# 屏蔽式自动编码器是可扩展的视觉学习器

何开明*, t 陈新雷* 谢赛宁 李阳浩 皮奥特-达拉尔 Ross Girshick

•平等的技术贡献 t 项目负责人

Facebook Al Research (FAIR)



## 摘要

*本文表明，遮蔽自动编码器（MAE）是计算机视觉中可扩展的se(f)监督学习器。我们的MAE方法很简单：我们对输入图像的随机斑块进行屏蔽，并重建缺失的像素。它基于两个核心设计。首先，我们开发了一个非对称性的编码器-解码器结构，其中编码器只对可见的斑块子集进行操作（不含掩码），同时还有一个轻量级的解码器来重建从潜在的表征和掩码中获取原始图像tokens。第二，我们发现，掩盖高比例的在输入图像中，例如75%，会产生一个不难的、有意义的监督任务。耦合这两个标志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具有良好泛化能力的高容量模型：例如，在只使用lmageNet-1K数据的方法中，一个虚构的ViT-Huge模型达到了最好的准确性（87.8%）。Tran.1:fer per formance在下游任务中的表现优于有监督的预训练，并显示出有希望的扩展行为。*

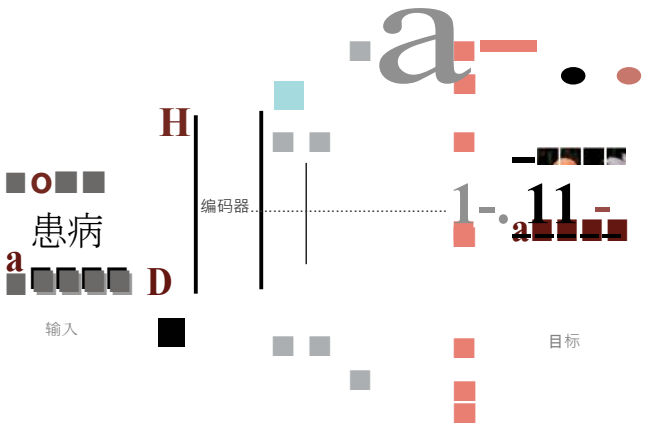



图1.我们的MAE架构。在预训练期间，一个大

## 1. 简介

深度学习见证了能力和容量持续增长的结构的爆炸f 33, 25, 571-在硬件快速增长的帮助下，今天的模型可以很容易地超过一百万张图像f131，并开始要求数以亿计的--通常是公众无法获得的--标记的图像f61。

在自然语言处理（NLP）中，这种对数据的渴求已经通过自监督的PRC训练成功解决。这些解决方案基于GPT f47, 48, 41中的自回归语言建模和BERT f141中的*屏蔽式自动编码*，概念上很简单：它们删除一部分数据并学习预测被删除的内容。这些方法现在能够训练包含一千多亿个参数的可推广的NLP模型f41-

屏蔽式自动编码系统的想法，是更普遍的DCnoising自动编码系统f581的一个组成部分，在计算机视觉中也是很自然和适用的。事实上，密切相关的研究

斑块和掩码标记被一个小型解码器处理，该解码器以像素重建原始图像。在预训练之后，解码器被丢弃，编码器被应用于未被破坏的图像（全套斑块）的识别任务。

视觉中的自动编码方法比BERT更早。然而，尽管随着BERT的成功，人们对这一想法产生了极大的兴趣，但视觉中的自动编码方法的进展却落后于NLP。我们问：*是什么让遮蔽式自动编码在视觉和语言之间变得不同？*我们试图从以下几个方面来回答这个问题。

(i) 直到最近，架构是不同的。在视觉方面，卷积网络f34l在过去十年中占主导地位f33l卷积通常在规则的网格上运行，将 "指标"如掩码标记f141或位置嵌入f57l集成到卷积网络中并不简单。然而，随着视觉变换器（ViT）的引入，这一架构上的差距已经得到了解决f161，应该不再是一个障碍。

(ii) 语言和视觉之间的信息密度是不同的。语言是人类产生的信号，具有高度的语义和信息密度。当训练一个模型来预测每个句子中仅有的几个缺失的单词时，这项任务似乎会诱发复杂的语言地位。相反，图像是具有严重空间冗余的自然信号--
例如，一个缺失的斑块可以由邻近的斑块重新覆盖，而这些斑块是由高层次的非------。

随机的图像补丁子集（*例如*，75%）被掩盖掉。编码器被应用于*可见斑块*的小子集。掩码标记*在*编码器*之后被*引入，整套en编码

图2.lmageNet*验证*图像的结果示例。对于每个三联体，我们显示了被遮蔽的图像（左），我们的MAE重构图（中），和地面实况（右）。屏蔽率为80%，196个斑块中只留下39个。更多的例子在附录中。 TA.1--

*没有损失计算出油的可见斑块，Nwdel输出的油的可见斑块在质量上更差。我们可以简单地将输出与可见斑块叠加，以提高视觉质量。我们故意选择这样做，这样我们可以更全面地展示mellwd的行为。*

图3.使用在lmageNet上训练的MAE（与图2中的模型权重相同）在COCO验证图像上的结果示例。观察最右边的两个例子的重建，虽然与地面真相不同，但在语义上是可信的。

对零件、物体和场景的理解。为了克服这种差异并鼓励学习有用的特征，我们表明一个简单的策略在计算机视觉中很有效：掩盖*非常高的*随机斑块的部分。这个策略在很大程度上减少了冗余，并创造了一个具有挑战性的自我监督任务，需要超越低层次图像统计的整体理解。为了对我们的重建任务有一个定性的认识，请看图2-4。

(iii) autocncodcr的*解码器*，将潜在的表征映射回输入，在重建文本和图像之间扮演着不同的角色。在视觉中，解码器重建*像素*，因此其输出的语义水平比普通的识别任务要低。这与语言不同，在语言中，解码器预测的是包含丰富语义信息的缺失*词*。虽然在BERT中，解码器可以是微不足道的（MLP）[141]，但我们发现，对于图像来说，解码器的设计在决定所学的潜在表征的语义水平方面起着关键作用。

