Masked Autoencoders Are Scalable Vision Learners

掩码自动编码器是可扩展的视觉学习者

1. Abstract

This paper shows that masked autoencoders (MAE) are scalable self-supervised learners for computer vision.

本文表明，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。

Our MAE approach is simple: we mask random patches of the input image and reconstruct the missing pixels.

我们的 MAE 方法很简单：我们随机屏蔽输入图像某些像素块并重建丢失的像素 。

It is based on two core designs. First, we develop an asymmetric encoder-decoder architecture, with an encoder that operates only on the visible subset of patches (without mask to-kens), along with a lightweight decoder that reconstructs the original image from the latent representation and mask tokens. Second, we find that masking a high proportion of the input image, e.g., 75%, yields a nontrivial and meaningful self-supervisory task.

它基于 关于两个核心设计。 首先，我们开发了一个不对称的 编码器-解码器架构，编码器仅在可见的**补丁子集**上运行（没有掩码标记），以及轻量级解码器从**隐层表示（中间层特征）和掩码标记**重建原始图像。 其次，我们发现掩蔽的比例很高的输入图像，例如 75%，产生一个非平凡的和有意义的自我监督任务。

Coupling these two designs enables us to train large models efficiently and effectively: we accelerate training (by 3× or more) and improve accuracy. Our scalable approach allows for learning，high-capacity models that generalize well: e.g., a vanilla ViT-Huge model achieves the best accuracy (87.8%) amongmethods that use only ImageNet-1K data. Transfer per-formance in downstream tasks outperforms supervised pre-training and shows promising scaling behavior.

结合这两种设计使我们能够有效地训练大型模型：我们加速训练（3 倍或更多）并提高准确性。 我们的可扩展方法**允许学习 泛化能力强的大容量模型**：例如，vanilla ViT-Huge 模型的准确率最高（87.8%）仅使用 ImageNet-1K 数据的方法。 下游任务中的迁移性能优于有监督的预训练，并显示出有希望的扩展行为。

2. Introduction

*Deep learning has witnessed an explosion of architectures of continuously growing capability and capacity**[33, 25, 57]. Aided by the rapid gains in hardware, models today can easily overfit one million images [13] and** begin to demand hundreds of millions of—often publicly inaccessible—labeled images [16]*

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 [33、25、57]。 借助硬件的快速增长，今天的模型可以轻松地过拟合一百万张图像 [13] 和 开始要求数亿——通常是公开的 无法访问 - 标记的图像

This appetite for data has been successfully addressed in natural language processing (NLP) by self-supervised pre-training. The solutions, based on autoregressive language modeling in GPT [47, 48, 4] and masked autoencoding in BERT [14], are conceptually simple: they remove a portion of the data and learn to predict the removed content. These methods now enable training of generalizable NLP models containing over one hundred billion parameters [4]

这种对数据的需求已成功解决 通过自我监督的预训练进行自然语言处理（NLP）。 基于自回归语言的解决方案 GPT [47, 48, 4] 中的建模和掩码自动编码 BERT [14] 在概念上很简单：它们删除了一部分 并学习预测删除的内容。 这些 方法现在可以训练可泛化的 NLP 模型 包含超过一千亿个参数 [4]

the idea of masked autoencoders, a form of more general denoising autoencoders [58], is natural and applicable in computer vision as well. Indeed, closely related research in vision [59, 46] preceded BERT. However, despite significant interest in this idea following the success of BERT, progress of autoencoding methods in vision lags behind NLP. We ask: what makes masked autoencoding different between vision and language? We attempt to answer this question from the following perspectives:

掩蔽自动编码器的想法，一种更通用的去噪自动编码器 [58] 的形式，在计算机视觉中也很自然且适用。 事实上，与视觉密切相关的研究 [59, 46] 早于 BERT。 然而，尽管随着 BERT 的成功对这一想法产生了浓厚的兴趣，但视觉自动编码方法的进展仍落后于 NLP。 我们问：**是什么让掩码自动编码在视觉和语言之间与众不同？** 我们试图从以下几个方面来回答这个问题：

(i) Until recently, architectures were different. In vision,convolutional networks [34] were dominant over the last decade [33]. Convolutions typically operate on regular grids and it is not straightforward to integrate ‘indicators’ such as mask tokens [14] or positional embeddings [57] into convolutional networks. This architectural gap, however, has been addressed with the introduction of Vision Transformers (ViT) [16] and should no longer present an obstacle

(i) 直到最近，架构还是不同的。 在视觉上， 卷积网络 [34] 在过去占主导地位 十年[33]。 **卷积通常在规则网格上运行，并且 整合“指标”并不简单，例如 将标记 [14] 或位置嵌入 [57] 掩码到卷积网络中。** 然而，这种架构差距已经 已通过引入 **Vision Transformers (ViT)** [16] 得到解决，不应再成为障碍

(ii) Information density is different between language and vision. Languages are human-generated signals that are highly semantic and information-dense. When training a model to predict only a few missing words per sentence, this task appears to induce sophisticated language understanding. Images, on the contrary, are natural signals with heavy spatial redundancy—e.g., a missing patch can be recovered from neighboring patches with little high-level understanding of parts, objects, and scenes. To overcome this difference and encourage learning useful features, we show that a simple strategy works well in computer vision: masking a very high portion of random patches. This strategy largely reduces redundancy and creates a challenging self-supervisory task that requires holistic understanding beyond low-level image statistics. To get a qualitative sense of our reconstruction task, see Figures 2 – 4.

(ii) 语言和视觉的**信息密度**不同。 语言是人类产生的信号是**高度语义和信息密集的。** 当训练一个模型只预测每个句子中的几个缺失词时， 这项任务似乎可以**诱导复杂的语言理解。** 相反，图像是自然信号**严重的空间冗余**——例如，可以从相邻的块中恢复缺失的块，而对部分、对象和场景的理解很少。 为了克服这种差异并鼓励学习有用的特征，我们证明了一种简单的策略在计算机视觉中效果很好：**掩盖很大一部分随机补丁。** 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。 要对我们的重建任务有一个定性的认识，请参见图 2-4。

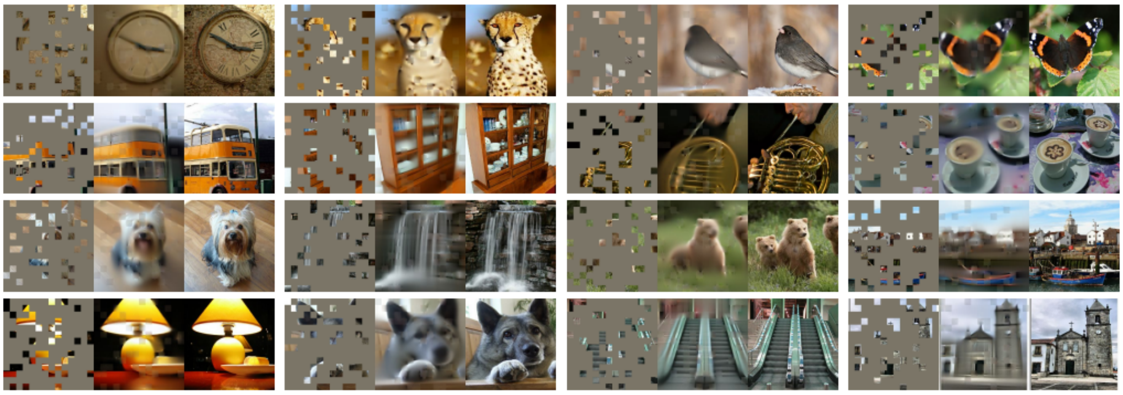


Figure 2. Example results on ImageNet *validation* images. For each triplet, we show the masked image (left), our MAE reconstruction[†] (middle), and the ground-truth (right). The masking ratio is 80%, leaving only 39 out of 196 patches. More examples are in the appendix.

[†] *As no loss is computed on visible patches, the model output on visible patches is qualitatively worse. One can simply overlay the output with the visible patches to improve visual quality. We intentionally opt not to do this, so we can more comprehensively demonstrate the method's behavior.*

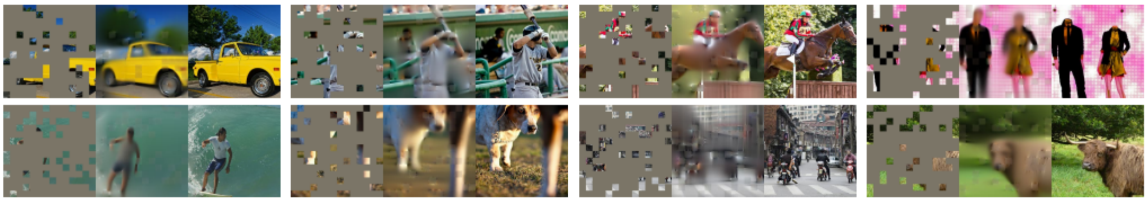


Figure 3. Example results on COCO validation images, using an MAE trained on ImageNet (the same model weights as in Figure 2). Observe the reconstructions on the two right-most examples, which, although different from the ground truth, are semantically plausible.

