Masked Autoencoders Are Scalable Vision Learners

Kaiming He　Xinlei Chen　Saining Xie　Yanghao Li　Piotr Dollár　Ross Girshick

摘要

掩码自动编码器（MAE）是用于计算机视觉的可扩展自监督学习器。我们屏蔽输入图像的随机块并重建丢失的像素。它基于两个核心设计。首先，我们开发了一个非对称的编码器-解码器架构，其中编码器仅对可见的像素块进行操作，解码器通过潜在表示和掩码标记中重建原始图像。其次，我们发现屏蔽大部分输入图像（例如 75%）会产生重要且有意义的自我监督任务。结合这两种设计使我们能够高效且有效地训练大型模型，我们能够加速训练并提高准确性。我们的模型具有很好的泛化能力，在仅使用 ImageNet-1K 数据集上，实现了 87.8% 的准确度，在下游任务上的性能也优于有监督模型。

关键词：MAE；自监督训练

1　引言

深度学习见证了能力和容量不断增长的架构的爆炸式增长[1]。在硬件快速发展的帮助下，如今的模型可以轻松地过拟合一百万张图像[2]，并开始需要数亿张的标记图像[3]。

这种对数据的需求已经通过自我监督预训练在自然语言处理（NLP）中得到成功解决。基于 GPT[4]中的自回归语言建模和 BERT[5]中的屏蔽自动编码的解决方案，在概念上很简单，它们删除了一部分数据并学习预测删除的内容。这些方法现在可以训练包含超过 1000 亿个参数的可泛化 NLP 模型[6]。掩码自动编码器的想法是一种更通用的去噪自动编码器[7]。它同样也适用于计算机视觉，甚至在视觉上的研究[8]先于 BERT，然而，尽管随着 BERT 的成功，人们对这个想法产生了浓厚的兴趣，但在视觉方向上，自动编码方法的进展却落后于 NLP。掩码自动编码器在视觉和语言之间的不同，我们试图从一下几个角度来解释。

第一，在视觉方面，卷积网络在过去十年中占据主导地位[9]。卷积通常在规则网格上运行，将掩码标记或位置嵌入等指标集成到卷积网络中并不简单。然而，这一架构差距已通过引入 Vision Transformers（ViT）[10]得到解决，不应再成为障碍。第二，语言和视觉之间的信息密度不同。语言是人类生成的信号，具有高度语义和信息密集性。当训练一个模型来预测每个句子中只有几个缺失的单词时，这个任务似乎会引发复杂的语言理解。相反，图像是具有大量空间冗余的自然信号。为了克服这种差异并鼓励学习有用的功能，我们证明了一种简单的策略可以在计算机视觉中效果很好：屏蔽很大一部分随机像素块。这种策略在很大程度上减少了冗余并创建了一项具有挑战性的自我监督任务，需要超越低级图像统计的整体理解。第三，自动编码器的解码器将潜在表示映射回输入，在重建文本和图像之间起着不同的作用。在视觉中，解码器重建像素，因此其输出的语义级别低于常见的识别任务。而在语言中，解码器预测包含丰富语义信息的缺失词。虽然在 BERT 中，解码器可能很简单[5]，但我们发现对于图像，解码器设计在确定学习的潜在表示的语义级别方面起着关键作用。

在这些分析下，我们提出了一种简单有效且可扩展的用于视觉表示学习的掩蔽自动编码器 (MAE) 形式。我们的 MAE 屏蔽了输入图像中的随机补丁，并在像素空间中重建缺失的补丁。它具有非对称

编码器解码器设计，我们的编码器仅作用于可见的像素块，而我们的解码器是轻量级的，并从潜在表示和掩码标记中重建输入。将掩码标记转移到我们的非对称编码器-解码器中的小型解码器会大大减少计算量。在这种设计下，非常高的屏蔽率（例如 75%）可以实现双赢：它优化了准确性，同时允许编码器仅处理一小部分（例如 25%）的补丁。这可以将整体预训练时间减少 3 倍或更多，同样减少内存消耗，使我们能够轻松地将 MAE 扩展到大型模型。