在这一分析的推动下，我们提出了一个简单、有效、可扩展的视觉表征学习的遮蔽自动编码（MAE）的形式。我们的MAE对输入图像的随机斑块进行屏蔽，并在像素空间中重建缺失的斑块。它有一个*不对称的*编码器和解码器设计。我们的编码器只对可见的斑块子集进行操作（没有掩码标记），而我们的解码器是轻量级的，并从潜在的rep
resentation与掩码标记一起重建输入（图I）。在我们的非对称编码器-
解码器中，将掩码标记转移到小的解码器，导致计算量大大减少。在这种设计下，一个非常高的屏蔽率（*如*75%）可以实现双赢的局面：它优化了准确性，同时降低了编码器只处理一小部分（*如*25%）的斑块。这可以使整个预训练时间减少3倍或更多，同样也可以减少内存消耗，使我们能够轻松地将我们的MAE扩展到大型模型。

我们的MAE可以学习到非常高容量的模型，并能很好地泛化。通过MAE的预训练，我们可以在ImageNct-
lK上训练像ViT-
Large/Hugc[161]这样的数据饥渴型模型，并提高泛化性能。用一个普通的ViT-Huge模型，在ImagcNct-
1K上进行微调时，我们达到了87.8%的准确性。这超过了以前所有只使用ImagcNct-
1K数据的结果。我们还评估了对象检测、实例分割和语义扫描的迁移学习。在这些任务中，我们的预训练取得了比它的监督预训练同行更好的结果，更重要的是，我们观察到通过扩大模型的规模获得了明显的收益。这些观察结果是一致的

与那些在NLP自我监督的预培训中所见证的
我们希望它们能使我们的领域在未来的日子里有
更多的机会。
探索一个类似的轨迹。

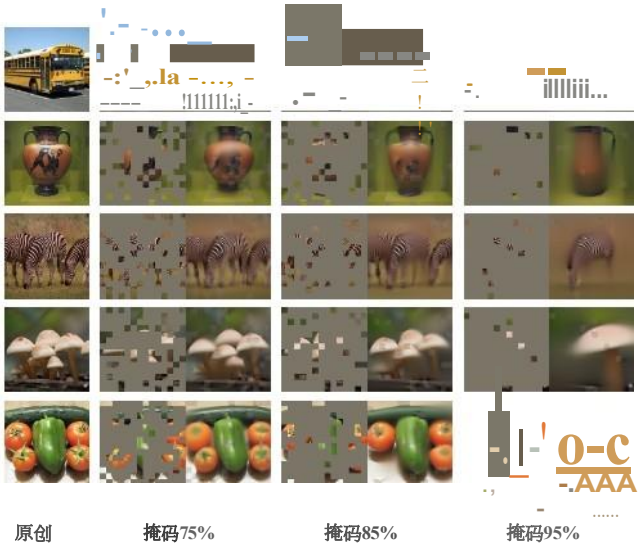



图4.使用预先训练好的掩蔽率为75%的MAE对ImageNet*验证*图像进行重建，但应用于掩蔽率更高的输入。预测结果与原始图像有合理的差异，表明该方法可以通用。

## 2. 相关工作

**掩蔽语言建模**及其自回归模型<；例如，BERT f141和GPT f47，48，41，是NLP中非常成功的预训练方法。这些方法保留了输入序列的一部分并训练模型
来预测缺失的内容。这些方法已被证明具有很好的扩展性，大量的证据表明，这些预先训练好的表征能使我们的生活更加丰富多彩。
对各种下游任务有很好的适应性。

**自动编码**是一种经典的学习代表的方法。它有一个将输入映射到潜在代表的编码器和一个重建输入的解码器。例如，PCA和K-mcans是自动编码，29l
Dcnoising自动编码（DAE）f581是一类自动编码，它破坏输入信号并学习重建原始的、未被破坏的信号。一系列的方法可以被认为是在不同的损坏情况下的广义DAE，*例如，*掩盖像素f59、46、61或去除颜色通道[701。我们的MAE是去噪自动编码的一种形式，但在许多方面与经典的DAE不同。

**屏蔽图像编码**方法从被屏蔽破坏的图像中学习表征。f591的开创性工作将遮蔽作为DAE的一种噪声类型。Context Encoder f461使用卷积网络对大的缺失区域进行涂抹。在NLP成功的激励下，最近相关的方法[6，16，21弧形基于变形器[571.
iGPT
f61对像素的序列进行操作，并预测未被发现的像素。已知的像素。ViT的论文f161研究了用于自我监督学习的遮蔽补丁预测。最近，BEiT
f21提出预测离散的标记f44, 501-

**自监督学习**方法在计算机视觉中受到了极大的关注，通常集中在不同的预训练任务上[15，61，42，70，45，171。最近，对比性学习f3，221很受欢迎，*例如*f62，43，23，
71，它为两个或多个视图之间的图像相似性和不相似性（或只有相似性f21，
81）建模。对比法和相关方法强烈地依赖于数据的增强[7，21，
81.自动编码追求的是一个完全不同的方向，它表现出不同的行为方式，我们将介绍。

## 3. 办法

我们的掩码自动编码(MAE)是一种简单的自动编码方法，可以根据部分观察结果重建原始信号。像所有的自动编码方法一样，我们的方法有一个编码器，将观察到的信号映射到一个潜在的表示，还有一个解码器，从潜在的表示中重建原始的信号。与经典的自动编码系统不同，我们采用了一种*非对称的*设计，允许编码器只对部分观察到的信号进行操作（没有掩码标记），而一个轻量级的解码器可以重新构建<；从潜像表示和掩码标记中获取完整的信号。图1说明了这个想法，接下来介绍。

**遮蔽。**按照ViT f 161，wc将一幅图像分成若干个不重叠的斑块。然后，我们对补丁的一个子集进行抽样，并对剩余的补丁进行遮蔽*（即，*移除）。我们的取样策略很简单：我们按照统一的分布，在没有替换的情况下随机取样斑块。我们简单地将其称为 "随机抽样"。

