" *arXiv preprint arXiv:2112.10740*, 2021. 6, 7, 13

[64] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, Y. Wei, Q. Dai, and H. Hu, "On data scaling in masked image modeling," *arXiv preprint arXiv:2206.04664*, 2022. 6

[65] H. Yang, M. Chen, Y. Wang, S. Tang, F. Zhu, L. Bai, R. Zhao, and W. Ouyang, "Domain invariant masked autoencoders for self-supervised learning from multi-domains," *arXiv preprint arXiv:2205.04771*, 2022. 6

[66] R. Bachmann, D. Mizrahi, A. Atanov, and A. Zamir, "Multimae: Multi-modal multi-task masked autoencoders," *arXiv preprint arXiv:2204.01678*, 2022. 6, 13

[67] Y. Tian, L. Xie, J. Fang, M. Shi, J. Peng, X. Zhang, J. Jiao, Q. Tian, and Q. Ye, "Beyond masking: Demystifying token-based pre-training for vision transformers," *arXiv preprint arXiv:2203.14313*, 2022. 6, 13

[68] J. Xie, W. Li, X. Zhan, Z. Liu, Y. S. Ong, and C. C. Loy, "Masked frequency modeling for self-supervised visual pre-training," *arXiv preprint arXiv:2206.07706*, 2022. 6

[69] S. Cao, P. Xu, and D. A. Clifton, "How to understand masked autoencoders," *arXiv preprint arXiv:2202.03670*, 2022. 6

[70] J. Pan, P. Zhou, and S. Yan, "Towards understanding why mask-reconstruction pretraining helps in downstream tasks," *arXiv preprint arXiv:2206.03826*, 2022. 6

[71] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," *ICCV*, 2021. 6, 7

[72] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in Neural Information Processing Systems*, 2020. 6

[73] I. Misra and L. v. d. Maaten, "Self-supervised learning of pretext-invariant representations," in *CVPR*, 2020. 7

[74] C. Tao, X. Zhu, G. Huang, Y. Qiao, X. Wang, and J. Dai, "Siamese image modeling for self-supervised vision representation learning," *arXiv preprint arXiv:2206.01204*, 2022. 7, 13

[75] Z. Li, Z. Chen, F. Yang, W. Li, Y. Zhu, C. Zhao, R. Deng, L. Wu, R. Zhao, M. Tang *et al.*, "Mst: Masked self-supervised transformer for visual representation," *NeurIPS*, 2021. 7, 13

[76] L. Wang, F. Liang, Y. Li, W. Ouyang, H. Zhang, and J. Shao, "Repre: Improving self-supervised vision transformer with reconstructive pre-training," *arXiv preprint arXiv:2201.06857*, 2022. 7, 13

[77] Y. Wei, H. Hu, Z. Xie, Z. Zhang, Y. Cao, J. Bao, D. Chen, and B. Guo, "Contrastive learning rivals masked image modeling in fine-tuning via feature distillation," *arXiv preprint arXiv:2205.14141*, 2022. 7

[78] L. Jing, J. Zhu, and Y. LeCun, "Masked siamese convnets," *arXiv preprint arXiv:2206.07700*, 2022. 7

[79] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong, "ibot: Image bert pre-training with online tokenizer," *ICLR*, 2022. 7, 13

[80] I. Kakogeorgiou, S. Gidaris, B. Psomas, Y. Avrithis, A. Bursuc, K. Karantzalos, and N. Komodakis, "What to hide from your students: Attention-guided masked image modeling," *arXiv preprint arXiv:2203.12719*, 2022. 7, 13

[81] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *arXiv preprint arXiv:2006.09882*, 2020. 7

[82] L. Zhou, H. Liu, J. Bae, J. He, D. Samaras, and P. Prasanna, "Self pre-training with masked autoencoders for medical image analysis," *arXiv preprint arXiv:2203.05573*, 2022. 8, 13

[83] Z. Chen, D. Agarwal, K. Aggarwal, W. Safta, M. M. Balan, V. Sethuraman, and K. Brown, "Masked image modeling advances 3d medical image analysis," *arXiv preprint arXiv:2204.11716*, 2022. 8, 13

[84] S. T. Ly, B. Lin, H. Q. Vo, D. Maric, B. Roysam, and H. V. Nguyen, "Student collaboration improves self-supervised learning: Dual-loss adaptive masked autoencoder for brain cell image analysis," *arXiv preprint arXiv:2205.05194*, 2022. 8, 13