MAE 具有很好的泛化性，通过 MAE 预训练，我们可以在 ImageNet-1K 上训练 ViT-Large/-Huge 等数据密集型模型，提高泛化性能。使用 ViT-Huge[10]模型，我们在 ImageNet-1K 上微调时达到 87.8% 的准确率。我们还评估了对象检测、实例分割和语义分割的迁移学习。在这些任务中，我们的预训练取得了比有监督的预训练更好的结果，更重要的是，我们通过扩大模型观察到显着的收益。这些观察结果与 NLP[5]中自我监督预训练中观察到的观察结果一致，我们希望它们将使我们的领域能够探索类似的轨迹。

2 相关工作

2.1 掩码语言模型

掩蔽语言建模及其自回归算法，例如 BERT 和 GPT，是 NLP 预训练中非常成功的方法。这些方法保留输入序列的一部分并训练模型来预测缺失的内容。这些方法已被证明可以很好地扩展[6]，并且大量证据表明这些预训练的表征可以很好地泛化到各种下游任务。

2.2 自动编码

自动编码是学习表示的经典方法。它有一个将输入映射到潜在表示的编码器和一个重建输入的解码器。例如，PCA 和 k-means 是自动编码器[11]。去噪自动编码器（DAE）[7]是一类自动编码器，它会破坏输入信号并学习重建原始的、未损坏的信号。这一系列方法可以被认为是不同情况下的广义 DAE，例如，对像素做掩码[12]或删除颜色通道[13]。我们的 MAE 是一种去噪自动编码形式，但在许多方面都不同于经典的 DAE。

2.3 掩码图像编码

掩码图像编码方法从被掩码的图像中学习表征。[14] 的开创性工作将掩蔽作为 DAE 中的一种噪声类型，上下文编码器使用卷积网络修复大量缺失区域。受 NLP 成功的推动，相关的最新方法基于 Transformers[1]。iGPT[12]对像素序列进行操作并预测未知像素。ViT[3]论文中研究了自监督学习的掩蔽补丁预测。最近，BEiT[15]提出预测离散标记[16]。

2.4 自监督学习

自监督学习方法在计算机视觉中应用广泛，通常侧重于不同的预训练代理任务[17]。最近，对比学习[18]很流行，例如[19]，它对两个或多个视图之间的图像相似性和不相似性（或仅相似性[20]）进行建模。对比学习和相关方法强烈依赖于数据增强[21]。自动编码追求概念上不同的方向，它表现出我们将要呈现的不同行为。

3 本文方法

3.1 本文方法概述

MAE 是一种简单的自动编码方法，可在给定部分观察的情况下重建原始信号。与所有自动编码器一样，我们的方法有一个将观察到的信号映射到潜在表示的编码器，以及一个从潜在表示重建原始信号的解码器。与经典自动编码器不同，我们采用非对称设计，允许编码器仅对部分观察到的信号（没有掩码标记）进行操作，并采用轻量级解码器从潜在表示和掩码标记中重建完整信号，图 1展示了这个想法。