具有*高*遮蔽率的随机抽样 *（即*被移除的斑块的比例）在很大程度上消除了冗余，从而创造了一个不容易通过从可见的相邻斑块推断来解决的任务（见图2-4）。均匀分布防止了潜在的中心偏向 *（即*在图像中心附近有更多的遮蔽斑块）。最后，高度稀疏的输入为设计一个高效的编码器创造了机会，接下来介绍。

**MAE编码器。**我们的编码器是一个ViT fl61，但只适用于*可见的*、*未被掩盖的斑块*。就像在标准的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，然后通过一系列的Transformer块来处理所得到的集合。然而，我们的编码器只对全集的一个小子集*（如*25%）进行操作。屏蔽的补丁会被移除；不使用屏蔽标记。这使我们可以用少量的计算和内存来训练非常大的编码器。全集由一个轻量级的解码器来处理，接下来介绍。

MAE解码器。MAE解码器的输入是由(i)编码的可见斑块组成的全套标记，以及(二)掩码令牌。Sec图1。每个掩码标记f141是一个共享的、学习过的向量，表示存在一个错误的------。

要预测的补丁。我们为这个完整的集合中的所有标记添加位置嵌入；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。我们将对这一完整集合中的所有标记添加位置嵌入；如果没有这一点，面具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变压器块。

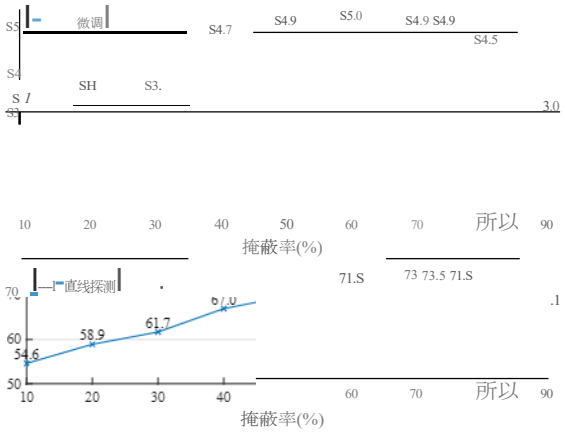
MAE解码器只在预训练期间用于执行图像重建ta,;;k（只有编码器用于产生用于识别的图像表示）。因此，解码器的结构可以以*独立*于编码器设计的方式灵活地设计。我们尝试使用非常小的解码器，比编码器更窄、更低。例如，我们的默认解码器*与*编码器*相比，*每个令牌的计算量<
I0%。通过这种不对称的设计，全套的标记只由轻量级的解码器负责，这大大减少了预训练时间。

**重构目标。**我们的MAE通过预测每个被遮蔽的补丁的*像素*值来重建输入。解码器输出中的每一个片段都是代表一个补丁的像素值的向量。解码器的最后一层是一个林耳投影，其输出通道的数量等于一个补丁中的像素值的数量。解码器的输出被重塑以形成一个重建的图像。我们的损失函数是计算重建图像和原始图像在像素空间中的平均平方误差（MSE）。我们只在被遮蔽的斑块上计算损失，与BERT rl4l相似。[1]

我们还研究了一个变体，其重建目标是每个被遮蔽的补丁的标准化像素值。具体来说，我们计算一个补丁中所有像素的平均值和标准差，并使用它们来规范这个补丁。在我们的实验中，使用归一化的像素作为重建目标可以提高表示质量。

**实施简单。**我们的MAE预训练可以有效地实施，而且重要的是，不需要任何专门的稀疏操作。首先，我们为每一个输入补丁生成一个标记（通过线性投影和一个附加的po
sitional嵌入）。接下来，我们根据掩蔽率，*随机洗刷*标记列表并*删除*列表中的最后一部分。这个过程为编码器产生了一个小的标记子集，相当于无替换地对补丁进行采样。编码后，我们在编码补丁的列表上附加一个掩码标记的列表，并*解开*这个完整的列表（倒置随机洗牌操作），使所有标记与它们的目标对齐。解码器被应用于这个完整的列表（加入位置嵌入）。如前所述，不需要任何稀疏操作。这个简单的实现引入了可以忽略不计的开销，因为洗牌和解除洗牌的操作是快速的。

[1]只在被遮蔽的斑块上计算损失，这与传统的去噪自动编码器[58]不同，后者在所有像素上计算损失。这种选择纯粹是由结果驱动的：在所有像素上计算损失会导致精度的轻微下降（*例如*，约0.5%）。




图S.

**屏蔽率。**高遮蔽率（75%）对微调（顶部）和线性探测（底部）都很有效。在本文的所有图中，它们的轴是lmageNet-1K的验证精度（%）。

## 4. 图像网实验

我们在lmagcNet-1K（INIK）r131训练集上进行自我监督的预训练。然后我们进行监督训练，用（i）端到端微调或（ii）线性探测来评估表征。我们报告了单个224x224作物的top-1验证精度。细节弧度见附录A.I。

基线。ViT-Large。我们在消融研究中使用ViT-Largc（ViT-L/16）r161作为骨干。ViT-L非常大（比RcsNct-50 r2sl大一个数量级），倾向于Ovcrfit。下面是ViT-L从头开始训练*与*从我们的基线MAE微调的比较。

| 划痕，原创 [16] | 划痕，我们的imp！。 | 基线MAE |
|---|---|---|
| 76.5 | 82.5 | 84.9 |

我们注意到，从头开始训练*有监督的*ViT-L是不容易的，需要一个有强大正则化的好配方（82.5%，见附录A.2）。即便如此，我们的MAE预训练还是贡献了很大的改进。这里的微调只针对50个epochs（*相对于*从头开始的200个），这意味着微调的准确性在很大程度上取决于预训练。

### 4.1. 主要属性

我们使用表一中的默认设置来消减我们的MAE（sec caption）。观察到了几个耐人寻味的特性。

**屏蔽率。**图5显示了遮蔽率的影响。最佳比率出乎意料地高。75%的比率对线性探测和微调都很好。这种行为与BERT r141相悖，;;t，其典型的屏蔽率为15%。我们的遮蔽率也远远高于计算机视觉领域的相关工作r6，16，21（20%到50%）。