[85] H. Quan, X. Li, W. Chen, M. Zou, R. Yang, T. Zheng, R. Qi, X. Gao, and X. Cui, "Global contrast masked autoencoders are powerful pathological representation learners," *arXiv preprint arXiv:2205.09048*, 2022. 8, 13

[86] J. An, Y. Bai, H. Chen, Z. Gao, and G. Litjens, "Masked autoencoders pre-training in multiple instance learning for whole slide image classification," in *Medical Imaging with Deep Learning*, 2022. 8, 13

[87] Y. Luo, Z. Chen, and X. Gao, "Self-distillation augmented masked autoencoders for histopathological image classification," *arXiv preprint arXiv:2203.16983*, 2022. 8, 13

[88] R. Wang, D. Chen, Z. Wu, Y. Chen, X. Dai, M. Liu, Y.-G. Jiang, L. Zhou, and L. Yuan, "Bevt: Bert pretraining of video transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 733–14 743. 8, 13

[89] H. Tan, J. Lei, T. Wolf, and M. Bansal, "Vimpac: Video pre-training via masked token prediction and contrastive learning," *arXiv preprint arXiv:2106.11250*, 2021. 8

[90] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *arXiv preprint arXiv:2203.12602*, 2022. 8, 13

[91] R. Girdhar, A. El-Nouby, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Omnimae: Single model masked pretraining on images and videos," *arXiv preprint arXiv:2206.08356*, 2022. 8, 13

[92] C. Feichtenhofer, H. Fan, Y. Li, and K. He, "Masked autoencoders as spatiotemporal learners," *arXiv preprint arXiv:2205.09113*, 2022. 8, 13

[93] A. Gupta, S. Tian, Y. Zhang, J. Wu, R. Martín-Martín, and L. Fei-Fei, "Maskvit: Masked visual pre-training for video prediction," *arXiv preprint arXiv:2206.11894*, 2022. 8

[94] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021. 8

[95] N. Mu, A. Kirillov, D. Wagner, and S. Xie, "Slip: Self-supervision meets language-image pre-training," *arXiv preprint arXiv:2112.12750*, 2021. 8

[96] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *ICML*, 2021. 8

[97] H. Bao, W. Wang, L. Dong, Q. Liu, O. K. Mohammed, K. Aggarwal, S. Som, and F. Wei, "Vlmo: Unified vision-language pre-training with mixture-of-modality-experts," *arXiv preprint arXiv:2111.02358*, 2022. 8

[98] X. Geng, H. Liu, L. Lee, D. Schuurams, S. Levine, and P. Abbeel, "Multimodal masked autoencoders learn transferable representations," *arXiv preprint arXiv:2205.14204*, 2022. 8

[99] Z.-Y. Dou, Y. Xu, Z. Gan, J. Wang, S. Wang, L. Wang, C. Zhu, P. Zhang, L. Yuan, N. Peng *et al.*, "An empirical study of training end-to-end vision-and-language transformers," in *CVPR*, 2022. 8

[100] Y. Bitton, G. Stanovsky, M. Elhadad, and R. Schwartz, "Data efficient masked language modeling for vision and language," *arXiv preprint arXiv:2109.02040*, 2021. 8

[101] J. Lu, C. Clark, R. Zellers, R. Mottaghi, and A. Kembhavi, "Unified-io: A unified model for vision, language, and multi-modal tasks," *arXiv preprint arXiv:2206.08916*, 2022. 8

[102] H. Bao, W. Wang, L. Dong, and F. Wei, "Vl-beit: Generative vision-language pretraining," *arXiv preprint arXiv:2206.01127*, 2022. 8

[103] Y. Pang, W. Wang, F. E. Tay, W. Liu, Y. Tian, and L. Yuan, "Masked autoencoders for point cloud self-supervised learning," *arXiv preprint arXiv:2203.06604*, 2022. 9, 13

[104] H. Liu, M. Cai, and Y. J. Lee, "Masked discrimination for self-supervised learning on point clouds," *arXiv preprint arXiv:2203.11183*, 2022. 9, 13

[105] R. Zhang, Z. Guo, P. Gao, R. Fang, B. Zhao, D. Wang, Y. Qiao, and H. Li, "Point-m2ae: Multi-scale masked autoencoders for hierarchical point cloud pre-training," *arXiv preprint arXiv:2205.14401*, 2022. 9, 13

