DeepL
订阅DeepL Pro以翻译大型文件。
欲了解更多信息，请访问www.DeepL.com/pro。

arXiv:2301.08243v3 [cs.CV] 2023年4月13日

利用联合嵌入预测架构从图像中进行自我监督学习

Mahmoud Assran[1,2,3*] **Quentin** Duval[1] **Ishan** Misra[1] **Piotr** Bojanowski[1] **Pascal** Vincent[1] **Michael** Rabbat[1,3] **Yann** LeCun[1,4] **Nicolas** Ballas[1]

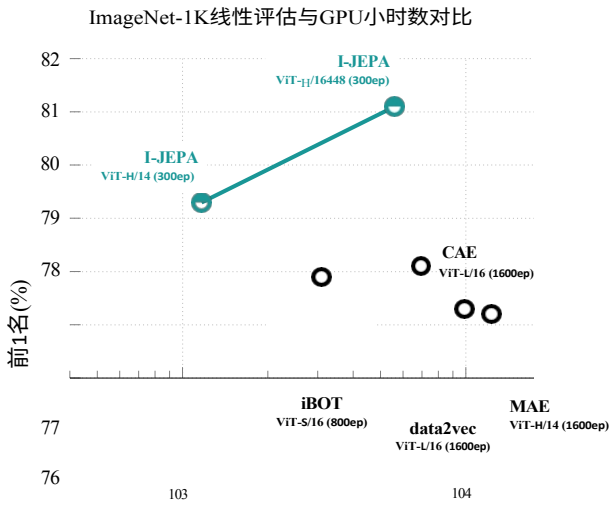
[1]元人工智能(FAIR) [2]麦吉尔大学 [3]魁北克人工智能研究所米拉 [4]纽约大学

摘要

本文展示了一种学习高语义图像表征的方法，而不依赖于手工制作的数据扩展。我们介绍了基于图像的联合嵌入预测架构（I-JEPA），这是一种用于从图像中自我监督学习的非生成性方法。I-JEPA 背后的想法很简单：从一个单一的上下文块中，预测各种目标的表征。

在同一图像中的块。指导I-JEPA产生语义表征的一个核心设计选择是掩蔽策略；具体而言，关键是要(a)对具有足够大的规模（语义）的焦油块进行采样，以及(b)

使用信息量足够大的（空间分布的）背景块。根据经验，当与视觉变换器结合时，我们发现I-JEPA具有高度的可扩展性。例如，我们使用 16 个 A100 GPU 在 72 小时内对 ImageNet 的 ViT-Huge/14 进行了训练，在从线性分类到物体计数和深度预测的各种任务中实现了强大的下游性能。



1. 简介

在计算机视觉中，有两个常见的系列方法用于从图像中进行自我监督学习：基于不变性的方法[1, 4, 10, 17, 18, 24, 35, 37, 74]和产生式方法[8, 28, 36, 57]。

基于不变性的预训练方法优化了编码器，使其对同一图像的两个或多个视图产生相似的嵌入[15, 20]，图像视图通常使用一组手工制作的数据增强来构建，如随机缩放、裁剪和颜色抖动[20]，等等[35]。这些预训练方法可以产生高语义水平的表征[4, 18]，但它们也引入了强烈的偏见，可能对某些下游任务或甚至对不同数据分布的预训练任务不利[2]。通常情况下，我们不清楚

*massran@meta.com

预培训GPU小时

图1.**ImageNet线性评估**。I-JEPA方法学习语义图像表征，在预训练期间不使用任何视图数据的增强。通过在表示空间中进行预测，I-JEPA产生了语义表示，同时使用的计算量比以前的方法少。

如何将这些偏见推广到需要不同抽象程度的任务中。例如，图像分类和实例分割不需要相同的变量[11]。此外，将这些特定于图像的增强功能归纳到其他模式（如音频）并不简单。

认知学习理论认为，生物系统中表征学习背后的驱动机制是内部模型的适应，以预先决定感觉输入的反应[31, 59]。这个想法是自监督生成方法的核心，它去除或破坏输入的部分，并学习预测被破坏的内容[9, 36, 57, 67, 68, 71]。特别是，掩蔽去噪方法通过从输入中重建随机掩蔽的斑块来学习表征，无论是在像素还是符号层面。遮蔽预训练任务比视图不变性方法需要更少的先验知识，并且很容易超越图像模式进行推广[8]。然而，这