该模型*推断出*缺失的斑块，以产生不同的、但合理

的输出（图4）。它使物体和场景的格式塔有了意义，这不能简单地通过扩展线条或纹理来完成。我们假设，这种类似推理的行为与我们的表征学习有关。

图5还显示，线性探测和微调rcsulL遵循*不同的*趋势。对于线性探测，ac-

| 大厦 | 呎吋 | 金 |
|---|---|---|
| I | 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) 解码器深度。深度解码器可以证明线性探测的准确性。

| 暗淡 | 呎吋 | 金 |
|---|---|---|
| 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) 解码器的宽度。解码器可以比编码器更窄（I024-d）。

| 案例 | 燃烧 | 金 | 荧光剂 (FLOPs) |
|---|---|---|---|
| 编码器，带[Ml | 84.2 | 59.6 | 3.3x |
| 编码器，无[Ml | 84.9 | 73.S | 1 x |

(c) **掩码令牌。**没有屏蔽令牌的编码器更准确、更快速（表2）。

| 案例 | 呎吋 | 金 |
|---|---|---|
| 像素(无规范) | 84.9 | 73.5 |
| 像素（含规范） | 85.4 | 73.9 |
| PCA | 84.6 | 72.3 |
| dVAE令牌 | 85.3 | 71.6 |

(d) 重建目标。像素一个重建目标是有效的。

| 案例 | rt | 金 |
|---|---|---|
| 没有 | 84.0 | 65.7 |
| 作物，固定尺寸 | 84.7 | 73.1 |
| 作物，兰特大小 | 84.9 | 73.S |
| 作物+颜色jit | 84.3 | 71.9 |

(e) 数据增强。我们的MAE在最小或没有增强的情况下工作。

| 案例 | 比例 | 呎吋 | 金 |
|---|---|---|---|
| 随机 | 75 | **84.9** | 73.5 |
| 块 | 50 | 83.9 | 72.3 |
| 块 | 75 | 82.8 | 63.9 |
| 网格 | 75 | 84.0 | 66.0 |

((掩码采样。随机取样的效果最好。视觉效果见图6。

表 一 .用ViT-U16在ImageNet-1K上进行的MAE消融实验。我们报告了微调 （ fl ）和线性探测（tin）的准确性（%）。如果没有指定，默认情况是：解码器深度为8，宽度为512，重建目标为非标准化像素，数据增强为随机调整大小的裁剪，遮蔽率为75%，预训练长度为800epochs。默认设置以灰色标注。

准确度随着遮蔽率的增加而稳步上升，直到甜蜜点：准确度差距高达约20%（54.6%与73.5%）。对于微调，结果对比率不那么敏感，广泛的掩蔽比率（40-80%）效果很好。图5中所有的微调结果都比从头开始训练的结果好（82.5%）。

**解码器设计。**我们的MAE解码器可以灵活地进行脱签，如表Ia和Ib中研究的那样。

表一a改变了解码器的深度（跨前块的数量）。一个足够深的解码器对线性探测很重要。这可以用像素重建任务和识别任务之间的差距来解释：自动编码中的最后几层对重建来说更加专业化，但对识别来说却不那么重要。一个合理的深度解码器可以考虑到重建的专业性，把潜在的表征留在一个更抽象的水平上。这种设计可以在线性探测中产生高达8%的改进（表la，'lin'）。然而，如果使用微调，编码器的最后几层可以被调整以适应识别任务。解码器的深度对改善微调的影响较小（表一a，'ft'）。

有趣的是，我们的*单块*解码器MAE在微调的情况下可以有很强的表现（84.8%）。请注意，一个单一的Transformer块是将信息从可见标记推广到屏蔽标记的最低要求。这样一个小型解码器可以进一步加快训练速度。

在表一中，我们研究了解码器的宽度（通道数）。我们默认使用512-d，在不进行微调和线性探测的情况下表现良好。更窄的解码器在微调时也表现良好。

总的来说，我们的默认MAE解码器是轻量级的。 它有8个块，宽度为512-d（表一中的灰色）。 *与*ViT-L（24个块，1024-d）*相比，* 它每个令牌只有9%的FLOPs。因此，虽然解码器处理所有的标记，但它仍然是整个计算的一小部分。

| 编码器 | 深度 | 呎吋王牌 | 小时 | 加速 |
|---|---|---|---|---|
| ViT-L, w/ [Ml | 8 | 84.2 | 42.4 | |
| 薇婷-L | 8 | 84.9 | 15.4 | 2.8x |
| 薇婷-L | I | 84.8 | ) '1.6 | 3.7x |
| 薇婷-H, w/ [Ml | 8 | | 119.M | |
| 薇塔-H | 8 | 85.8 | 34.5 | 3.5x |
| 薇塔-H | I | 85.9 | 29.3 | 4.lx |

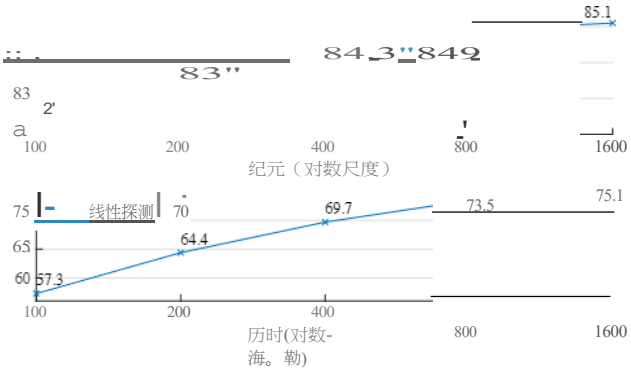
表2.我们的MAE训练（800个历时）的挂钟时间，在128个TPU-v3核心中用TensorFlow进行了基准测试。加速是相对于编码器有掩码标记的条目而言的（灰色）。解码器宽度为512，掩码率为75%，t。这个emry是通过训练十个epochs来估计的。

**掩码标记。**我们的MAE的一个重要设计是在编码器中跳过掩码标记[Ml]，并在后面的轻量级解码器中应用它。表Ic研究了这种设计。

如果编码器*使用*掩码标记，它的表现*更糟：*在线性探测中，它的准确性下降了14%。在这种情况下，预训练和部署之间存在着差距：在预训练中，这个编码器的输入中有很大一部分掩码标记，而这些标记在未损坏的图像中并不存在。这种差距可能会降低部署时的准确性。通过从编码器中移除掩码标记，我们将编码器约束为总是秒杀*真实的*斑块，从而提高准确率。