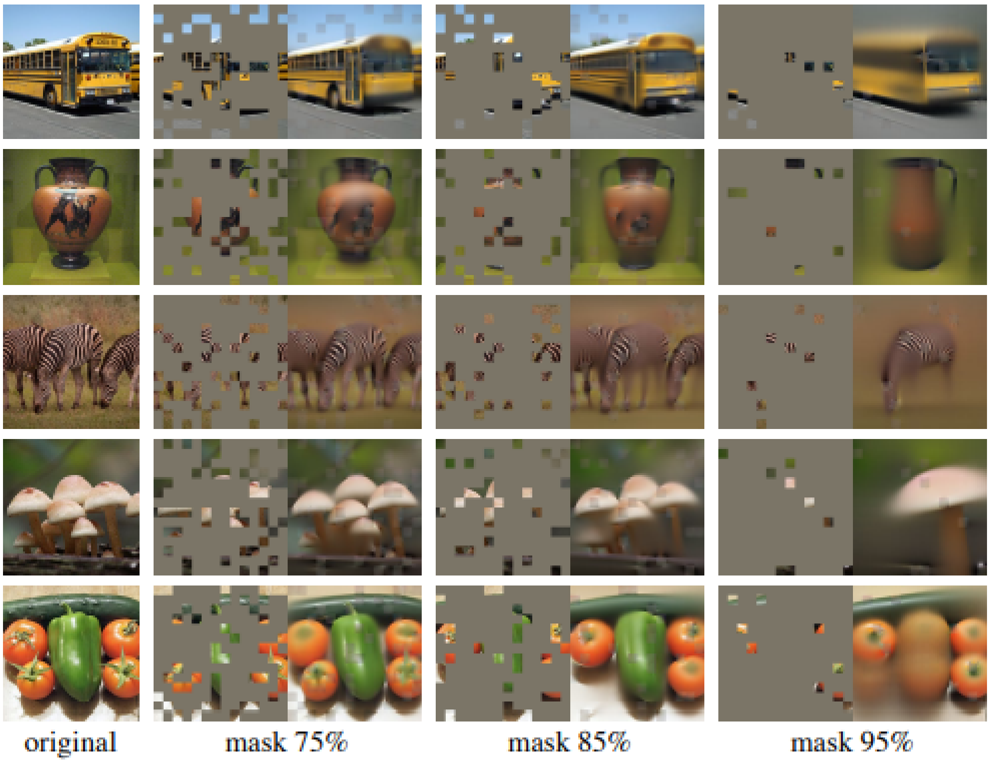


Figure 4. Reconstructions of ImageNet *validation* images using an MAE pre-trained with a masking ratio of 75% but applied on inputs with higher masking ratios. The predictions differ plausibly from the original images, showing that the method can generalize.

(iii) The autoencoder's decoder, which maps the latent representation back to the input, plays a different role between reconstructing text and images. In vision, the decoder reconstructs pixels, hence its output is of a lower semantic level than common recognition tasks. This is in contrast to language, where the decoder predicts missing words that contain rich semantic information. While in BERT the

decoder can be trivial (an MLP) [14], we found that for images, the decoder design plays a key role in determining the semantic level of the learned latent representations. Driven by this analysis, we present a simple, effective, and scalable form of a masked autoencoder (MAE) for visual representation learning. Our MAE masks random patches from the input image and reconstructs the missing patches in the pixel space. It has an asymmetric encoder-decoder design. Our encoder operates only on the visible subset of patches (without mask tokens), and our decoder islightweight and reconstructs the input from the latent representation along with mask tokens (Figure 1). Shifting the mask tokens to the small decoder in our asymmetric encoder-decoder results in a large reduction in computation. Under this design, a very high masking ratio (e.g., 75%) can achieve a win-win scenario: it optimizes accuracy while allowing the encoder to process only a small portion (e.g., 25%) of patches. This can reduce overall pre-training time by 3× or more and likewise reduce memory consumption, enabling us to easily scale our MAE to large models. Our MAE learns very high-capacity models that generalize well. With MAE pre-training, we can train data-hungry models like ViT-Large/-Huge [16] on ImageNet-1K with improved generalization performance. With a vanilla ViT-Huge model, we achieve 87.8% accuracy when fine tuned on ImageNet-1K. This outperforms all previous results that use only ImageNet-1K data. We also evaluate transfer learning on object detection, instance segmentation, and semantic segmentation. In these tasks, our pre-training achieves better results than its supervised pre-training counterparts, and more importantly, we observe significant gains by scaling up models. These observations are aligned with those witnessed in self-supervised pre-training in NLP [14, 47, 48, 4] and we hope that they will enable our field to explore a similar trajectory.

(iii) **自动编码器的解码器**，它**映射潜在的表示**返回到输入，在重建文本和图像之间起着不同的作用。**在视觉上，解码器 重建像素，因此其输出的语义水平低于普通识别任务。这和语言是相反的，解码器预测缺失的单词 包含丰富的语义信息。**虽然在 BERT 中，解码器可以是**微不足道**的（一个 MLP）[14]，但我们发现对于图像，**解码器设计在决定学习潜在信息的语义级别起着重要作用。** 在此分析的推动下，我们提出了一种简单、有效、 和可扩展形式的掩码自动编码器 (MAE) 视觉表征学习。我们的 MAE 掩码随机 来自输入图像的补丁并重建缺失的 像素空间中的补丁。它具有**非对称**编码-解码器设计。我们的编码器只在可见的 补丁的子集（没有掩码标记），我们的解码器是轻量级的，并从潜在表示中重构输入以及掩码标记（图 1）。**从掩码标记到小的解码器，在我们的非对称的编码器-解码器中导致计算量大大减少。** 在这种设计下，一个非常高的掩码率（例如，75%）可以 实现双赢：**它优化了准确性，同时允许编码器只处理一小部分（例如， 25%) 的补丁。**这可以**减少整体的预训练时间 3 倍或更多**，同样**减少内存消耗**， 使我们能够轻松地将我们的 MAE 扩展到大型模型。 我们的 MAE 学习了泛化能力非常高的模型。通过 MAE 预训练，我们可以在 ImageNet-1K 上训练需要大量数据的模型，例如 ViT-Large/-Huge [16] 具有改进的泛化性能。 vanilla ViT-Huge 模型，我们在 ImageNet-1K 上微调时达到 87.8% 的准确率。这优于之前仅使用 ImageNet-1K 数据的所有结果。我们还评估 对象检测、实例分割的迁移学习， 和语义分割。在这些任务中，我们的预训练 比其监督的预训练同行取得更好的结果，更重要的是，我们观察到显着的收益 通过扩大模型。这些观察结果一致 与那些在 NLP 的自我监督预训练中见证的人 [14, 47, 48, 4] 我们希望它们能让我们的领域 探索类似的轨迹。

3. Related Work

***Masked language modeling** and its autoregressive counterparts, e.g., BERT [14] and GPT [47, 48, 4], are highly successful methods for pre-training in NLP. These methodshold out a portion of the input sequence and train models to predict the missing content. These methods have beenshown to scale excellently [4] and a large abundance of evidence indicates that these pre-trained representations generalize well to various downstream tasks.*

掩码语言建模及其自回归对应物，例如 BERT [14] 和 GPT [47, 48, 4]，是高度 NLP预训练的成功方法。这些方法 保留一部分输入序列并训练模型 预测缺失的内容。 这些方法已 显示出很好的扩展性[4]，并且大量证据表明这些预训练的表示可以很好地推广到各种下游任务。

**Autoencoding** is a classical method for learning representations. It has an encoder that maps an input to a latent representation and a decoder that reconstructs the input. For example, PCA and k-means are autoencoders [29]. Denoisingautoencoders (DAE) [58] are a class of autoencoders thatcorrupt an input signal and learn to reconstruct the original, uncorrupted signal. A series of methods can be thought of as a generalized DAE under different corruptions, e.g., masking pixels [59, 46, 6] or removing color channels [70].Our MAE is a form of denoising autoencoding, but differentfrom the classical DAE in numerous ways.**

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构输入的解码器。 例如，PCA 和 k-means 是自动编码器 [29]。 去噪 自动编码器（DAE）[58] 是一类自动编码器， 破坏输入信号并学习重建原始的未破坏信号。 可以想到一系列方法 作为不同损坏下的广义 DAE，例如， 屏蔽像素 [59, 46, 6] 或移除颜色通道 [70]。 我们的 MAE 是一种去噪自动编码的形式，但不同的是 以多种方式从经典的 DAE 中提取。

**Masked image encoding** methods learn representations from images corrupted by masking. The pioneering work of [59] presents masking as a noise type in DAE. Context Encoder [46] inpaints large missing regions using convolutional networks. Motivated by the success in NLP, related recent methods [6, 16, 2] are based on Transformers [57]. iGPT [6] operates on sequences of pixels and predicts unknown pixels. The ViT paper [16] studies masked patch prediction for self-supervised learning. Most recently, BEiT [2] proposes to predict discrete tokens [44, 50]

掩码图像编码方法从被掩码损坏的图像学习信息。 **开创性的工作 [59] 将掩码作为 DAE 中的一种噪声类型。** **语境 编码器** [46] 使用卷积网络**修复大的缺失区域**。 受到 NLP 成功的激励，相关 最近的方法 [6, 16, 2] 基于 Transformer [57]。 iGPT [6] **对像素序列进行操作并预测未知像素。** ViT 论文 [16] 研究了掩码补丁 自我监督学习的预测。 最近，BEiT [2] 提出**预测离散标记** [44, 50]

**Self-supervised learning** approaches have seen significant interest in computer vision, often focusing on different pretext tasks for pre-training [15, 61, 42, 70, 45, 17]. Recently, contrastive learning [3, 22] has been popular, e.g.,[62, 43, 23, 7], which models image similarity and dissimilarity (or only similarity [21, 8]) between two or more views. Contrastive and related methods strongly depend on data augmentation [7, 21, 8]. Autoencoding pursues a conceptually different direction, and it exhibits different behaviors as we will present.

自我监督学习方法已取得显著成效 对计算机视觉感兴趣，通常专注于预训练的不同**修饰任务**[15、61、42、70、45、17]。 最近，**对比学习** [3, 22] 很流行，例如， [62, 43, 23, 7]，它模拟两个或多个图像之间的相似性和不相似性（或仅相似性 [21, 8]） 意见。 对比和相关方法强烈依赖于 数据增强 [7, 21, 8]。 **自动编码追求一个概念上不同的方向，它表现出我们将要呈现的不同行为。**

4. Approach

our masked autoencoder (MAE) is a simple autoencoding approach that reconstructs the original signal given its partial observation. Like all autoencoders, our approach has an encoder that maps the observed signal to a latent representation, and a decoder that reconstructs the original signal from the latent representation. Unlike classical autoencoders, we adopt an asymmetric design that allows the encoder to operate only on the partial, observed signal (without mask tokens) and a lightweight decoder that reconstructs the full signal from the latent representation and mask tokens. Figure 1 illustrates the idea, introduced next.