[106] C. Min, D. Zhao, L. Xiao, Y. Nie, and B. Dai, "Voxel-mae: Masked autoencoders for pre-training large-scale point clouds," *arXiv e-prints*, pp. arXiv–2206, 2022. 9

[107] Q. Tan, N. Liu, X. Huang, R. Chen, S.-H. Choi, and X. Hu, "Mgae: Masked autoencoders for self-supervised learning on graphs," *arXiv preprint arXiv:2201.02534*, 2022. 9, 13

[108] H. Chen, S. Zhang, and G. Xu, "Graph masked autoencoder," *arXiv preprint arXiv:2202.08391*, 2022. 9, 13

[109] Z. Hou, X. Liu, Y. Dong, C. Wang, J. Tang *et al.*, "Graphmae: Self-supervised masked graph autoencoders," *arXiv preprint arXiv:2205.10803*, 2022. 9, 13

[110] J. Li, R. Wu, W. Sun, L. Chen, S. Tian, L. Zhu, C. Meng, Z. Zheng, and W. Wang, "Maskgae: Masked graph modeling meets graph autoencoders," *arXiv preprint arXiv:2205.10053*, 2022. 9, 13

[111] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel *et al.*, "Soft actor-critic algorithms and applications," *arXiv preprint arXiv:1812.05905*, 2018. 9

[112] T. Tao, D. Reda, and M. van de Panne, "Evaluating vision transformer methods for deep reinforcement learning from pixels," *arXiv preprint arXiv:2204.04905*, 2022. 9

[113] M. Laskin, K. Lee, A. Stooke, L. Pinto, P. Abbeel, and A. Srinivas, "Reinforcement learning with augmented data," *Advances in neural information processing systems*, vol. 33, pp. 19 884–19 895, 2020. 9

[114] T. Xiao, I. Radosavovic, T. Darrell, and J. Malik, "Masked visual pre-training for motor control," *arXiv preprint arXiv:2203.06173*, 2022. 9

[115] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017. 9

[116] Y. Seo, D. Hafner, H. Liu, F. Liu, S. James, K. Lee, and P. Abbeel, "Masked world models for visual control," *arXiv preprint arXiv:2206.14244*, 2022. 9

[117] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Masked spectrogram modeling using masked autoencoders for learning general-purpose audio representation," *arXiv preprint arXiv:2204.12260*, 2022. 9

[118] A. Baade, P. Peng, and D. Harwath, "Mae-ast: Masked autoencoding audio spectrogram transformer," *arXiv preprint arXiv:2203.16691*, 2022. 9

[119] M. Zha, "Time series generation with masked autoencoder," *arXiv preprint arXiv:2201.07006*, 2022. 9

[120] K. Majmundar, S. Goyal, P. Netrapalli, and P. Jain, "Met: Masked encoding for tabular data," *arXiv preprint arXiv:2206.08564*, 2022. 9

[121] H. Xu, S. Ding, X. Zhang, H. Xiong, and Q. Tian, "Masked autoencoders are robust data augmentors," *arXiv preprint arXiv:2206.04846*, 2022. 9, 13

[122] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *CVPR*, 2017. 13

[123] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *NeurIPS*, 2017. 13

[124] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *ICLR*, 2017. 13

[125] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *ICLR*, 2018. 13

[126] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" *ICLR*, 2019. 13

7 APPENDIX

TABLE 2
Summary of works with maksed autoencoder in vision.