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le中，我们将整个训练FLOPs减少了3.3倍，这导致我们的实现有2.8倍的壁时钟加速（secta ble 2）。对于较小的解码器（I块）、较大的编码器（ViT-H）或两者，壁时钟加速甚至更大（3.5-4.1倍）。请注意，在掩码率为75%的情况下，速度提升可以超过4倍，部分原因是自我注意的复杂性是二次的。此外，内存也大大减少，这可以使训练更大的模型或通过大批量训练来加快速度。时间和内存效率使我们的MAE有利于训练非常大的模型。


图6.掩膜采样策略决定了借口任务djfficulty，影响了重建质量和表示方法（表11）。这里的输出来自于用特定的掩蔽策略训练的MAE。左边：随机采样（我们的默认值）。中dle：块状取样[2]，去除大的随机块。右图：网格化取样，每四个斑块中保留一个。图片来自验证集。




211仅使用裁剪增强就会降低13%的准确率

**重构目标。**我们在表Id中比较了不同的重建目标。到目前为止，我们的结果是基于没有（每块）标准化的像素。使用*具有*归一化*的*像素可以提高精确度。这种按片归一化增强了局部的对比度。**在**另一个变体**中**，我们在补丁空间进行PCA，并使用最大的PCA系数。(这里是96)作为目标。这样做会降低准确度。这两个实验表明，高频成分在我们的方法中是有用的。

我们还比较了预测令牌的MAE变体，*即*BEiT f21中使用的目标，具体到trus变体，我们使用DALLE预训练的dVAE f501作为tokcnizcr，遵循f21。MAE解码器使用交叉熵损失预测骰子中的标记。这种标记化*与*未规范化的像素*相比*，微调精度提高了0.4%，但*与*nom1alizcd像素*相比*没有优势。它也降低了线耳探测的准确性。**在**第5节中，我们进一步表明，在迁移学习中，标记化是没有必要的。

我们*基于像素的*MAE比tokeniza tion简单得多。dVAE tokcnizcr需要多一个预训练阶段，这可能取决于额外的数据（250M图像f501）。dVAE编码器是一个大的卷积网络（ViT-L的40%FLOPs），增加了不小的开销。使用pix els docs不会有这些问题。

**数据增强。**表le说明了数据增强对我们的MAE预训练的影响。

我们的MAE在使用固定尺寸或随机尺寸（都有随机的水平翻转）的*仅裁剪*的增量时效果良好。添加颜色的jittcring会降低结果，因此我们在其他实验中不使用它。

令人惊讶的是，即使*不*使用*数据增强*（只有中心裁剪，没有翻转），我们的MAE也表现得很好。这一特性与对比学习和相关方法f62, 23, 7, 211有很大的不同，后者严重依赖数据增强。据观察，f

图7.**训练时间表。一个**较长的训练计划可以带来明显的改善。这里的每一点都是一个完整的训练计划。该模型为ViT-L，其默认设置见表I。

和28%，BYOL f2ll和SimCLR f7l此外，没有证据表明对比学习可以在没有增强的情况下工作：图像的两个视图弧度相同，很容易满足一个琐碎的解决方案。

**在MAE中**，数据增强的作用主要是通过随机掩码（接下来会说明）对n,cd进行PCR。每一次迭代的掩码都是不同的，因此它们产生新的训练样本，与数据增强无关。掩码使借口任务变得困难，需要较少的增强来规范化训练。

**掩膜采样策略。**在表I中，我们比较了不同的掩膜采样策略，如图6所示。

f21中提出的*块状*遮蔽策略倾向于去除大块（图6中）。在比例为50%的情况下，我们的MAE效果相当好，但在比例为75%的情况下，效果就会下降。这项任务比随机抽样更难，因为观察到的训练损失更高。实验的结果也比较模糊。

我们还研究了*网格化*取样，即定期保留每四个斑块中的一个（图6右）。这是一个更容易的任务，而且训练损失更低。表征的清晰度更高。然而，表示质量较低。

简单的随机抽样对我们的**MAE**来说效果最好。它允许更高的掩蔽率，在享受良好的准确性的同时，也提供了更大的提速效益。

**训练时间表。**到目前为止，我们的消融是基于800-poch的预训练。图7显示了训练计划长度的影响。训练时间越长，精确度越高。事实上，即使在1600个历时中，我们也没有观察到线性探测精度的饱和。这种行为与对比学习方法不同，*例如*MoCo v3 f91 saniratcs在300 epochs的ViT-L。请注意，**MAE**编码器在每个历时中只对25%的斑块进行编码，而在对比性学习中，编码器在每个历时中对200%（两茬）甚至更多（多茬）斑块进行编码。

| 方法 | 预训练数据 | ViT-8 | ViT-L | ViT-H | ViT-H.i,ig |
|---|---|---|---|---|---|
| scratch, our impl. | | 82.3 | 82.6 | 83.1 | |
| DINO [SJ | !"NIK | 82.8 | | | |
| MoCo v3 (9) | !"NIK | 83.2 | 84.1 | | |
| BEiT [2] | INIK+DALLE | 83.2 | 85.2 | | |
| 闵中 | !"NIK | 83.6 | 85.9 | 86.9 | 87.8 |

表3. 与以前在ImageNet上的结果的比较

**IK。**预训练数据是ImageNet-1K训练集（除了BEiT中的标记器是在250M DALLE数据上预训练的（50））。所有的自监督方法都是通过端到端的微调进行评估的。ViT模型为B/16，UI6，H/14（16）。每一列的最佳结果都有下划线。所有的结果都是在224的图像大小上得出的，除了ViT-H在448上有额外的结果。在这里，我们的MAE重建了正常化的像素，并预先训练了1600个epochs。