我们的掩码自动编码器 (MAE) 是一种简单的自动编码方法，可以根据其重构原始信号 部分观察。 像所有自动编码器一样，我们的方法 具有将观察到的信号映射到潜在信号的编码器 表示，以及从潜在表示中重建原始信号的解码器。 不同于经典自动编码器，我们采用非对称设计，允许 编码器仅对部分观察到的信号进行操作 （没有掩码标记）和一个轻量级解码器，它从潜在表示重建完整信号和 掩码令牌。 图 1 说明了这个想法，接下来介绍。

Masking*. Following ViT [16], we divide an image into regular non-overlapping patches. Then we sample a subset of patches and mask (i.e., remove) the remaining ones. Our sampling strategy is straightforward: we sample random patches without replacement, following a uniform distribution. We simply refer to this as “random sampling”. Random sampling with a high masking ratio (i.e., the ratio of removed patches) largely eliminates redundancy, thus creating a task that cannot be easily solved by extrapolation from visible neighboring patches (see Figures 2 – 4). The uniform distribution prevents a potential center bias (i.e., more masked patches near the image center). Finally, the highly sparse input creates an opportunity for designing an efficient encoder, introduced next.*

掩码。 在 ViT [16] 之后，我们将图像划分为规则的非重叠块。 然后我们采样一个子集 补丁和掩码（即删除）其余的。 我们的 抽样策略很简单：我们随机抽样 补丁无需更换，遵循均匀分布。 我们简单地将其称为“随机抽样”。 具有高掩蔽率（即移除的补丁的比率）的随机采样在很大程度上消除了冗余，因此创建一个无法通过外推轻松解决的任务 从可见的相邻斑块（见图 2-4）。 这 均匀分布可防止潜在的中心偏差（即 图像中心附近有更多蒙版补丁）。 最后， 高度稀疏的输入为设计一个 高效编码器，接下来介绍。

MAE encoder*. Our encoder is a ViT [16] but applied only on visible, unmasked patches. Just as in a standard ViT, our encoder embeds patches by a linear projection with added positional embeddings, and then processes the resulting set via a series of Transformer blocks. However, our encoder only operates on a small subset (e.g., 25%) of the full set. Masked patches are removed; no mask tokens are used. This allows us to train very large encoders with only a fraction of compute and memory. The full set is handled by a lightweight decoder, described next.*

MAE编码器。 我们的编码器是 ViT [16]，**但仅适用于 在可见的、未遮盖的补丁上。** 就像在标准 ViT 中一样，我们的 **编码器通过线性投影嵌入补丁，并添加 位置嵌入**，**然后处理结果集 通过一系列转换块。** 但是，我们的编码器 只对整个集合的一小部分（例如 25%）进行操作。 **被屏蔽的补丁被移除； 不使用掩码标记。** 这使我们能够只用一小部分计算和内存来训练非常大的编码器。 全套由一个处理 轻量级解码器，接下来描述。

MAE decoder. The input to the MAE decoder is the full set of tokens consisting of (i) encoded visible patches, and (ii) mask tokens. See Figure 1. Each mask token [14] is a shared, learned vector that indicates the presence of a missing patch to be predicted. We add positional embeddings to all tokens in this full set; without this, mask tokens would have no information about their location in the image. The decoder has another series of Transformer blocks. The MAE decoder is only used during pre-training to perform the image reconstruction task (only the encoder is used to produce image representations for recognition). Therefore, the decoder architecture can be flexibly designed in a manner that is independent of the encoder design. We experiment with very small decoders, narrower and shallower than the encoder. For example, our default decoder has <10% computation per token vs. the encoder. With this asymmetrical design, the full set of tokens are only processed by the lightweight decoder, which significantly reduces pre-training time.

MAE解码器。 MAE解码器的输入是完整的 由 (i) 编码可见补丁组成的标记集，以及 (ii) 掩码标记。参见图 1。每个掩码标记 [14] 是一个 共享的、学习的向量，表示存在缺失 要预测的补丁。我们将位置嵌入添加到 全套中的所有代币；没有这个，掩码尿急会 没有关于他们在图像中的位置的信息。这 解码器有另一个系列的转换块。 MAE 解码器仅在预训练期间使用 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，解码器架构可以灵活设计 以独立于编码器设计的方式。我们 尝试使用非常小的解码器，比编码器更窄更浅。例如，我们的默认解码器 与编码器相比，每个标记的计算量 <10%。有了这个 非对称设计，全套token仅由轻量级解码器处理，显着减少预训练时间。

Reconstruction target. *Our MAE reconstructs the input by predicting the pixel values for each masked patch. Each element in the decoder’s output is a vector of pixel values representing a patch. The last layer of the decoder is a linear projection whose number of output channels equals the number of pixel values in a patch. The decoder’s output is reshaped to form a reconstructed image. Our loss function computes the* *mean squared error (MSE) between the reconstructed and original images in the pixel space. We compute the loss only on masked patches, similar to BERT [14].1 We also study a variant whose reconstruction target is the normalized pixel values of each masked patch. Specifically, we compute the mean and standard deviation of all pixels in a patch and use them to normalize this patch. Using normalized pixels as the reconstruction target improves representation quality in our experiments.*

重建目标。 我们的 MAE 重构输入 通过预测每个蒙面补丁的像素值。 每个 解码器输出中的元素是像素值的向量 代表一个补丁。 解码器的最后一层是线性投影，其输出通道数等于 补丁中的像素值数量。 解码器的输出是 重新整形以形成重建图像。 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)。 我们计算 仅在蒙面补丁上的损失，类似于 BERT [14].1 我们还研究了一个变体，其重建目标是 每个掩码补丁的归一化像素值。 具体来说，我们计算所有的均值和标准差 补丁中的像素并使用它们来规范化该补丁。 使用归一化像素作为重建目标可以提高 我们实验中的表示质量。

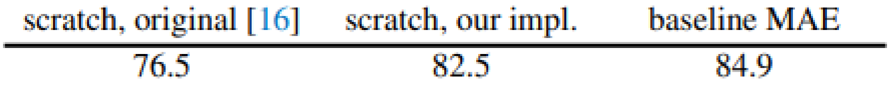
Simple implementation. *Our MAE pre-training can be implemented efficiently, and importantly, does not require any specialized sparse operations. First we generate a token for every input patch (by linear projection with an added positional embedding). Next we randomly shuffle the list of tokens and remove the last portion of the list, based on the masking ratio. This process produces a small subset of tokens for the encoder and is equivalent to sampling patches without replacement. After encoding, we append a list of mask tokens to the list of encoded patches, and unshuffle this full list (inverting the random shuffle operation) to align all tokens with their targets. The decoder is applied to this full list (with positional embeddings added). As noted, no sparse operations are needed. This simple implementation introduces negligible overhead as the shuffling and unshuffling operations are fast.*

简单的实现。 我们的 MAE 预训练可以高效实施，重要的是，不需要任何 专门的稀疏运算。 首先我们生成一个token 每个输入补丁（通过添加位置嵌入的线性投影）。 接下来我们随机打乱列表 标记并删除列表的最后一部分，基于 掩蔽率。 此过程为编码器生成一小部分令牌，相当于采样补丁 无需更换。 编码后，我们附加一个列表 将令牌掩码到编码补丁列表中，并取消随机播放 这个完整列表（反转随机洗牌操作）对齐 所有令牌及其目标。 解码器应用于此 完整列表（添加了位置嵌入）。 如前所述，没有 需要稀疏操作。 这个简单的实现 引入了可忽略的开销，因为改组和取消改组操作很快。

5.ImageNet Experiments

We do self-supervised pre-training on the ImageNet-1K (IN1K) [13] training set. Then we do supervised training to evaluate the representations with (i) end-to-end fine-tuning or (ii) linear probing. We report top-1 validation accuracy of a single 224×224 crop. Details are in Appendix A.1. Baseline: ViT-Large. We use ViT-Large (ViT-L/16) [16] as the backbone in our ablation study. ViT-L is very big (an order of magnitude bigger than ResNet-50 [25]) and tends to overfit. The following is a comparison between ViT-L trained from scratch vs. fine-tuned from our baseline MAE:

我们在 ImageNet-1K 上进行自我监督的预训练 (IN1K) [13] 训练集。 然后我们进行监督训练 通过 (i) 端到端微调评估表示 或 (ii) 线性探测。 我们报告 top-1 验证准确性 单个 224×224 图像。 详细信息在附录 A.1 中。 基线：ViT-大。 我们使用 ViT-Large (ViT-L/16) [16] 作为我们消融研究的骨干。 ViT-L 非常大（一个 数量级大于 ResNet-50 [25]）并且趋于 过拟合。 下面是ViT-L之间的对比 从头开始训练与从我们的基线 MAE 微调：



scratch, original [16]	scratch, our impl.	baseline MAE
76.5	82.5	84.9

We note that it is nontrivial to train supervised ViT-L from scratch and a good recipe with strong regularization is needed (82.5%, see Appendix A.2). Even so, our MAE pre-training contributes a big improvement. Here fine-tuning is only for 50 epochs (vs. 200 from scratch), implying that the fine-tuning accuracy heavily depends on pre-training.

我们注意到训练有监督的 ViT-L 从 从头开始，一个具有强正则化的好方法是 需要（82.5%，见附录 A.2）。 尽管如此，我们的 MAE 预训练还是做出了很大的改进。 这里微调是 仅适用于 50 个 epoch（相对于从头开始的 200 个），这意味着 微调精度很大程度上取决于预训练。

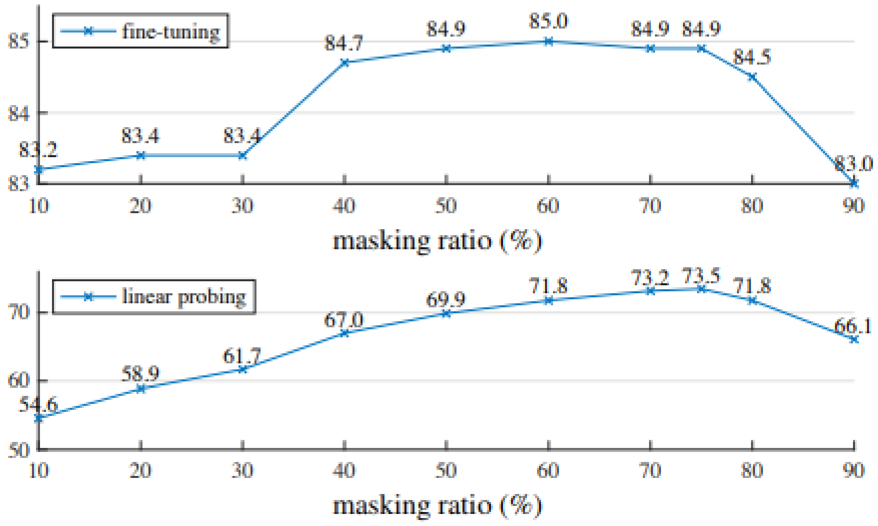
4.1. Main Properties

We ablate our MAE using the default settings in Table 1(see caption). Several intriguing properties are observed.