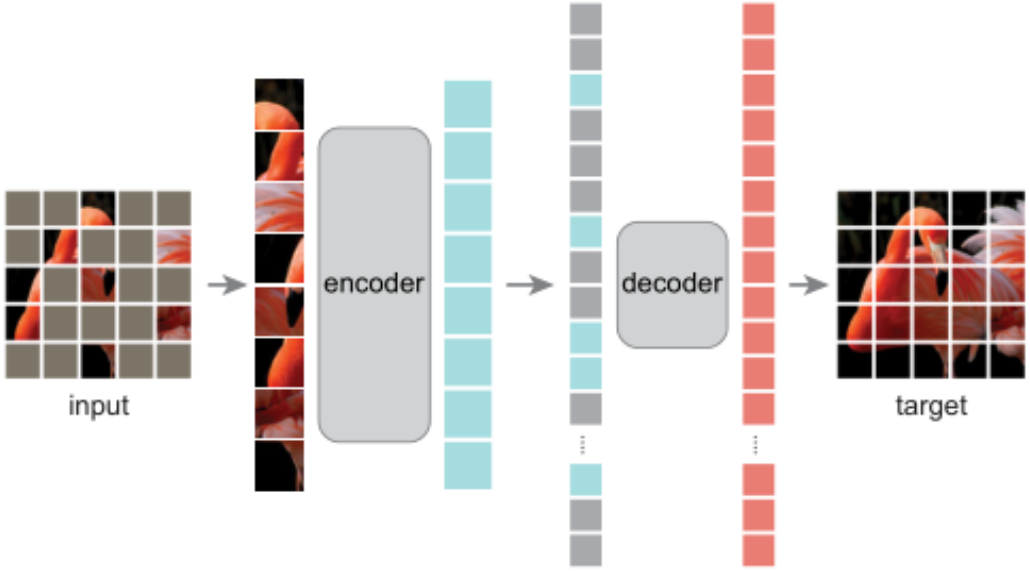

图 1: MAE 模型架构

3.2 掩码操作

跟 ViT[10]一样，我们将图像划分为规则的非重叠像素块。然后我们对像素块的一个子集进行采样并对其余的做掩码操作，即删除。我们的抽样策略很简单，在不放回的情况下按照均匀分布对随机像素块进行抽样。具有高掩蔽率（即移除像素块的比率）的随机采样在很大程度上消除了冗余，从而创建了一个无法通过从可见的相邻像素块外推轻松解决的任务（见图 2 - 4）。均匀分布可防止潜在的中心偏差（即，图像中心附近有更多掩码块）。最后，高度稀疏的输入为设计高效编码器创造了机会。