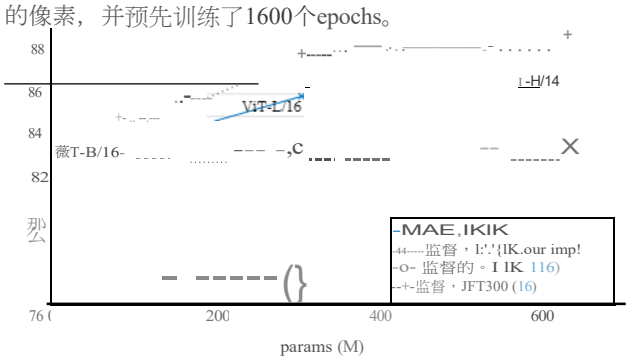



图8.**MAE预训练**与**监督预训练**，在ImageNet-**1K**（224大小）中通过微调进行评估。我们与在IN I Kor JFT300M中训练的原始ViT结果[16]进行比较。

## 4.2. 与以往结果的比较

**与自监督方法的比较。**在表3中，我们比较了自监督的ViT模型的微调结果。对于ViT-B，所有方法的表现都很接近。对于ViT-L，各方法之间的差距较大，这表明更大的模型面临的一个挑战是如何减少过度拟合。

我们的MAE可以很容易地扩大规模，并且从更大的模型中显示出稳定的改进。我们使用ViT-H（224大小）获得了86.9%的准确性。通过对448尺寸的微调，我们*仅使用IN1K数据就*获得了**87.8%的**准确率。在所有只使用INIK数据的方法中，以前的最佳准确率是87.1%（512大小）f671，基于高级网络。在竞争激烈的INIK基准（无外部数据）中，我们比最先进的方法有不小的改进。我们的结果是基于*vanilla* ViT的，我们期望先进的网络会表现得更好。

与BEiT f2l相比，我们的MAE*更准确*，同时也*更简单*和*快速*。我们的方法是重构像素，而BEiT则是预测标记。当用ViT-B重构像素时，BEiT的f21下降了1.8%。[2]我们不需要dVAE预训练。此外，我们的MAE比BEiT快得多（每历时3.5*倍*），原因见表Ic。

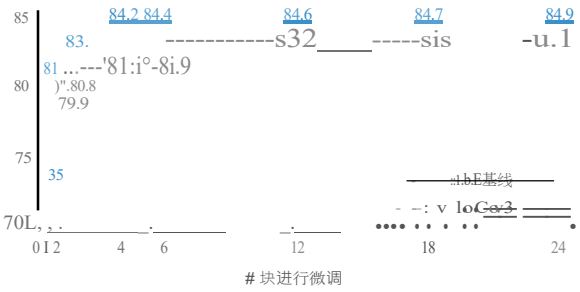



图9.在表I的默认设置下，ViT-L的**部分微调**结果与被微调的变压器块数有关。调整O块是线性探测；24块是完全微调。我们的MAE表征的线性可分性较差，但如果有一个或多个块被调谐，则始终比MoCo v3更可信。

[2]我们观察到使用ViT-L的BEiT也出现了退化：它产生了85.2%（代币）和83.5%（像素），这是从官方代码复制的。

表3中的MAE模型弧形预训练了1600个历时，以获得更好的准确性（图7）。即便如此，在相同的硬件上训练时，我们的总预训练时间也比其他方法*少*。例如，在128个TPU-v3cores上训练ViT-
L，我们的MAE的训练时间是31小时，1600个历时，MoCo v3的训练时间是36小时，300个历时f91

*与*有监督的预训练进行比较。在ViT的原始论文f]61中，ViT-
L在INIK中训练时出现退化。我们的有监督训练的实现（第A.2节）效果更好，但准确率达到了饱和。图8.

我们的MAE预训练，只使用INI
K，可以更好地概括：对于高容量的模型，比从头开始训练的收益更大。它的趋势与JFf-
300M的*监督*预训练相似，在f
161。这种比较表明，我们的MAE可以帮助扩大模型的规模。

### 4.3. 部分微调

表一显示，线性探测和微调的结果基本*不相关*。*在*过去的几年里，线性探测一直是一个流行的协议；然而，它错过了追求*强大但非线性*特征的机会--
这确实是深度学习的一个优势。作为一个中间地带，我们研究了一个*部分微调*协议：微调最后七层而冻结其他层。这个协议也被用于早期的工作中，*例如*f65，70,421。

图9显示了结果。值得注意的是，只微调了一个变换器块将准确性从73.5%大幅提升到81.0%。此外，如果我们只对最后一个区块（即其MLP子区块）的"一半"进行微调，我们可以凝胶79.1%，比线性探测法好得多。这个变体本质上是微调一个MLP头。微调几个块（*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
f91进行了比较，这是一种具有ViT-L结果的CON
Lrnslivc方法。MoCo
v3具有更高的线性探测精度；然而，其所有的部分微调结果都比MAE差。当调整4个块时，差距为2.6%。虽然MAE表征的线性分离度较低，但它们的*非线性*特征较强，在调谐非线性头时表现良好。

| | | ÃÃÃ | | J\pma.,k | |
|---|---|---|---|---|---|
| 方法 | 预训数据 | YiT-8 | YiT-L | YiT-8 | YiT-L |
| 被监督的 | 镍币，带标签 | 47.9 | 49.3 | 42.9 | 43.9 |
| MoCo v3 | !"NIK | 47.9 | 49.3 | 42.7 | 44.0 |
| 蓓儿丹娣 | INIK+DALLE | 49.8 | 53.3 | 44.4 | 47.1 |
| MAE | !"NIK | 50.3 | 53.3 | **44.9** | 47.2 |

| 方法 | 预训数据 | YiT-8 | YiT-L |
|---|---|---|---|
| 被监督的 | 镍钴合金，带标签 | 47.4 | 49.9 |
| MoCov3 | INIK | 47.3 | 49.1 |
| 蓓儿丹娣 | INIK+DALLE | 47.1 | 53.3 |
| MAE | !"NIK | 48.1 | 53.6 |

表4.使用ViT Mask R-CNN基线的COCO物体检测和分割。所有条目都是基于我们的实现。自我监督的条目使用[*没有*标签*的*NIK数据。掩膜AP的趋势与盒式AP相似。