Masking ratio.

*Figure 5 shows the influence of the masking ratio. The optimal ratios are surprisingly high. The ratio of 75% is good for both linear probing and fine-tuning.This behavior is in contrast with BERT [14], whose typicamasking ratio is 15%. Our masking ratios are also muchhigher than those in related works [6, 16, 2] in computervision (20% to 50%).The model infers missing patches to produce different,yet plausible, outputs (Figure 4). It makes sense of thegestalt of objects and scenes, which cannot be simply completed by extending lines or textures. We hypothesize thatthis reasoning-like behavior is linked to the learning of useful representations.**Figure 5 also shows that linear probing and fine-tuning results follow different trends. For linear probing, the accuracy increases steadily with the masking ratio until the sweet point: the accuracy gap is up to ~20% (54.6% vs. 73.5%). For fine-tuning, the results are less sensitive to the ratios, and a wide range of masking ratios (40–80%) work well. All fine-tuning results in Figure 5 are better than training from scratch (82.5%).*

4.1。 主要属性 我们使用表 1 中的默认设置消融我们的 MAE （见标题）。 观察到几个有趣的特性。 掩蔽率。 图 5 显示了掩蔽率的影响。 最佳比率惊人地高。 75% 的比率对线性探测和微调都有好处。 这种行为与 BERT [14] 形成对比，BERT [14] 的典型 掩蔽率为 15%。 我们的掩蔽率也很高 高于计算机相关著作 [6, 16, 2] 视力（20% 到 50%）。 该模型推断缺失的补丁产生不同的， 看似合理的输出（图 4）。这很有意义 物体和场景的完形，不能简单地通过延伸线条或纹理来完成。 我们假设 这种类似推理的行为与学习有用的表示有关。图 5 还显示了线性探测和微调 结果遵循不同的趋势。 对于线性探测，精度随着掩蔽率稳步增加，直到 甜蜜点：准确率差距高达 ~20%（54.6% vs. 73.5%）。 对于微调，结果对 比率，以及范围广泛的掩蔽率 (40–80%) 工作 好吧。 图 5 中的所有微调结果都优于从头开始训练 (82.5%)。



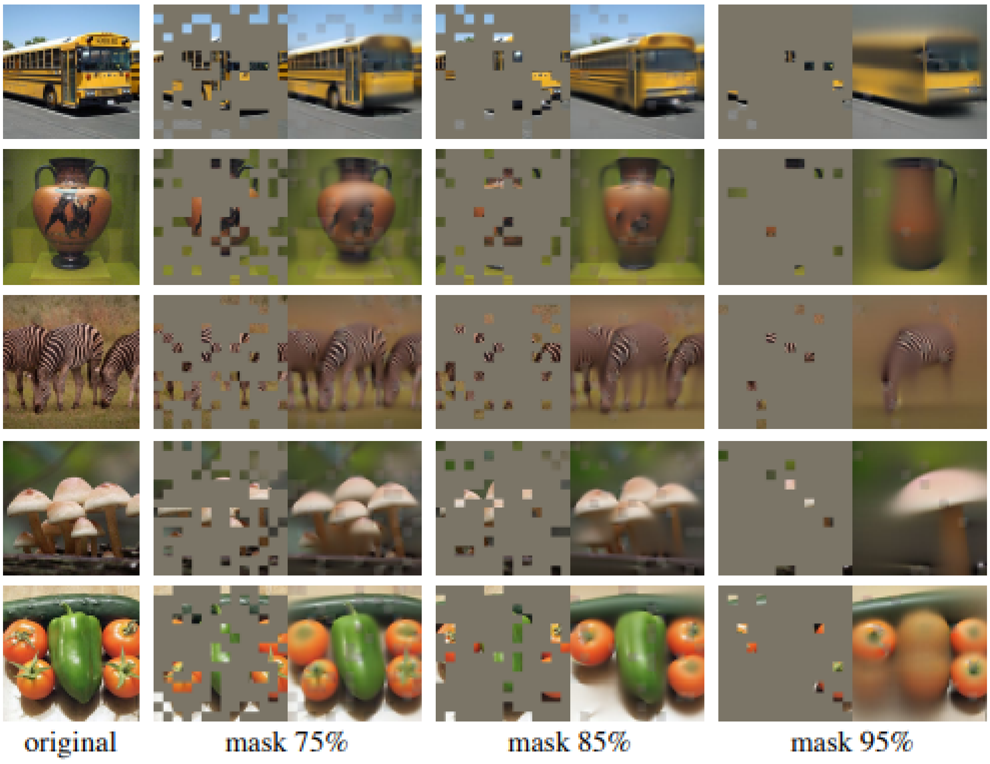


Figure 4. Reconstructions of ImageNet *validation* images using an MAE pre-trained with a masking ratio of 75% but applied on inputs with higher masking ratios. The predictions differ plausibly from the original images, showing that the method can generalize.

Decoder design.

Our MAE decoder can be flexibly designed, as studied in Table 1a and 1b.Table 1a varies the decoder depth (number of Transformer blocks). A sufficiently deep decoder is important for linear probing. This can be explained by the gap between a pixel reconstruction task and a recognition task: the last several layers in an autoencoder are more specialized for reconstruction, but are less relevant for recognition. A reasonably deep decoder can account for the reconstruction specialization, leaving the latent representations at a more abstract level. This design can yield up to 8% improvement in linear probing (Table 1a, 'lin'). However, if fine-tuning is used, the last layers of the encoder can be tuned to adapt to the recognition task. The decoder depth is less influential for improving fine-tuning (Table 1a, 'ft'). Interestingly, our MAE with a single-block decoder can perform strongly with fine-tuning (84.8%). Note that a single Transformer block is the minimal requirement to propagate information from visible tokens to mask tokens. Such a small decoder can further speed up training. In Table 1b we study the decoder width (number of channels). We use 512-d by default, which performs well under fine-tuning and linear probing. A narrower decoder also works well with fine-tuning. Overall, our default MAE decoder is lightweight. It has 8 blocks and a width of 512-d (gray in Table 1). It only has 9% FLOPs per token vs. ViT-L (24 blocks, 1024-d). As such, while the decoder processes all tokens, it is still a small fraction of the overall compute.

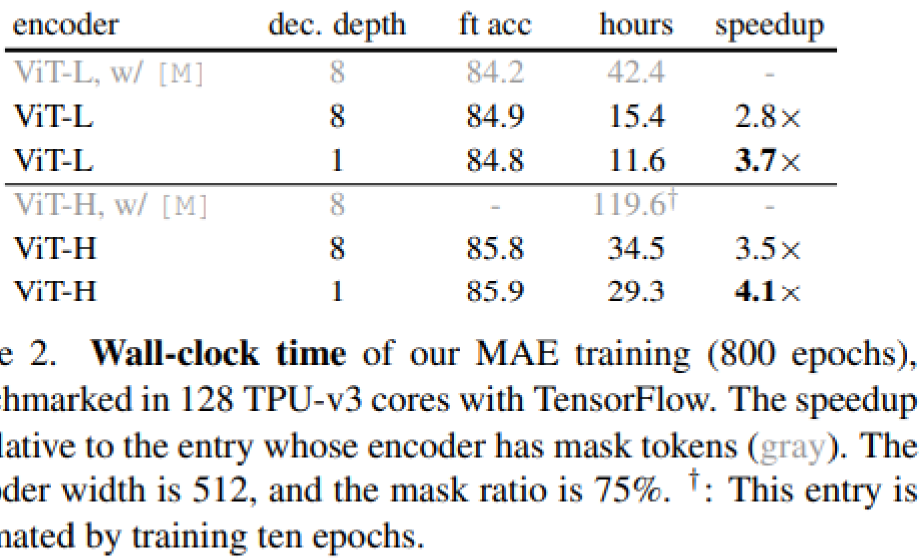
解码器设计。我们的 MAE 解码器可以灵活设计，如表 1a 和 1b 所示。 表 1a 改变了解码器深度（变压器块的数量）。足够深的解码器很重要 用于线性探测。这可以通过像素重建任务和识别任务之间的差距来解释： 自动编码器中的最后几层更专业 用于重建，但与识别不太相关。一种 相当深的解码器可以解释重建 专业化，将潜在表示留在更多 抽象层次。这种设计可以产生高达 8% 的改进 在线性探测中（表 1a，“lin”）。但是，如果微调 使用时，可以调整编码器的最后几层以适应 到识别任务。解码器深度影响较小 用于改进微调（表 1a，'ft'）。 有趣的是，我们的带有单块解码器的 MAE 可以 通过微调

（84.8%）表现强劲。请注意，单个 Transformer 块是将信息从可见令牌传播到掩码令牌的最低要求。这样的 一个小的解码器可以进一步加快训练速度。 在表 1b 中，我们研究了解码器宽度（通道数）。我们默认使用 512-d，它在微调和线性探测下表现良好。更窄的解码器 与微调配合得很好。 总的来说，我们默认的 MAE 解码器是轻量级的。**它有 8 个块，宽度为 512-d**（表 1 中的灰色）。它只是 与 ViT-L（24 个区块，1024-d）相比，每个令牌有 9% 的 FLOP。 因此，虽然解码器处理所有令牌，但它仍然是一个 整体计算的一小部分。

Mask token.

An important design of our MAE is to skip the mask token [M] in the encoder and apply it later in the lightweight decoder. Table 1c studies this design. If the encoder uses mask tokens, it performs worse: its accuracy drops by 14% in linear probing. In this case, there is a gap between pre-training and deploying: this encoder has a large portion of mask tokens in its input in pretraining, which does not exist in uncorrupted images. This gap may degrade accuracy in deployment. By removing the mask token from the encoder, we constrain the encoder to always see real patches and thus improve accuracy. Moreover, by skipping the mask token in the encoder, we greatly reduce training computation. In Table 1c, we reduce the overall training FLOPs by 3.3×. This leads to a 2.8× wall-clock speedup in our implementation (see Table 2). The wall-clock speedup is even bigger (3.5–4.1×), for a smaller decoder (1-block), a larger encoder (ViT-H), or both. Note that the speedup can be >4× for a masking ratio of 75%, partially because the self-attention complexity is quadratic. In addition, memory is greatly reduced, which can enable training even larger models or speeding up more by large-batch training. The time and memory efficiency makes our MAE favorable for training very large models.