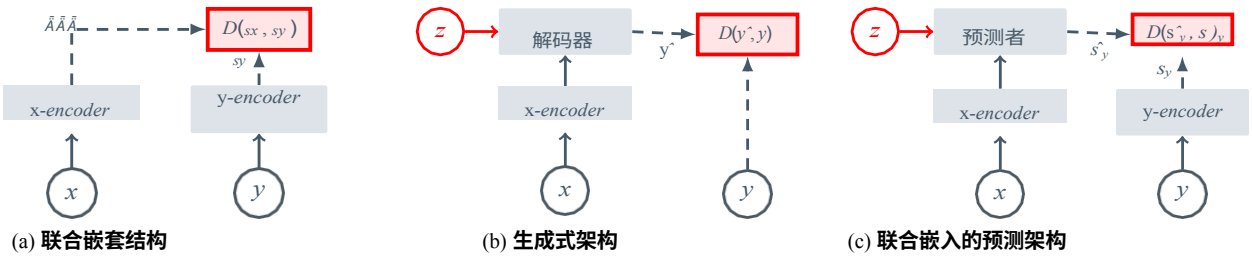


图2.自我监督学习的常见架构，其中系统学习捕捉其输入之间的关系。其目的是为不兼容的输入分配高能量（大标度值），为兼容的输入分配低能量（低标度值）。**(a)** 联合嵌入架构学习为兼容的输入x、y输出相似的嵌入，为不兼容的输入输出不相似的嵌入。**(b)** 生成式架构学习从兼容信号x直接重建信号y，使用解码器网络，该网络以额外的（可能是潜在的）变量z为条件，以促进重建。**(c)** 联合嵌入预测架构学习从兼容信号x中预测信号y的嵌入，使用预测器网络，该网络以额外的（可能是潜伏的）变量z为条件来促进预测。

由此产生的表征通常具有较低的语义水平，并且在现成的评估（例如线性探测）中和在语义分类任务的有限监督下的转移设置中，表现得不如基于不变性的预训练[4]。因此，需要一个更多的适应机制（例如，端到端的微调）来收获这些方法的全部优势。

在这项工作中，我们探讨了如何在不使用通过图像转换编码的额外先验知识的情况下提高自监督表征的语义水平。为此，我们为图像引入了一个联合嵌入预测架构[48]（I-JEPA）。图3是对该方法的说明。I-JEPA背后的想法是预测抽象代表空间中缺失的信息；例如，给定一个单一的上下文块，预测同一图像中各种目标块的代表，其中目标代表是由一个学习的目标编码器网络计算的。

与在像素/符号空间中预测的生成方法相比，I-JEPA利用抽象的预测目标，对其而言，不必要的像素级细节可能被消除，从而导致模型学习更多的语义特征。指导I-JEPA产生语义表征的另一个核心设计选择是所提出的多区块遮蔽策略。具体来说，我们证明了使用信息丰富的（空间分布的）背景块来预测图像中足够大的目标块的重要性。

通过广泛的实证评估，我们证明了这一点：

- I-JEPA学习了强大的现成表征，而没有使用手工制作的视图增强（参见图1）。I-JEPA在ImageNet-1K的线性探测、半监督1%的ImageNet-1K和语义转移任务中，优于MAE[36]等像素重建方法。
- I-JEPA与视图不变的预训练具有竞争性

在语义任务上，I-JEPA与其他方法不同，在低层次的视觉任务（如物体计数和深度预测）上取得了更好的表现（第5和6节）。通过使用一个较简单的模型和较不严格的归纳偏见，I-JEPA适用于更广泛的任务集。

- I-JEPA也是可扩展和高效的（第7节）。在ImageNet上预训练一个ViT-H/14需要不到1200个GPU小时，这比用iBOT[79]预训练的ViT-S/16快2.*5*倍以上，比用MAE预训练的ViT-H/14高效10*倍*以上。表征空间的预测大大减少了自监督预训练所需的计算量。

2. 背景介绍

自我监督学习是一种代表学习的方法，在这种方法中，一个系统学习捕捉其输入之间的关系。这个目标可以用基于能量的模型（EBM）[49]的框架来描述，其中自我监督的目标是为不相容的输入赋予高能量，并为相容的输入赋予低能量。许多现有的生成性和非生成性的自监督学习方法确实可以在这个框架中得到体现；见图2。