这些观察结果表明，线性可分离性并不是评价表征质量的唯一指标。还有人观察到 *（例如，* [81]），线性探测与迁移学习的性能没有很好的相关性，例如，对于目标检测。据我们所知，在NLP中，线性评价并不经常被用来作为预训练的基准。

## 5. 转移学习实验

我们使用表3中的预训练模型在下游任务中评估转移学习。

物体检测和分割。我们对Mask R-CNN [241]在COCO [371-ViT主干线上进行了微调，以便与FPN [361]一起使用（A.3节）。我们对表4中的所有条目都采用这种方法。我们报告了用于物体检测的盒式AP和用于实例分割的面具AP。

*与*有监督的预训练相比，我们的MAE在所有配置下都表现得更好（表4）。对于较小的ViT-B，我们的MAE比有*监督的*预训练高2.4分（50.3 *对*47.9，AP""）。更重要的是，在较大的ViT-L中，我们的MAE预训练比监督预训练高出4.0分（53.3 *vs.* 49.3）。

*基于像素的*MAE优于或与*基于标记的*BEiT持平，而MAE则更简单、更快速。MAE和BEiT的弧度都优于MoCo v3，MoCo v3与监督预训练相当。

**语义分割。**我们使用UpcrNet[631]在ADE20K[721]上进行了实验（第A.4节）。表5显示，我们的预训练比*有监督的*PRC训练明显提高了结果，*例如，* ViT-L提高了3.7分。我们基于像素的MAE也优于基于令牌的BEiT。这些观察结果与COCO中的观察结果一致。

分类任务。表6研究了iNaturalists[561]和Places[711]任务中的转移学习（second A.5）。在iNat上，我们的方法显示出很强的扩展行为：随着模型的增大，准确率也大大提升。我们的结果*以很大的幅度*超过了以前的最佳结果。*在*Places上。我们的MAE超过了以前的最佳结果f19, 401。这些数据是通过对数十亿张图像的预训练获得的。

像素*与令*牌。表7比较了作为MAE重建目标的像素*和*标记。虽然使用dVAE标记比使用*未归一化的*像素要好，但在我们测试的所有情况下，它与使用*归一化的*像素在统计学上是相似的。这再次表明，对于我们的MAE来说，标记化是没有必要的。

表5.**ADE20K语义分割**（mloU）使用Uper Net。BEiT结果是使用官方代码再现的。其他条目是基于我们的实现。自我监督的条目使用*没有*标签*的*IN IK数据。

| 数据集 | YiT-8 | YiT-L | YiT-H | YiT-H44g | 最好的 |
|---|---|---|---|---|---|
| iNat 2017 | 70.5 | 75.7 | 79.3 | 83.4 | 75.4 (55] |
| iNat2018 | 75.4 | 80.1 | 83.0 | 86.8 | 81.2 [54] |
| iNat 2019 | 80.5 | 83.4 | 85.7 | 88.3 | 84.1 [54] |
| 场所205 | 63.9 | 65.8 | 65.9 | 66.8 | 66.0[19Jf |
| 场所365 | 57.9 | 59.4 | 59.8 | 60.3 | 58.0[40Jt |

表6.在分类数据集上的转移学习精度，使用MAE在INIK上进行预训练，然后进行微调。我们提供了与以往最佳结果的系统级比较。

t：在10亿张图像上进行了预训练。[1]：在35亿张图像上进行了预训练。

| | ÄÄÄ | | | 椰子 | | ADE20K | |
|---|---|---|---|---|---|---|---|
| | 薇塔-B | YiT-L | 薇T-11 | ViT-B | ViT-L | ViT-B | ViT-L |
| 像素(不含nonn) | 83.3 | 85.1 | 86.2 | 49.5 | 52.8 | 48.0 | 51.8 |
| 像素(w/n) | 83.6 | 85.9 | 86.9 | 50.3 | 53.3 | 48.1 | 53.6 |
| dYAE令牌 | 83.6 | 85.7 | 86.9 | 50.3 | 53.2 | 48.1 | 53.4 |
| 6. | 0.0 | -0.2 | 0.0 | 0.0 | -0.1 | 0.0 | -0.2 |

表7.作为MAE重建目标的像素*与*令牌。6.是使用dVAE标记和使用归一化像素之间的差异。差异在统计学上是不显著的。

## 6. 讨论和结论

简单的算法可以很好地扩展，是深度学习的核心。在NLP中，简单的自我监督学习方法*（例如，*[47, 14, 48, 41]）能够从指数级扩展的模型中获益。在计算机视觉中，尽管自监督学习取得了进展，但实际的预训练范式仍以监督为主导*（例如*[33, 51, 25, 161]）。在这项研究中，我们在ImageNct和转移学习中观察到，自动编码--
一种类似于NLP技术的简单自我监督方法--
提供了可扩展的好处。视觉中的自我监督学习现在可能正走在与NLP类似的轨道上。

另一方面，我们注意到，图像和语言*的性质不同*，这种差异必须认真对待。图像仅仅是记录下来的光，*没有*语义分解为文字的视觉类似物。我们不是试图去除objccL，而是重新移动那些最有可能*不*形成语义scgn1cnt的随机斑块。同样地，我们的MAE重建了像素，这些像素*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、整体的rcconslTuc
tions，表明它已经学会了许多视觉概念，*即*语义。我们假设，这种行为是通过MAE内部丰富的隐藏表征发生的。我们希望这个观点能对未来的工作有所启发。

**更广泛的影响。**建议的方法是基于训练数据集的学习统计，因此会反映出这些数据的偏差，包括具有负面社会影响的数据。该模型可能产生不存在的内容。这些问题值得进一步研究和考虑，在这项工作的基础上生成图像。

## 参考文献


llJ Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton.Layer normalization. *arXiv:1607.06450,* 2016.

l2J Hangbo Bao, Li Dong, and Furu Wei.BEiT: 图像变换器的BERT预训练. *arXiv:2106.08254,*2021.*Accessed in.Jun.e 2021.*

l3J Suzanna Becker and Geoffrey E Hinton.发现随机点状图中的表面的自组织神经网络。*Na ture,* 1992.