model	masking strategy	input type	prediction target	pretrain dataset(resolution)	loss	encoder	decoder	Finetune Accuracy	arxiv time	publish status
BEiT [10]	block-wise	all patches	tokens	ImageNet-1K(224)	dVAE	ViT-B ViT-L	dVAE	86.2% 85.2%	2021.06.15	ICLR 2022
MAE [1]	random	visible	pixel	ImageNet-1K(224)	MSE	ViT-B ViT-L ViT-H	8 block,width 512	83.6% 85.9% 86.9%	2021.11.11	CVPR 2022
CAE [46]	block-wise	visible	tokens	ImageNet-1K(224)	CE(MIM), MSE(align)	ViT-S ViT-B ViT-L	4 blocks 4 blocks 4 blocks	82.0% 83.9% 86.3%	2022.02.07	arxiv
SimMIM [25]	random	all patches	pixel	ImageNet-1K(224)		ViT-B	linear layer	83.8%	2021.11.18	CVPR 2022
Peco [44]	random	all	token from proposed tokenizer	ImageNet-1K (800epoch)	same as BEiT	ViT-B	decoder	84.5%	21.11.24	
With joint embedding					**loss**					
iBOT [29]	block-wise	visible	tokens	ImageNet-1K(224)		ViT-B	None	87.8%	2021.11.15	ICLR2022
SplitMask [63]	block-wise	all patches	tokens	ImageNet-1K(224) ImageNet-1K(224)	CE(MIM),InfoNCE	ViT-S ViT-B	2 blocks 2 blocks	81.5% 83.6%	2021.12.20	arxiv
AttMask [80]	attention-guided mask =object-bias mask	masked(student)&unmasked(teacher)	teacher's feature	ImageNet-1K	distillation(KLdiv)	iBOT, ViT-S for attention	None	Incomparable	22.05.23	arXiv
ConMIM [22]	random	masked(student)&unmasked(teacher)	teacher's feature	ImageNet-1K(300)	InfoNCE	ViT-S ViT-B	decoder	82.0% 83.7%	22.05.19	arXiv
SIM [74]	random	masked(student)&unmasked(teacher)	teacher's feature	dataset	cosine similarity	ViT-B	decoder	Incomparable	22.06.02	arXiv
Contrastive leaning	**with recon loss**				**loss**					
MST [29]	no masking	all	pixel	ImageNet-1K (300)	DINO+l1 recon	DeiT-S	CNN	76.9%	21.01.10	NeurIPS 2021
Repre [76]	random	all	pixel	ImageNet-1K (800) ImageNet-1K (800)	DINO+l1 recon	ViT-S ViT-B	CNN	77.9% 79.2%	22.01.18	arXiv
Efficient and Effective	**=fast pretraining**			(epoch)						
LoMaR [54]	random window&patchs	all in windows	pixel	ImageNet-1K (300) ImageNet-1K (1600)	MSE	ViT-B	MLP (ReconHead)	83.3% 84.1%	22.06.01	arxiv
ObjMAE [55]	random,object-aware	visible in object	pixel in object	ImageNet100	MSE	sameMAE,ViT-B	sameMAE,ViT-B	Incomparable	22.05.28	arxiv
Beyond Masking [67]	random mask+ro(ex,zoom-in..)	visible	pixel	ImageNet-1K (300)	MSE	sameMAE,ViT-B	sameMAE,ViT-B	83.2%(mask+zoomin)	22.05.27	arxiv
Lightweight architecture										
GreenMaeTrans [51]	group window attention	visible	Pixel	ImageNet-1K (800)	MSE	Swin-B Swin-L	ViT	83.7% 85.1%	22.05.26	arxiv
Uniform Masking [52]	proposed uniform masking	all, 25%mask	pixel	ImageNet-1K (200)	MSE	PVT-S Swin-T	ViT	82.04% 82.0%	22.05.20	arxiv
MixMIM [57]	No mask	A mixture of two images	Two image pixel	ImageNet-1K (300) ImageNet-1K (300) ImageNet-1K (600) ImageNet-1K (600) ImageNet-1K (300)	MSE	ViT-B ViT-L Swin-B Swin-L PVT-L	8 blocks	83.2% 85.0% 84.4% 85.7% 83.2%	22.05.26	arxiv
HiViT [53]	random	visible	pixel	ImageNet-1K (300) ImageNet-1K (800)	MSE	(proposed) HiViT-B	6 blocks	83.8% 84.2%	22.05.30	arxiv
MAE augmentor/Others										
CIM [67]	reconed=corrupted	all corrupted image	all pixel	ImageNet-1K	L1+L2+discriminative Generator: small BEiT	ViT-S ViT-B ResNet-50	none	81.6% 83.3% 80.4%	22.02.07	arxiv
MRA [121]	n	n	n	n	n	MAE	MAE	Incomparable	22.06.10	arxiv
Medical images										
Self-MAE [82]	random	visible	pixel	ChestX-ray14 (224)	MAE	ViT-B	ViT	lung disease classification 81.5%	2022.3.10	arxiv
MIM-Medical [83]	random	all	raw voxel	BTCV, TCIA-COVID19 (9)	l1,l2	ViT-B/L	linear layer	multi-organ segmentation 76.03%	2022.04.25	arxiv
DAMA [84]	adaptive	visible	pixel,feature	ImageNet-1K(224)	MSE?	ViT-B	?	500 epoch, 83.17	2022.5.10	arxiv
GCMAE [85]	random	visible	pixel	Camelyon16(224)	MSE	ViT	8 blocks transformer	83.29%	2022.5.18	arxiv
SD-MAE [87]	random	visible	pixel	PatchCamelyon,NCT-CRC-HE,ImageNet-100 (224)	MSE	ViT-S	4 blocks 192d, transformer	ImageNet-100,84.63%	2022.03.21	arxiv
MAE-MIL [86]	random	visible	pixel	Camelyon16(1024)	MSE	?	?	61%	/	MIDL 2022
Multi-modal										
MultiMAE [68]	random	visible	pixel	ImageNet-1K(224)	loss	ViT-B		83.3%	2022.04.04	arxiv