掩码令牌。我们的 MAE 的一个**重要设计是跳过 编码器中的掩码标记 [M] 并稍后在 轻量级解码器应用。** 表 1c 研究了这种设计。 如果编码器使用掩码标记，它的性能会更差：它的 线性探测的精度下降 14%。在这种情况下， 预训练和部署之间存在差距：该编码器在预训练的输入中有很大一部分掩码标记，这在未损坏的图像中不存在。这 差距可能会降低部署的准确性。**通过删除 从编码器屏蔽令牌，我们将编码器约束为 总是看到真正的补丁，从而提高准确性。** 此外，**通过跳过编码器中的掩码标记， 我们大大减少了训练计算**。在表 1c 中，我们 将整体训练 FLOPs 减少 3.3 倍。这导致 在我们的实现中实现了 2.8 倍的挂钟加速（见表 2）。挂钟加速甚至更大（3.5-4.1×）， 对于较小的解码器（1-block），较大的编码器（ViT-H）， 或两者。请注意，对于掩蔽，加速可以大于 4 倍 75% 的比例，部分原因是自注意力的复杂性 是二次的。此外，内存大大减少，这 可以训练更大的模型或加速更多 通过大批量训练。时间和内存效率 使我们的 MAE 有利于训练非常大的模型。



encoder	dec. depth	ft acc	hours	speedup
ViT-L, w/ [M]	8	84.2	42.4	-
ViT-L	8	84.9	15.4	2.8×
ViT-L	1	84.8	11.6	**3.7**×
ViT-H, w/ [M]	8	-	$119.6^{\dagger}$	-
ViT-H	8	85.8	34.5	3.5×
ViT-H	1	85.9	29.3	**4.1**×

Table 2. **Wall-clock time** of our MAE training (800 epochs), benchmarked in 128 TPU-v3 cores with TensorFlow. The speedup is relative to the entry whose encoder has mask tokens (gray). The decoder width is 512, and the mask ratio is 75%. †: This entry is estimated by training ten epochs.

Reconstruction target.

We compare different reconstruction targets in Table 1d. Our results thus far are based on pixels without (per-patch) normalization. Using pixels with normalization improves accuracy. This per-patch normalization enhances the contrast locally. In another variant, we perform PCA in the patch space and use the largest PCA coefficients (96 here) as the target. Doing so degrades accuracy. Both experiments suggest that the high-frequency components are useful in our method. We also compare an MAE variant that predicts tokens, the target used in BEiT [2]. Specifically for this variant, we use the DALLE pre-trained dVAE [50] as the tokenizer, following [2]. Here the MAE decoder predicts the token indices using cross-entropy loss. This tokenization improves fine-tuning accuracy by 0.4% vs. unnormalized pixels, but has no advantage vs. normalized pixels. It also reduces linear probing accuracy. In §5 we further show that tokenization is not necessary in transfer learning. Our pixel-based MAE is much simpler than tokenization. The dVAE tokenizer requires one more pre-training stage, which may depend on extra data (250M images [50]). The dVAE encoder is a large convolutional network (40% FLOPs of ViT-L) and adds nontrivial overhead. Using pixels does not suffer from these problems.

重建目标。我们在表 1d 中比较了不同的重建目标。到目前为止，我们的结果基于 没有（每个补丁）归一化的像素。**使用像素 归一化提高了准确性**。这种逐块归一化在局部增强了对比度。在另一个变体中，我们 在**补丁空间中执行 PCA 并使用最大的 PCA 系数**（此处为 96）作为目标。**这样做会降低准确性**。两个实验都表明**高频 组件在我们的方法中很有用。** 我们还比较了预测标记的 MAE 变体， BEiT [2] 中使用的目标。专门针对此变体， 我们使用 DALLE 预训练的 dVAE [50] 作为分词器， 遵循[2]。这里 MAE 解码器使用交叉熵损失来预测令牌索引。这种标记化改进了 与未归一化像素相比，微调精度为 0.4%，但与归一化像素相比没有优势。它还降低了线性探测精度。在第 5 节中，**我们进一步表明在迁移学习中不需要标记化。 我们基于像素的 MAE 比标记化简单得多。** dVAE 分词器需要再进行一次预训练 阶段，这可能取决于额外的数据（250M 图像 [50]）。 dVAE 编码器是一个大型卷积网络（40% ViT-L 的 FLOPs）并增加了不平凡的开销。使用像素不会遇到这些问题。

Data augmentation.

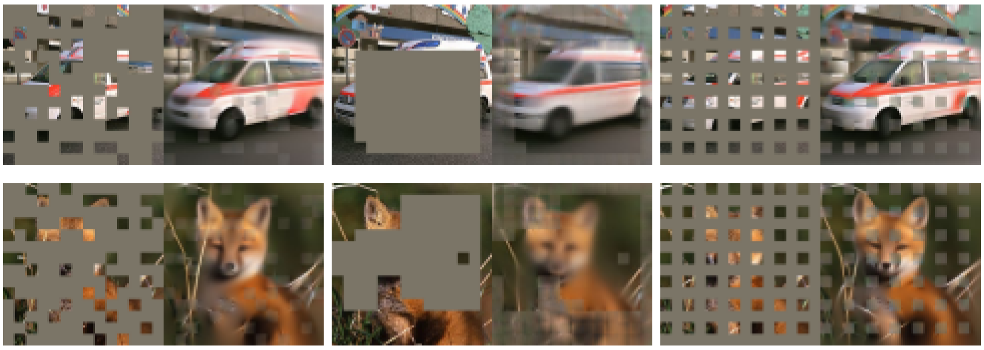
Table 1e studies the influence of data augmentation on our MAE pre-training. Our MAE works well using cropping-only augmentation, either fixed-size or random-size (both having random horizontal flipping). Adding color jittering degrades the results and so we do not use it in other experiments. Surprisingly, our MAE behaves decently even if using no data augmentation (only center-crop, no flipping). This property is dramatically different from contrastive learning and related methods [62, 23, 7, 21], which heavily rely on data augmentation. It was observed [21] that using cropping-only augmentation reduces the accuracy by 13%and 28% respectively for BYOL [21] and SimCLR [7]. In addition, there is no evidence that contrastive learning can work without augmentation: the two views of an image are the same and can easily satisfy a trivial solution. In MAE, the role of data augmentation is mainly performed by random masking (ablated next). The masks are different for each iteration and so they generate new training samples regardless of data augmentation. The pretext task is made difficult by masking and requires less augmentation to regularize training.

数据增强。表 1e 研究了数据的影响 增强我们的 MAE 预训练。 **我们的 MAE 使用仅裁剪增强效果很好，无论是固定大小还是随机大小（都具有随机 水平翻转）。添加颜色抖动会降低结果，因此我们不会在其他实验中使用它。** 令人惊讶的是，我们的 MAE 表现得很好，即使使用 没有数据增强（只有中心裁剪，没有翻转）。这 属性与对比学习有很大不同 和相关的方法 [62, 23, 7, 21]，它们严重依赖 关于数据增强。据观察[21]，使用 仅裁剪增强将 BYOL [21] 和 SimCLR [7] 的准确率分别降低了 13% 和 28%。在此外，**没有证据表明对比学习可以 无需增强即可工作**：图像的两个视图是 相同并且可以很容易地满足一个平凡的解决方案。 在 MAE 中，**数据增强的作用主要是通过随机掩码（ablated next）来完成的。掩码是 每次迭代都不同**，因此它们会生成新的训练 样本，**无论数据增强如何。借口任务 通过掩蔽变得困难并且需要较少的增强 使培训常态化。**

Mask sampling strategy.

In Table 1f we compare different mask sampling strategies, illustrated in Figure 6. The block-wise masking strategy, proposed in [2], tends to remove large blocks (Figure 6 middle). Our MAE with block-wise masking works reasonably well at a ratio of 50%, but degrades at a ratio of 75%. This task is harder than that of random sampling, as a higher training loss is observed. The reconstruction is also blurrier. We also study grid-wise sampling, which regularly keeps one of every four patches (Figure 6 right). This is an easier task and has lower training loss. The reconstruction is sharper. However, the representation quality is lower. Simple random sampling works the best for our MAE. It allows for a higher masking ratio, which provides a greater speedup benefit while also enjoying good accuracy.

掩码采样策略。 在表1f中，我们比较了不同的 掩码采样策略，如图6所示。 [2]中**提出的分块掩蔽策略倾向于 删除大块**（图6中间）。 我们的MAE逐块掩码的效果相当好，比例为50%，但以75%的比例降解。 这个任务比较难 相比于随机抽样，因为更高的训练损失是 观察到的。 重建也更加模糊。 我们还研究了网格抽样，它定期保持 每四个补丁之一（图6右）。 这是一项更容易的任务，并且训练损失更低。 重建是 更锋利。 但是，表示质量较低。 **简单随机抽样最适合我们的MAE。 它 允许更高的掩蔽率，从而提供更大的 加速的好处，同时也享有良好的准确性。**



random 75% block 50% grid 75%

Figure 6. **Mask sampling strategies** determine the pretext task difficulty, influencing reconstruction quality and representations (Table 1f). Here each output is from an MAE trained with the specified masking strategy. Left: random sampling (our default). Middle: block-wise sampling [2] that removes large random blocks. Right: grid-wise sampling that keeps one of every four patches. Images are from the validation set.

Training schedule.

Our ablations thus far are based on 800-epoch pre-training. Figure 7 shows the influence of the training schedule length. The accuracy improves steadily with longer training. Indeed, we have not observed saturation of linear probing accuracy even at 1600 epochs. This behavior is unlike contrastive learning methods, e.g., MoCo v3 [9] saturates at 300 epochs for ViT-L. Note that the MAE encoder only sees 25% of patches per epoch, while in contrastive learning the encoder sees 200% (two crop) or even more (multi-crop) patches per epoch.