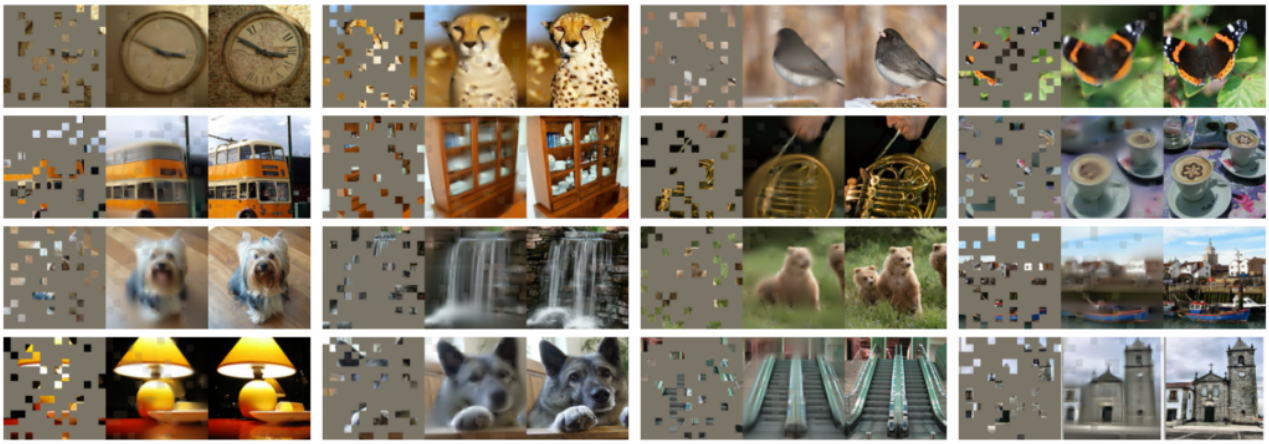

图 2: ImageNet 验证图像的示例结果。对于每个三元组，我们展示了掩码图像（左）、我们的 MAE 重建（中）和真实图像（右）

图 3: COCO 验证图像的示例结果，使用在 ImageNet 上训练的 MAE

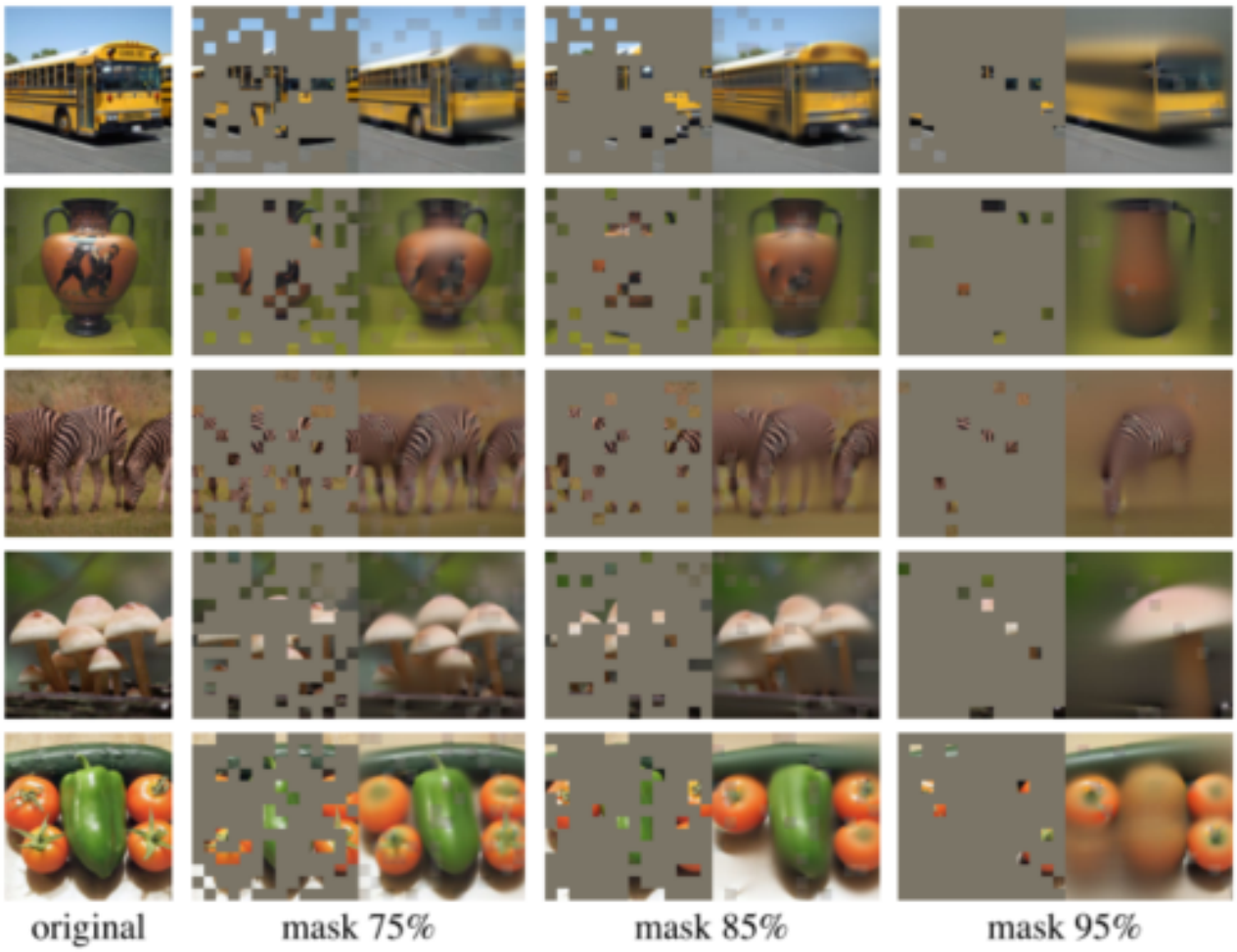

图 4: 使用经过 75% 掩蔽率预训练但应用于具有更高掩蔽率的输入的 MAE 重建 ImageNet 验证图像。

3.3 MAE 编码器

MAE 编码器来源于 ViT，仅作用于可见的、未掩码的像素块。就像在标准 ViT 中一样，我们的编码器通过添加位置嵌入的线性投影嵌入像素块，然后通过一系列 Transformer 块处理结果集。然而，我们的编码器只对整个集合的一小部分（例如 25%）进行操作。掩码补丁被移除，不使用掩码标记。这使我们能够仅使用一小部分计算和内存来训练非常大的编码器。

3.4 MAE 解码器

MAE 解码器的输入是完整的像素块，包括通过编码器的可见像素块和掩码处理的像素块。每个掩码标记[5]都是一个共享的学习向量，指示存在要预测的缺失补丁。我们将位置嵌入添加到这个完整集合中的所有标记；如果没有这个，掩码标记将没有关于它们在图像中的位置的信息。解码器有另一系列的 Transformer 块。MAE 解码器仅在预训练期间用于执行图像重建任务，仅编码器用于生成用于识别的图像表示。因此，可以以独立于编码器设计的方式灵活地设计解码器架构。我们用非常小的解码器进行实验，比编码器更窄更浅。例如，与编码器相比，解码器每个标记的计算量小于 10%。通过

这种不对称设计，所有像素块仅由轻量级解码器处理，从而显着减少了预训练时间。

3.5 重建目标

MAE 通过预测每个掩码块的像素值来重建输入。解码器输出中的每个元素都是代表一个块的像素值向量。解码器的最后一层是线性投影，其输出通道数等于块中的像素值数。解码器的输出被重塑以形成重建图像。我们的损失函数计算像素空间中重建图像和原始图像之间的均方误差（MSE）。我们仅计算掩蔽块的损失，类似于 BERT。我们还研究了一种变体，其重建目标是每个掩蔽块的归一化像素值。具体来说，我们计算一个块中所有像素的均值和标准差，并使用它们对这个块进行归一化。在实验中使用归一化像素作为重建目标提高了表示质量。