联合嵌入架构。基于不变性的预训练可以在EBM的框架内使用联合嵌入架构（JEA），它可以学习为兼容的输入x、y***输出***相似的嵌入，为不兼容的输入输出不相似的嵌入；见图2a。在基于图像的预训练中，兼容的x、y对通常是通过对同一输入图像随机应用手工制作的数据增强而构建的[20]。

JEAs的主要挑战是表示崩溃，其中的能量景观是平坦的（即，无论输入如何，编码器都会产生一个恒定的输出）。在过去的几年里，有几种方法已经被研究出来。

为防止表征崩溃，如对比性损失，明确推开负面例子的嵌入[15,24,37]，非对比性损失，最小化整个嵌入的信息冗余[10, 74]，以及基于聚类的方法，最大化平均嵌入的熵[4, 5, 18]。还有一些启发式的方法，利用*X-编码器*和*Y-编码器*之间的不对称结构去标志，以避免collapse [8, 24, 35] 。

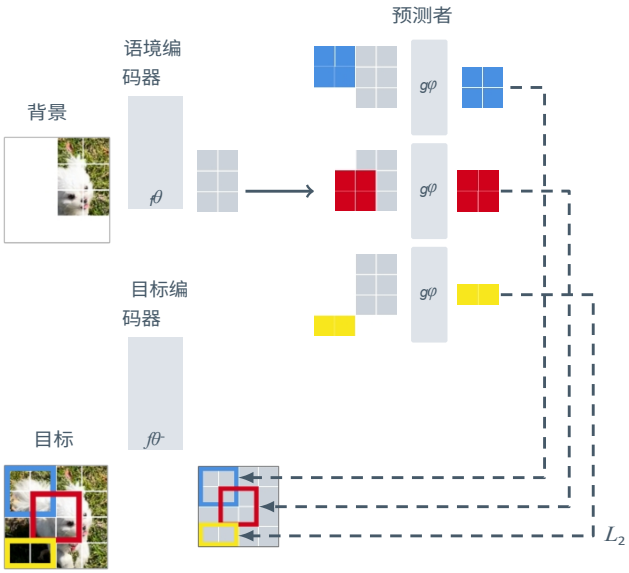
生成式架构。基于重构的自我监督学习方法也可以在EBM的框架内使用生成式架构来实现；见图2b。生成结构学习直接从一个兼容的信号$\boldsymbol{x}$中重新构建一个信号$\boldsymbol{y}$，使用一个解码器网络，该网络以一个额外的（可能是潜在的）变量z为条件来促进重建。在基于图像的预训练方面，计算机视觉中的一种常见方法是使用以下方法产生兼容的$\boldsymbol{x}$、$\boldsymbol{y}$对
屏蔽[9, 38]，其中$\boldsymbol{x}$是图像$\boldsymbol{y}$的副本，但其中一些斑块被屏蔽。然后，调节变量z对应于一组（可能是可学习的）掩码和位置标记，它向解码器指定了要重建的图像斑块。只要z的信息容量与信号$\boldsymbol{y}$***相比较***低，这些结构就不会出现表征崩溃的问题。

联合嵌入预测结构。如图2c所示，联合嵌入预测结构[48]在概念上与生成结构相似；然而，一个关键的区别是，损失函数是在嵌入空间而不是输入空间中应用。JEPA学习从兼容的信号$\boldsymbol{x}$中预测信号$\boldsymbol{y}$的嵌入，使用一个预测网络，该网络以一个附加的（可能是潜在的）变量z为条件，以促进预测。我们提出的 I-JEPA 提供了这种架构在使用掩码的图像背景下的实例；见图 3 。

与联合嵌入架构相反，JEPA并不寻求对一组手工制作的数据增强的不变的表征，而是寻求在附加信息***Z的***条件下相互预测的表征。然而，与联合嵌入架构一样，表征崩溃也是JEPA的一个问题；我们利用***X***和***Y编码器***之间的非对称架构来避免表征崩溃。