训练安排。 到目前为止，我们的消融是基于800 epoch预训练。 图7显示了 训练计划长度。 精度稳步提高 训练时间更长。 事实上，即使在1600个epoch时，我们也没有观察到线性探测精度的饱和。 这种行为不同于对比学习方法，例如， MoCo v3 [9]在ViT-L的300个epoch处饱和。 注意MAE编码器每个epoch只能看到25%的补丁， 而在对比学习中，编码器在每个时期看到200%（两次裁剪）甚至

更多（多裁剪）的补丁。

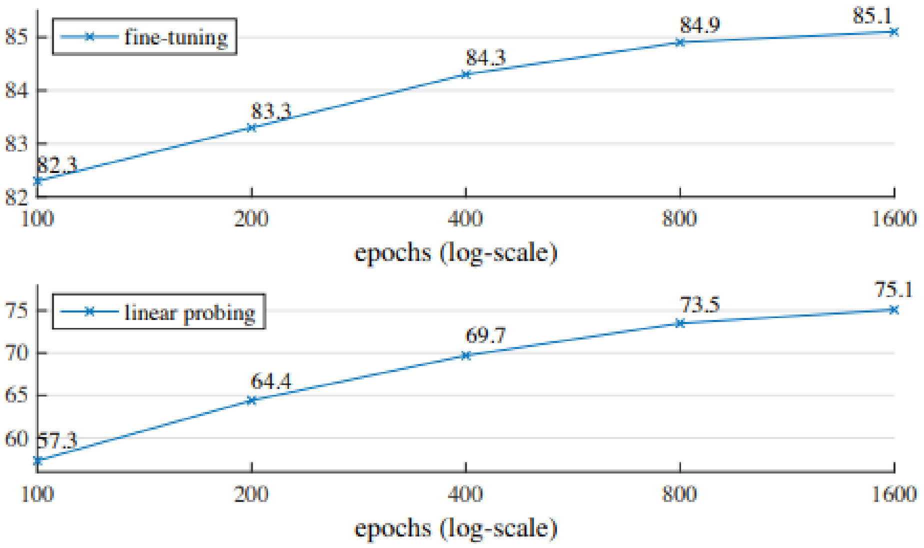
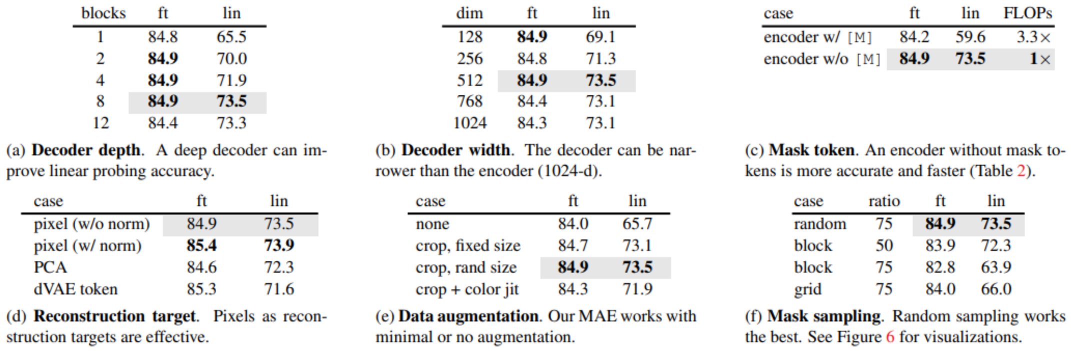


Figure 7. **Training schedules**. A longer training schedule gives a noticeable improvement. Here each point is a full training schedule. The model is ViT-L with the default setting in Table 1.



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

(a) **Decoder depth**. A deep decoder can improve linear probing accuracy.

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

(b) **Decoder width**. The decoder can be narrower than the encoder (1024-d).

case	ft	lin	FLOPs
encoder w/ `[M]`	84.2	59.6	3.3×
encoder w/o `[M]`	**84.9**	**73.5**	**1**×

(c) **Mask token**. An encoder without mask tokens is more accurate and faster (Table 2).

case	ft	lin
pixel (w/o norm)	84.9	73.5
pixel (w/ norm)	**85.4**	**73.9**
PCA	84.6	72.3
dVAE token	85.3	71.6

(d) **Reconstruction target**. Pixels as reconstruction targets are effective.

case	ft	lin
none	84.0	65.7
crop, fixed size	84.7	73.1
crop, rand size	**84.9**	**73.5**
crop + color jit	84.3	71.9

(e) **Data augmentation**. Our MAE works with minimal or no augmentation.

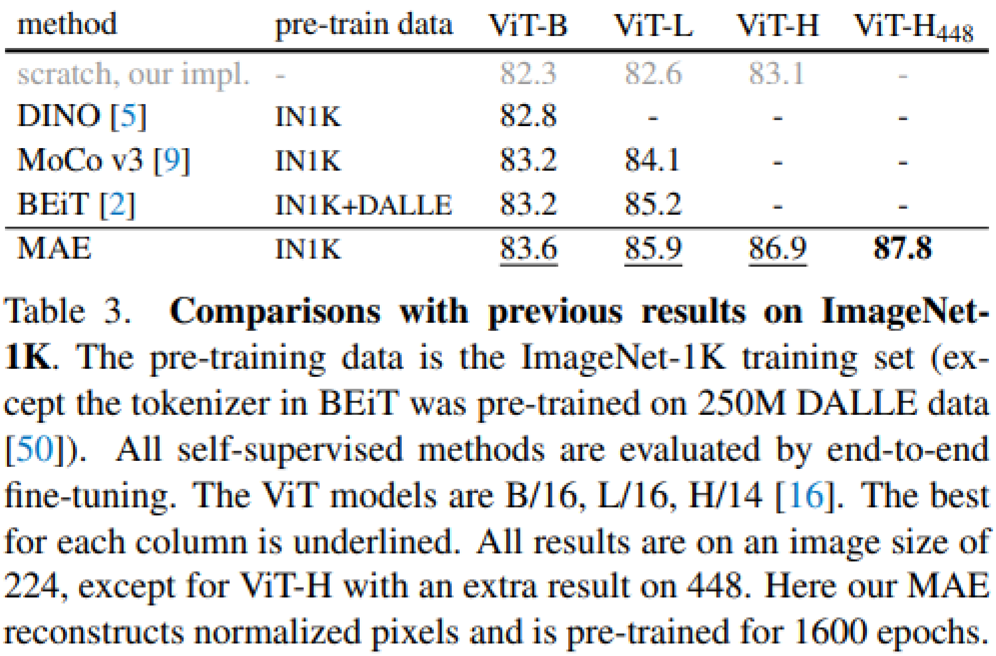
case	ratio	ft	lin
random	75	**84.9**	**73.5**
block	50	83.9	72.3
block	75	82.8	63.9
grid	75	84.0	66.0

(f) **Mask sampling**. Random sampling works the best. See Figure 6 for visualizations.

4.2. Comparisons with Previous Results

*Comparisons with self-supervised methods. In Table 3 we compare the fine-tuning results of self-supervised ViT models. For ViT-B, all methods perform closely. For ViT-L,the gaps among methods are bigger, suggesting that a challenge for bigger models is to reduce overfitting. Our MAE can scale up easily and has shown steady improvement from bigger models. We obtain 86.9% accuracy using ViT-H (224 size). By fine-tuning with a 448 size, we achieve 87.8% accuracy, using only IN1K data. The previous best accuracy, among all methods using only IN1K data, is 87.1% (512 size) [67], based on advanced networks. We improve over the state-of-the-art by a nontrivial margin in the highly competitive benchmark of IN1K (no external data). Our result is based on vanilla ViT, and we expect advanced networks will perform better. Comparing with BEiT [2], our MAE is more accurate while being simpler and faster. Our method reconstructs pixels, in contrast to BEiT that predicts tokens: BEiT reported a 1.8% degradation [2] when reconstructing pixels with ViT-B.2 We do not need dVAE pre-training. Moreover, our MAE is considerably faster (3.5× per epoch) than BEiT, for the reason as studied in Table 1c.**The MAE models in Table 3 are pre-trained for 1600 epochs for better accuracy (Figure 7). Even so, our total pre-training time islessthan the other methods when trained on the same hardware. For* *example, training ViT-L on 128 TPU-v3 cores, our MAE's training time is 31 hours for 1600 epochs and MoCo v3's is 36 hours for 300 epochs [9].*

4.2.与先前结果的比较 与自我监督方法的比较。在表 3 我们比较了自监督 ViT 的微调结果 楷模。对于 ViT-B，所有方法的性能都非常接近。对于 ViT-L， 方法之间的差距更大，这表明更大模型的挑战是减少过度拟合。 我们的 MAE 可以轻松扩展，并且从更大的模型中显示出稳定的改进。我们获得了 86.9% 的准确率 使用 ViT-H（224 尺寸）。通过微调 448 大小，我们 达到 87.8% 的准确率，仅使用 IN1K 数据。以前最好的精度，在所有只使用 IN1K 的方法中 数据，是 87.1%（512 大小）[67]，基于先进的网络。我们以不平凡的幅度改进了最先进的技术 在极具竞争力的 IN1K 基准测试中（无需外部 数据）。我们的结果基于 vanilla ViT，我们期望 先进的网络会表现得更好。 与 BEiT [2] 相比，我们的 MAE 更准确 同时更简单，更快。我们的方法重构 像素，与预测令牌的 BEiT 相比：BEiT 在重建像素时报告了 1.8% 的退化 [2] 使用 ViT-B.2 我们不需要 dVAE 预训练。此外，我们的 MAE 比每个 epoch 快得多（每 epoch 3.5 倍） BEiT，原因如表 1c 所示。表 3 中的 MAE 模型针对 1600 进行了预训练 epochs 以获得更好的准确性（图 7）。即便如此，我们的总 训练时预训练时间少于其他方法 在相同的硬件上。例如，在 128 上训练 ViT-L TPU-v3 核心，我们的 MAE 的训练时间是 1600 的 31 小时 epochs 和 MoCo v3 是 36 小时，300 个 epochs [9]。



method	pre-train data	ViT-B	ViT-L	ViT-H	ViT-H$_{448}$
scratch, our impl.	-	82.3	82.6	83.1	-
DINO [5]	IN1K	82.8	-	-	-
MoCo v3 [9]	IN1K	83.2	84.1	-	-
BEiT [2]	IN1K+DALLE	83.2	85.2	-	-
MAE	IN1K	83.6	85.9	86.9	**87.8**

Table 3. **Comparisons with previous results on ImageNet-1K**. The pre-training data is the ImageNet-1K training set (except the tokenizer in BEiT was pre-trained on 250M DALLE data [50]). All self-supervised methods are evaluated by end-to-end fine-tuning. The ViT models are B/16, L/16, H/14 [16]. The best for each column is underlined. All results are on an image size of 224, except for ViT-H with an extra result on 448. Here our MAE reconstructs normalized pixels and is pre-trained for 1600 epochs.