TABLE 3
Summary of works with masked autoencoder on videos.

model	masking strategy	input type	prediction target	pretrain dataset	image size	Test Set	encoder	decoder	Finetune Accuracy	arxiv time	publish status
BEVT [88]	block-wise/tube	all patches	token	ImageNet-1K, HowTo100M	224	SSv2 Diving48 K400	Video Swin-Base	CNN-1, Linear-1	70.6% 86.7% 80.6%	2021.12.02	CVPR 2022
MaskFeat [23]	cube masking	all patches	features (HOG)	K400	224	K400	MViT-S MViT-L	None	82.2% 84.3%	2021.12.16	CVPR 2022
VideoMAE [90]	tube masking	visible	pixel	SSv2 K400	224	SSv2 K400	ViT-B	4 block, 384d	69.3% 79.45%	2022.03.23	arxiv
MAE-video [92]	random	visible	pixel	K400	224	K400 SSv2	ViT-L	4 block, 512d	84.8% 72.1%	2022.05.18	arxiv
OmniMAE [91]	random	all patches	pixel	ImageNet-1K, SSv2	224	ImageNet-1K SSv2 ImageNet-1K SSv2	ViT-B ViT-B ViT-L ViT-L	4 block, 384d	82.8% 69.0% 84.7% 73.4%	2022.06.16	arxiv

TABLE 4
Summary of works with masked autoencoder on point cloud and graph.

model	masking strategy	input type	prediction target	encoder	decoder	pretrain dataset	Finetune Accuracy	arxiv time	publish status
Point Cloud									
Point-BERT [88]	random	all tokens	Tokens extracted from tokenizer (dVAE)	Standard Transformer encoder [28]	PointNet (MLP layers) [122]	ModelNet40 ScanObjectNN ShapeNetPart	93.8% OBJ-BG:87.43, OBJ-ONLY:88.12, PB-T50-RS:83.07 84.11 mIoUc, 85.6 mIoUi	21.11.29	CVPR 2022
Point-MAE [103]	random	visible tokens	Raw point patches	Standard transformer encoder	Standard transformer decoder	ModelNet40 ScanObjectNN ShapeNetPart	93.8% OBJ-BG:90.02, OBJ-ONLY:88.29, PB-T50-RS:85.18 86.1 mIoUi	220313	arXiv
MaskPoint [104]	random	Encoder: visible tokens Decoder: real/fake points	Real or Fake	Standard transformer encoder	Standard transformer decoder	ModelNet40 ScanObjectNN ShapeNetPart	93.8% OBJ-BG:88.1, OBJ-ONLY:89.3, PB-T50-RS:84.3 84.4 mIoUc, 86.0 mIoUi	220321	arXiv
Point-M2AE [105]	multi-scale, random	visible tokens	Raw point patches	Proposed hierarchical transformer	Proposed hierarchical transformer	ModelNet40 ScanObjectNN ShapeNetPart	94.0% OBJ-BG:91.22, OBJ-ONLY:88.81, PB-T50-RS:86.43 84.86 mIoUc, 86.51 mIoUi	220328	arXiv
Graph						**Node classification**			
MGAE [107]	random edge	visible	masked edges	SAGE [123],GCN [124]	dot product, MLP	Cora, CiteSeer, PubMed (classify 70% masked edge)	86.15%, 74.60%, 86.91%	220107	arXiv
GMAE [108]	random node	visible	masked nodes	Graph Transformers	Graph Transformers	Cora, CiteSeer, PubMed	81.14%, 69.25%, 81.40%	220217	arXiv
GraphMAE [109]	random node	all graph	target	GNN(GCN [124], GAT [125], GIN [126])	GNN(GAT, GIN)	Cora, CiteSeer, PubMed	84.2%, 73.4%, 81.1%	220522	KDD'22
MaskGAE [110]	random edge	visible	masked edges, degree regression	GCN	MLP	Cora, CiteSeer, PubMed	84.05%, 73.49%, 83.06%	220520	arXiv