4 复现细节

4.1 与已有开源代码对比

与已有开源代码相比，实现基于 MAE 模型使用 UPerNet 的方法对数据集 ADE20K 实现语义分割的任务。结合 UPerNet，把训练完成的 MAE 编码器保留下来，舍弃 MAE 的解码器，仅使用编码器来生成用于识别的图像表示，从数据集中采样样本，提取到样本的潜在特征作为输入，输入到 UPerNet 网络模型中经行语义信息划分和输出。

4.2 实验环境搭建

该实验所使用的环境如下：

系统版本：Linux version 5.15.0-43-generic

GPU 版本：2 × NVIDIA GeForce RTX3090 24G

torch 版本：1.12.1

tensorboard：2.10.1

matplotlib：3.5.3

4.3 实验设计细节

实验首先在 cifar100 的数据集上进行了自监督训练，对 cifar100 上的数据实现重构，使用 pytorch 实现的 MAE-vit 模型，由于时间和机器的限制，训练的是 ViT 的 tiny 版，在输入图像的尺寸上采取 224*224 大小，patches 数量为 16。编码器的设计，维度设为 192，网络深度设为 12，transformer 中的多头注意力机制数量设为 3，mask 比例为 75%。解码器的设计，维度为 512，网络深度为 8，transformer 中的多头注意力机制数量设为 16。足够深的解码器对于线性探测很重要。这可以用像素重建任务和识别任务之间的差距来解释：自动编码器中的最后几层更专门用于重建，但与识别的相关性较低。一个相当深的解码器可以解释重建专业化，将潜在表示留在更抽象的层次上。此设计可使线性探测性能提高高达 8%。但是，如果使用微调，则可以调整编码器的最后一层以适应识别任务, 解码器深度对改进微调的影响较小。对于超参数的设置，学习率设为 1.5e-4，batch_size 设为 256。在 1500 个迭代次数之后，训练基本达到收敛。