Comparisons with supervised pre-training.

In the original ViT paper [16], ViT-L degrades when trained in IN1K.Our implementation of supervised training (see A.2) works
better, but accuracy saturates. See Figure 8.Our MAE pre-training, using only IN1K, can generalize better: the gain over training from scratch is bigger forhigher-capacity models. It follows a trend similar to the JFT-300M supervised pre-training in [16]. This comparison shows that our MAE can help scale up model sizes.

练，仅使用 IN1K，可以更好地泛化：从头开始训练的增益对于 更高容量的型号。它遵循类似的趋势 [16] 中的 JFT-300M 监督预训练。这种比较表明，我们的 MAE 可以帮助扩大模型大小。

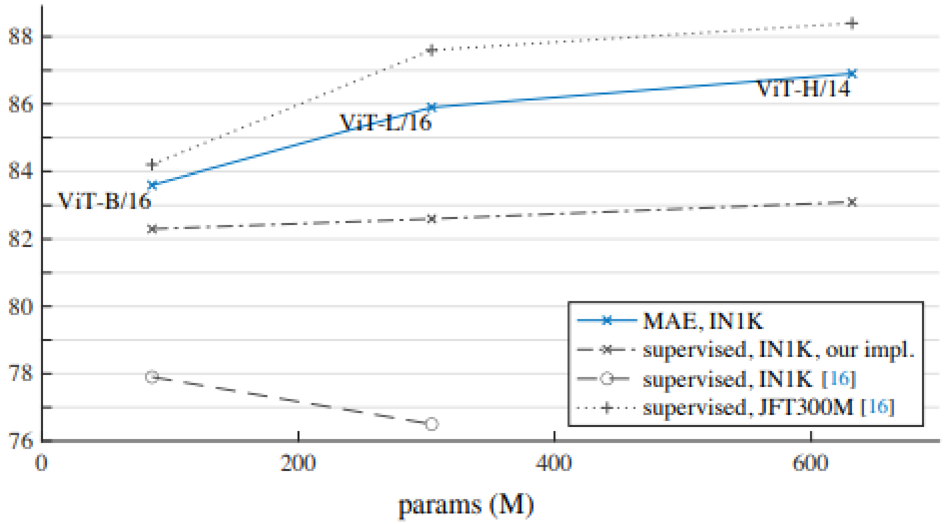


Figure 8. **MAE pre-training *vs.* supervised pre-training**, evaluated by fine-tuning in ImageNet-1K (224 size). We compare with the original ViT results [16] trained in IN1K or JFT300M.

4.3. Partial Fine-tuning

Table 1 shows that linear probing and fine-tuning results are largely uncorrelated. Linear probing has been a popular protocol in the past few years; however, it misses the opportunity of pursuing strong but non-linear features—which is indeed a strength of deep learning. As a middle ground, we study a partial fine-tuning protocol: fine-tune the last several layers while freezing the others. This protocol was also used in early works, e.g., [65, 70, 42]. Figure 9 shows the results. Notably, fine-tuning only one.Transformer block boosts the accuracy significantly from 73.5% to 81.0%. Moreover, if we fine-tune only "half" of the last block (i.e., its MLP sub-block), we can get 79.1%, much better than linear probing. This variant is essentially fine-tuning an MLP head. Fine-tuning a few blocks (e.g., 4 or 6) can achieve accuracy close to full fine-tuning. In Figure 9 we also compare with MoCo v3 [9], a contrastive method with ViT-L results available. MoCo v3 has higher linear probing accuracy; however, all of its partial finetuning results are worse than MAE. The gap is 2.6% when tuning 4 blocks. While the MAE representations are less linearly separable, they are stronger non-linear features and perform well when a non-linear head is tuned.These observations suggest that linear separability is notthe sole metric for evaluating representation quality. It hasalso been observed (e.g., [8]) that linear probing is not wellcorrelated with transfer learning performance, e.g., for object detection. To our knowledge, linear evaluation is not often used in NLP for benchmarking pre-training.

4.3.部分微调 表 1 显示了线性探测和微调结果 很大程度上是不相关的。**线性探测**一直很流行的protocol 过去几年的；然而，它错过了追求强大但非线性特征的机会——这是 确实是深度学习的强项。作为中间地带，**我们 研究一个部分微调协议：微调最后几层，同时冻结其他层。**该协议也是 用于早期作品，例如 [65, 70, 42]。 图 9 显示了结果。值得注意的是，**微调只有一个 变压器块显着提高了精度 73.5% 至 81.0%。此外，如果我们只微调"一半" 最后一个区块（即它的 MLP 子区块），我们可以得到 79.1%，比线性探测好得多。这种变体本质上是 微调 MLP 头。微调几个块（例如，4 或 6) 可以达到接近完全微调的精度。** 在图 9 中，我们还与 MoCo v3 [9] 进行了比较，这是一种具有 ViT-L 结果的对比方法。 MoCo v3有 更高的线性探测精度；然而，它的所有部分 微调结果比 MAE 差。差距为 2.6% 调整 4 块时。虽然 MAE 表示是 线性可分性较差，它们是更强的非线性特征 并且在调谐非线性磁头时表现良好.这些观察表明线性可分性不是 评估表示质量的唯一指标。 它有 还观察到（例如，[8]）线性探测效果不佳与迁移学习性能相关，例如，用于对象检测。 据我们所知，线性评估不是 经常在 NLP 中用于对预训练 进行基准测试。

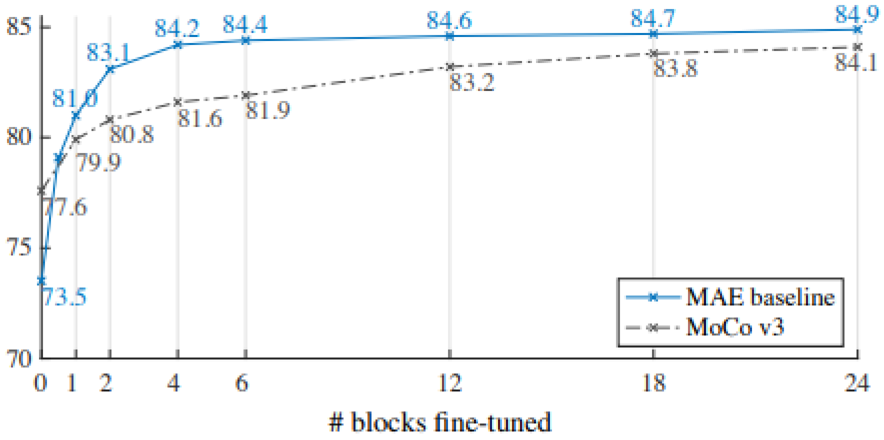


Figure 9. **Partial fine-tuning** results of ViT-L w.r.t. the number of fine-tuned Transformer blocks under the default settings from Table 1. Tuning 0 blocks is linear probing; 24 is full fine-tuning. Our MAE representations are less linearly separable, but are consistently better than MoCo v3 if one or more blocks are tuned.

6.Transfer Learning Experiments

We evaluate transfer learning in downstream tasks using the pre-trained models in Table 3.

Object detection and segmentation.

We fine-tune Mask R-CNN [24] end-to-end on COCO [37]. The ViT backbone is adapted for use with FPN [36] (see A.3). We apply this approach for all entries in Table 4. We report box AP for object detection and mask AP for instance segmentation. Compared to supervised pre-training, our MAE performs better under all configurations (Table 4). With the smaller ViT-B, our MAE is 2.4 points higher than supervised pre-training (50.3 vs. 47.9, APbox). More significantly, with the larger ViT-L, our MAE pre-training outperforms supervised pre-training by 4.0 points (53.3 vs. 49.3). The pixel-based MAE is better than or on par with the token-based BEiT, while MAE is much simpler and faster. Both MAE and BEiT are better than MoCo v3 and MoCo v3 is on par with supervised pre-training.

Semantic segmentation.

We experiment on ADE20K [72] using UperNet [63] (see A.4). Table 5 shows that our pre-training significantly improves results over supervised pre-training, e.g., by 3.7 points for ViT-L. Our pixel-based MAE also outperforms the token-based BEiT. These observations are consistent with those in COCO.

Classification tasks.

Table 6 studies transfer learning on the iNaturalists [56] and Places [71] tasks (see A.5). On iNat, our method shows strong scaling behavior: accuracy improves considerably with bigger models. Our results surpass the previous best results by large margins. On Places, our MAE outperforms the previous best results [19, 40], which were obtained via pre-training on billions of images.

Pixels vs. tokens.

Table 7 compares pixels vs. tokens as the MAE reconstruction target. While using dVAE tokens is better than using unnormalized pixels, it is statistically similar to using normalized pixels across all cases we tested. It again shows that tokenization is not necessary for our MAE.