l4J Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neclakantan, Pranav Shyam, Girish Sastry, Amanda Askcll, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Hcnighan, Rcwon Child, Aditya Ramc.h. Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Matcusz Litwin, Scott Gray, Ben jamin Chess, Jack Clark, Christopher Bemer, Sam McCandlish, Alce Radford, llya Sutskcver, and Dario Amodei。语言模型弧形的少许学习者。In *NeurJPS,* 2020.

l5J Mathilde Caron, Hugo Touvron, lshan Misra, Herve Jcgou, Julien Maira！, Piotr Bojanowski, and Armand Joulin.自监督视觉变换器的新特性。In */CCV,*2021.

l6J Mark Chen, Alce Radford, Rcwon Child, Jeffrey Wu, Hccwoo Jun, David Luan, and llya Sutskcvcr.来自像素的生成性训练。In *JCML,* 2020.

l7J Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton.A simple framework for contrastive learning of visual rcp rc. cntations.In *JCML,* 2020.

l8J Xinlci Chen and Kaiming He.探索简单的暹罗语代表学习。In *CVPR,*2021.

l9J 陈新立, 谢赛宁, 和何开明.训练自我监督的视觉变形器的经验研究.In */CCV,*2021.

[I0J Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning.ELECTRA：预先训练文本编码器作为判别器而不是生成器。In */CU?,*2020.

[IJ Corinna Corte. and Vladimir Vapnik.支持-向量网络。*机器学习,* 1995.

[ 12J Ekin D Cubuk, Barret Zoph,Jonathon Shlcns, and Quoc V Le.Ran daugmcnt:实用的自动数据增强与减少的搜索空间。In *CVPR Workslwps,* 2020.

[13J Jia Deng, Wei Dong, Richard Sochcr, Li-Jia Li, Kai Li, and Li Fci-Fci. lmagcNct:一个大规模的分层图像数据库。In *CVPR,*2009.

[14J Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.BERT：用于语言理解的深度双向变换器的预训练。在*NAACL，* 2019年。

[15] Carl Docrsch, Abhinav Gupta, and Alexei A Efros.Unsupervised visual rcprc. Cntation learning by context prediction.In */CCV,* 2015.

LI6J Alexcy Dosovil kiy, Lucas Beyer, Alexander Kolcsnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Untcrthincr, Mostafa Dehghani, Matthias Mindcrcr, Georg Hcigold, Sylvain Gclly, Jakob

Uszkorcit, and Neil Houlsby.一张图片值I6x16个字。规模化图像识别的变形。在*JCLR,* 2021.

[17J Spyros Gidaris, Pravccr Singh, and Nikos Komodakis.通过预测图像旋转进行无超视距表示学习。In *ICLR,* 2018.

[ 18J Xavier Glorot和Yoshua Bengio。了解训练深度前馈神经网络的难度。In *A/STAn-,* 20 I0.

[19J Priya Goyal, Mathilde Caron, Benjamin Lcfaudeux, Min Xu, Pcngchao Wang, Vivck Pai, Mannat Singh, Vitaliy Liptchinsky, ls han Misra, Armand Joulin, and Piotr Bojanowski.*arXiv:2103.01988,*2021.自监督的野外视觉特征训练。

[20J Priya Goyal, Piotr Dollar, Ross Girshick, Pieter Noordhuis, Lukasz Wc. olowski, Aapo Kyrola, Andrew I\.tlloch, Yangqing Jia, and Kaiming He.准确的大批量SGD：在一小时内训练lmagcNct. *arXiv:1706.02677,* 2017.

[21J Jean-Bastien Grill, Florian Strub, Floren!Altchc, Corcntin Tallcc, Pierre Richemond, Elena Buchal kaya, Carl Docrsch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Ghc. hlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rcmi Munos, and Michal Valko.Boot strap yourown latent - 一种自我监督学习的新方法。在*Neur/PS,* 2020年。

[22J Raia Hadsell, Sumit Chopra, and Yann Lcun.通过学习不变的映射来降低维度。In *CVPR,*2006.

[23J Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Gir shick.动量对比用于无监督的视觉rcprc.cntation学习。In *CVPR,*2020.

[24J 何凯明，Georgia Gkioxari, Piotr Dollar, 和Ross Girshick。掩码R-CNN。In */CCV,*2017.

[25J 何开明, 张翔宇, 任少卿, 和孙健.用于图像识别的深度残差学习。In *CVPR,*2016.

[26J Dan Hcndrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Dc. ai, Tyler Zhu, Samyak Parajuli, Mike Guo, ct al. The many faces of robustnc. s:对分布外泛化的批判性分析。In */CCV,*2021.

[27J Dan Hcndrycks and Thomas Dicttcrich.对常见损坏和扰动的神经网络工作鲁棒性的基准测试。In *ICLR,* 2019.

[28J Dan Hcndrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song.自然对抗性的例子。In *CVPR,* 2021.

[29J Geoffrey E Hinton and Richard S Zemel.Autocncoders, minimum dc. cription length, and hclmholtz free energy.In *NeurJPS,* 1994.

[30J Gao Huang, YuSun,Zhuang Liu, Daniel Sedra, and Kilian Q Wein berger.具有随机深度的深度网络。In *ECCV,* 2016.

[31J Scrgcy loffc and Christian Szcgedy.批量归一化。通过减少内部协变量偏移来加速深度网络训练。在*ICML,* 2015.

[32J lnsoo Kim, Seung.ju Han, Ji-won Back, Scong-Jin Park, Jac-Jun Han, and Jinwoo Shin.通过可验证解码器的质量诊断图像识别。In *CVPR,* 2021.

[33J Alex Krizhcvsky, llya Sutskcvcr, and Geoff Hinton. lmagcnct clas sification with deep convolutional neural networks.In *Neur/PS,* 2012.

[34J Yann LcCun, Bernhard Boser, John S Denker, Donnie Hender son, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel.应用于手写邮政编码识别的反向传播法。*Neu ral compulation,* 1989.

[35J李阳浩，谢赛宁，陈新磊，Piotr Dollar, 何开明, 和Ross Girshick。用视觉变换器进行检测转移学习的基准。*正在准备中*，2021年。