blocks	ft	lin
1	84.8	65.5
2	**84.9**	70.0
4	**84.9**	71.9
8	**84.9**	**73.5**
12	84.4	73.3

dim	ft	lin
128	**84.9**	69.1
256	84.8	71.3
512	**84.9**	**73.5**
768	84.4	73.1
1024	84.3	73.1

图 5: 解码器深度及宽度

之后的实验是基于官方开源的 MAE 预训练模型，结合 UPerNet 的方法在 ADE20K 数据集上完成语义分割的迁移任务。在输入图像的尺寸上采取 512*512 大小，patches 数量为 16。在模型架构上，舍弃了 MAE 中的解码器部分，只采用编码器来提取潜在特征，维度设为 768，网络深度设为 12，transformer 中的多头注意力机制数量设为 12，对于超参数的设置，学习率设为 1e-4。之后再通过 UPerNet 对得到的潜在特征处理，完成语义分割的任务，得到效果输出。

5 实验结果分析

图 6: cifar100 验证图像的示例结果

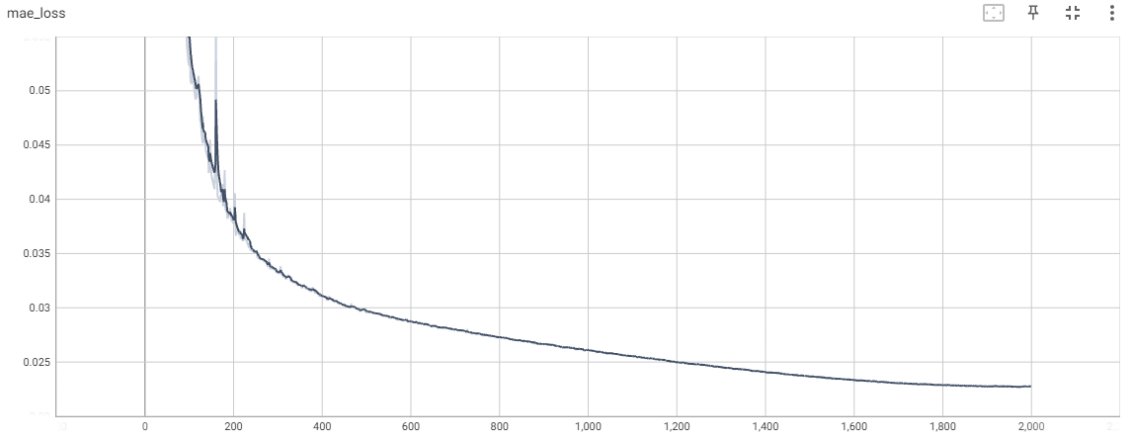

图 7: MAE 训练 loss 曲线图

在 cifar10 数据集上完成 MAE 自监督训练，图 6 展示的是在 2000 个迭代次数之后，MAE 重构原始图像的效果图，可以看出在大体的轮廓和色彩上，重构图像都比较接近于真实图像，表现出来的效果还是很好的。图 7 表示的是在训练过程中 loss 的曲线图，可以看到在 1600 个迭代次数之后，模型