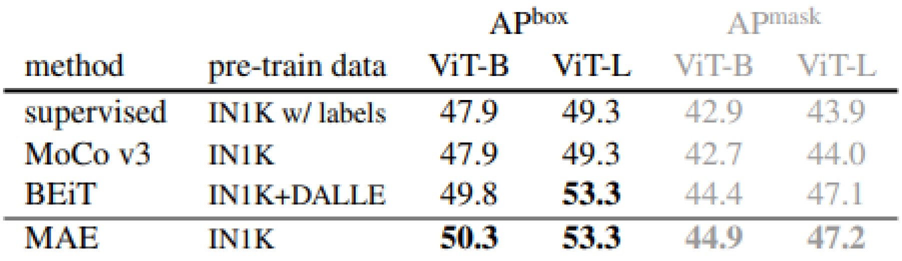
5.迁移学习实验 我们使用以下方法评估下游任务中的迁移学习 表 3 中的预训练模型。

对象检测和分割。我们微调 Mask R-CNN [24] 在 COCO [37] 上端到端。 ViT 骨干 适用于 FPN [36]（见 A.3）。我们应用这个 表 4 中所有条目的方法。我们报告框 AP 为 对象检测和掩码 AP 用于实例分割。 与有监督的预训练相比，我们的 MAE 执行 在所有配置下都更好（表 4）。随着较小 ViT-B，我们的 MAE 比监督预训练高 2.4 点（50.3 对 47.9，APbox）。更重要的是，随着 更大的 ViT-L，我们的 MAE 预训练优于有监督的 预训练 4.0 分（53.3 对 49.3）。 基于像素的 MAE 优于或与 基于令牌的 BEiT，而 MAE 更简单、更快。 MAE 和 BEiT 都优于 MoCo v3 和 MoCo v3 与有监督的预训练相当。

语义分割。我们在 ADE20K [72] 上进行实验 使用 UperNet [63]（见 A.4）。表 5 显示，我们的预训练显着提高了监督预训练的结果，例如，ViT-L 提高了 3.7 分。我们基于像素的 MAE 也优于基于代币的 BEiT。这些观察 与COCO中的一致。

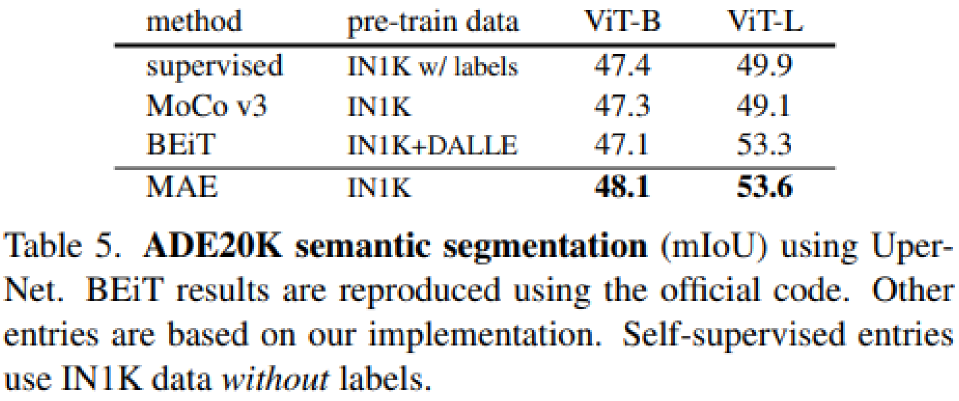
分类任务。表 6 研究迁移学习 iNaturalists [56] 和 Places [71] 任务（见 A.5）。在 iNat，我们的方法显示出强大的缩放行为：准确度 更大的模型显着改善。我们的结果大大超过了以前的最佳结果。在地方，我们的 MAE 优于之前的最佳结果 [19, 40]， 这是通过对数十亿张图像进行预训练获得的。

像素与令牌。表 7 比较了像素与标记作为 MAE 重建目标。使用 dVAE 令牌时 比使用非归一化像素更好，它在统计上类似于在我们测试的所有情况下使用归一化像素。它 再次表明我们的 MAE 不需要标记化。



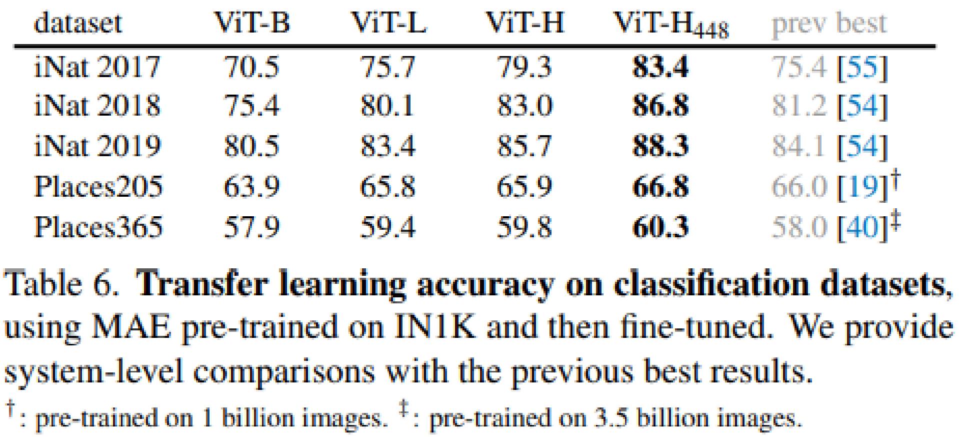
		AP^{box}		AP^{mask}	
method	pre-train data	ViT-B	ViT-L	ViT-B	ViT-L
supervised	IN1K w/ labels	47.9	49.3	42.9	43.9
MoCo v3	IN1K	47.9	49.3	42.7	44.0
BEiT	IN1K+DALLE	49.8	**53.3**	44.4	47.1
MAE	IN1K	**50.3**	**53.3**	**44.9**	**47.2**

Table 4. **COCO object detection and segmentation** using a ViT Mask R-CNN baseline. All entries are based on our implementation. Self-supervised entries use IN1K data *without* labels. Mask AP follows a similar trend as box AP.



method	pre-train data	ViT-B	ViT-L
supervised	IN1K w/ labels	47.4	49.9
MoCo v3	IN1K	47.3	49.1
BEiT	IN1K+DALLE	47.1	53.3
MAE	IN1K	**48.1**	**53.6**

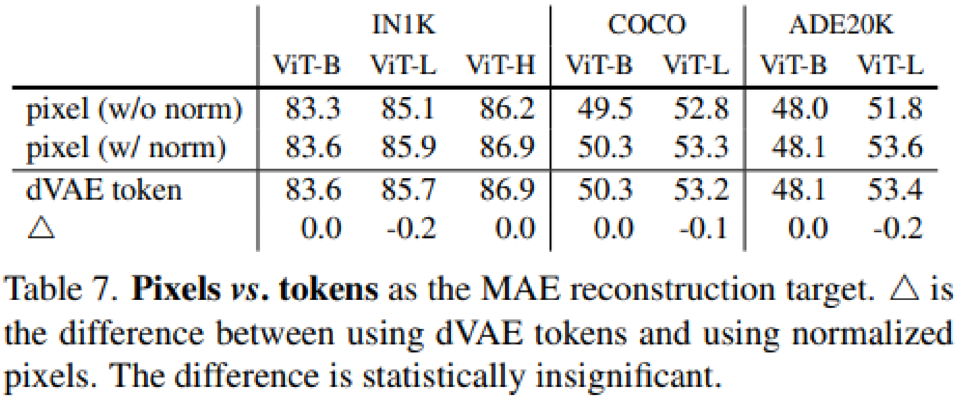
Table 5. **ADE20K semantic segmentation** (mIoU) using UperNet. BEiT results are reproduced using the official code. Other entries are based on our implementation. Self-supervised entries use IN1K data *without* labels.



dataset	ViT-B	ViT-L	ViT-H	ViT-H_{448}	prev best
iNat 2017	70.5	75.7	79.3	**83.4**	75.4 [55]
iNat 2018	75.4	80.1	83.0	**86.8**	81.2 [54]
iNat 2019	80.5	83.4	85.7	**88.3**	84.1 [54]
Places205	63.9	65.8	65.9	**66.8**	66.0 [19]†
Places365	57.9	59.4	59.8	**60.3**	58.0 [40]‡

Table 6. **Transfer learning accuracy on classification datasets**, using MAE pre-trained on IN1K and then fine-tuned. We provide system-level comparisons with the previous best results.

†: pre-trained on 1 billion images. ‡: pre-trained on 3.5 billion images.



	IN1K			COCO		ADE20K	
	ViT-B	ViT-L	ViT-H	ViT-B	ViT-L	ViT-B	ViT-L
pixel (w/o norm)	83.3	85.1	86.2	49.5	52.8	48.0	51.8
pixel (w/ norm)	83.6	85.9	86.9	50.3	53.3	48.1	53.6
dVAE token	83.6	85.7	86.9	50.3	53.2	48.1	53.4
△	0.0	-0.2	0.0	0.0	-0.1	0.0	-0.2

Table 7. **Pixels *vs*. tokens** as the MAE reconstruction target. △ is the difference between using dVAE tokens and using normalized pixels. The difference is statistically insignificant.

7.Discussion and Conclusion

Simple algorithms that scale well are the core of deep learning. In NLP, simple self-supervised learning methods (e.g., [47, 14, 48, 4]) enable benefits from exponentially scaling models. In computer vision, practical pre-training paradigms are dominantly supervised (e.g. [33, 51, 25, 16]) despite progress in self-supervised learning. In this study, we observe on ImageNet and in transfer learning that an autoencoder—a simple self-supervised method similar to techniques in NLP—provides scalable benefits. Self-supervised learning in vision may now be embarking on a similar trajectory as in NLP. On the other hand, we note that images and languages are signals of a different nature and this difference mustbe addressed carefully. Images are merely recorded light without a semantic decomposition into the visual analogue of words. Instead of attempting to remove objects, we remove random patches that most likely do not form a semantic segment. Likewise, our MAE reconstructs pixels, which are not semantic entities. Nevertheless, we observe (e.g., Figure 4) that our MAE infers complex, holistic reconstructions, suggesting it has learned numerous visual concepts, i.e., semantics. We hypothesize that this behavior occurs by way of a rich hidden representation inside the MAE. We hope this perspective will inspire future work.

Broader impacts.

The proposed method predicts content based on learned statistics of the training dataset and as such will reflect biases in those data, including ones with negative societal impacts. The model may generate inexistent content. These issues warrant further research and consideration when building upon this work to generate images.

讨论与结论 可扩展性好的简单算法是深度算法的核心 学习。在 NLP 中，简单的自我监督学习方法 （例如，[47, 14, 48, 4]）能够以指数方式受益 缩放模型。在计算机视觉中，实用的预训练 范式受到主要监督（例如 [33, 51, 25, 16]） 尽管在自我监督学习方面取得了进展。在这项研究中， 我们在 ImageNet 和迁移学习中观察到 自动编码器——一种简单的自我监督方法，类似 到 NLP 中的技术——提供可扩展的好处。视觉中的自我监督学习现在可能正在开始 与 NLP 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，这种差异必须 认真对待。图像只是记录的光 没有语义分解成视觉类比 的话。我们没有尝试删除对象，而是删除了最有可能不形成语义段的随机补丁。同样，我们的 MAE 重建像素，这不是语义实体。然而，我们观察到（例如， 图 4）我们的 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。我们假设这种行为发生 通过 MAE 内部丰富的隐藏表示。我们 希望这种观点能激发未来的工作。 更广泛的影响。所提出的方法预测内容 基于训练数据集的学习统计数据，因此将反映这些数据中的偏见，包括具有负面社会影响的数据。该模型可能会生成不存在的 内容。在此工作的基础上生成图像时，这些问题值得进一步研究和考虑。