就基本收敛，但随着迭代次数的增加，loss 还有继续下降的趋势。由此证明，在图像上的自监督学习也是可以表现的很好的，能够达到有监督训练模型的效果。

图 8: ADE20K 验证图像语义分割的示例结果

基于 MAE 预训练模型，结合 UPerNet 的方法在 ADE20K 数据集上完成语义分割的迁移任务。图 8 是本实验的语义分割效果图，可以看出能达到比较出色的分割效果。对于主体的内容，都能很好的捕获，并于背景部分区分开来。下表展示了 MAE 模型在 ADE20K 上的实验结果，平均准确率能达到 59.36%，mIoU 指标可以达到 48.13%，证明了 MAE 在迁移学习上同样能表现的很出色，具有很强的泛化能力，不逊色于在特定数据集上训练的有监督模型。

aAcc	mIoU	mAcc
82.99	48.13	59.36

6 总结与展望

可扩展的算法是深度学习的核心。在 NLP 中，简单的自监督学习方法可以从指数缩放模型中获益。在计算机视觉中，尽管自我监督学习取得了进展，但实用的预训练范式主要受到监督。在这项研究中，我们在 ImageNet 数据集和迁移学习中观察到自动编码器，一种类似于 NLP 技术的简单自监督方法，便可以提供了可扩展的优势。视觉中的自我监督学习现在可能正在走上与 NLP 中类似的轨迹。

另一方面，我们注意到图像和语言是不同性质的信号，必须谨慎处理这种差异。图像只是记录下来的光，没有语义分解成文字的视觉模拟。我们没有尝试删除对象，而是删除了很可能不形成语义段的随机补丁。同样，MAE 重建像素，这些像素不是语义实体。尽管如此，我们观察到 MAE 推断出复杂的整体重建，表明它已经学习了许多视觉概念，即语义。我们假设这种行为是通过 MAE 内部丰富的隐藏表示形式发生的。这也是未来工作可以研究探讨的方向。

参考文献

[1] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[2] DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. 2009: 248-255.

[3] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[4] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[J]., 2018.

[5] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[6] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.

[7] VINCENT P, LAROCHELLE H, BENGIO Y, et al. Extracting and composing robust features with denoising autoencoders[C]//Proceedings of the 25th international conference on Machine learning. 2008: 1096-1103.

[8] PATHAK D, KRAHENBUHL P, DONAHUE J, et al. Context encoders: Feature learning by inpainting [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2536-2544.

[9] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[10] BOGDANOV S, LÜLLMANN C, MARTIN P, et al. Honey quality and international regulatory standards: review by the International Honey Commission[J]. Bee world, 1999, 80(2): 61-69.

[11] HINTON G E, ZEMEL R. Autoencoders, minimum description length and Helmholtz free energy[J]. Advances in neural information processing systems, 1993, 6.

[12] CHEN M, RADFORD A, CHILD R, et al. Generative pretraining from pixels[C]//International conference on machine learning. 2020: 1691-1703.

[13] ZHANG R, ISOLA P, EFROS A A. Colorful image colorization[C]//European conference on computer vision. 2016: 649-666.

[14] VINCENT P, LAROCHELLE H, LAJOIE I, et al. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.[J]. Journal of machine learning research, 2010, 11(12).

[15] BAO H, DONG L, WEI F. Beit: Bert pre-training of image transformers[J]. arXiv preprint arXiv:2106.08254, 2021.

[16] VAN DEN OORD A, VINYALS O, et al. Neural discrete representation learning[J]. Advances in neural information processing systems, 2017, 30.

[17] DOERSCH C, GUPTA A, EFROS A A. Unsupervised visual representation learning by context prediction[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1422-1430.

[18] BECKER S, HINTON G E. Self-organizing neural network that discovers surfaces in random-dot stereograms[J]. Nature, 1992, 355(6356): 161-163.

[19] WU Z, XIONG Y, YU S X, et al. Unsupervised feature learning via non-parametric instance discrimination[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 3733-3742.

[20] GRILL J B, STRUB F, ALTCHÉ F, et al. Bootstrap your own latent-a new approach to self-supervised learning[J]. Advances in neural information processing systems, 2020, 33: 21271-21284.

[21] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations[C]//International conference on machine learning. 2020: 1597-1607.