3. 方法

我们现在描述一下拟议的基于图像的联合嵌入预测



结构（I-JEPA），如图3所示。其总体目标如下：给定一个上下文区块，预测各种目标区块的表示方法

图3.**I-JEPA。**基于图像的联合嵌入预测架构使用单一的上下文块来预测来自同一图像的各种目标块的代表。上下文编码器是一个视觉变换器（ViT），它只处理可见的上下文斑块。预测器是一个狭窄的ViT，它接受上下文编码器的输出，并以po-sitional tokens（彩色显示）为条件，预测目标块在特定位置的表现。目标表征与目标编码器的输出相呼应，其权重在每次迭代中通过上下文编码器权重的指数移动平均值进行更新。

在同一图像中。我们使用视觉转化器[29, 63](ViT)架构来处理上下文编码器、目标编码器和预测器。一个ViT是由一堆转换层组成的，每个转换层都是由一个全连接的MLP的自我注意[66]操作组成的。我们的编码器/预测器架构让人联想到生成性屏蔽自动编码器（MAE）[36]方法。然而，一个关键的区别是，I-JEPA方法是非生成性的，预测是在表示空间进行的。

目标。我们首先描述我们如何在I-JEPA框架中产生目标：在I-JEPA中，目标对应于图像块的表示。给定一个输入图像$\boldsymbol{y}$，我们把它转换成N个非重叠斑块的序列，并通过目标编码器$f_{\bar{\vartheta}}$来获得相应的斑块级表示$\boldsymbol{s}_y = \{\boldsymbol{s}_y{}_1, \ldots, \boldsymbol{s}_y{}_N\}$，其中$\boldsymbol{s}_y$ k是与k^{th} 补丁相关的表示。为了获得我们的损失目标，我们随机抽取M个（可能是重叠的）块，从

$_y$我们用B_i 表示i^{th} 块的掩码反应，用$\boldsymbol{s}_y(i) = \{\boldsymbol{s}_y\}_{j \in B}$

其

原始　　背景　　目标

图4.我们的背景和目标掩蔽策略的例子。

给定一个图像，我们随机抽取4个目标块，其比例范围为（0.15, 0.2），长宽比范围为（0.75, 1.5）。接下来，我们随机抽取一个上下文块，其比例范围为

(0.85, 1.0)并删除任何重叠的目标块。在这种策略下，目标区块是相对有语义的，而上下文的

块的信息量很大，但却很稀疏（处理效率高）。

补丁级表示。通常情况下，我们将M设置为4，并在(0.75, 1.5)范围内以随机长宽比和(0.15, 0.2)范围内的随机比例对块进行采样。请注意，目标块是通过掩盖目标编码器的*输出*，而不是输入获得的。这一区别对于确保高语义水平的目标表示至关重要；例如，见[8]。

语境。回顾一下，I-JEPA背后的目标是通过单一的背景块来预测目标块的表征。为了获得I-JEPA的上下文，我们首先从图像中抽出一个单一的块$\boldsymbol{x}$，其随机比例范围为（0.85，1.0），长宽比为单位。我们用B_x表示与上下文块$\boldsymbol{x}$相关的掩码。由于目标块是独立于上下文块采样的，因此可能会有大量的重叠。为了确保一个非琐碎的预测任务，我们从上下文块中删除任何重叠的区域。图4显示了实践中各种背景和目标块的例子。接下来是被屏蔽的上下文块、

$\boldsymbol{x}$，通过上下文编码器f_ϑ，得到相应的补丁级表示$\boldsymbol{s}_x = \{\boldsymbol{s}_{x\,j}\}_{j\in B_x}$。

预测。鉴于上下文编码器的输出，$\boldsymbol{s}_x$，我们希望预测M个目标块的表征$\boldsymbol{s}_y(1), \ldots, \boldsymbol{s}_y(M)$。为此，对于一个给定的目标块$\boldsymbol{s}_y(i)$对应的目标掩码$B_i$，预测器$g_\phi(\cdot,\cdot)$将上下文编码器的输出$\boldsymbol{s}_x$和我们希望预测的每个补丁的掩码标记作为输入、

$\{\boldsymbol{m}_j\}_{j\in B_i}$，并输出一个补丁级预测$\hat{\mathbf{s}}_y(i) = \{\hat{\boldsymbol{s}}_{y\,j}\}_{j\in B_i} = g_\phi(\boldsymbol{s}_x, \{\boldsymbol{m}_j\}_{j\in B_i})$。掩码令牌是参数是一个共享的可学习向量，有一个附加的位置嵌入。由于我们希望对M个目标块进行预测，我们应用我们的预测器M次，每次都对我们希望预测的目标块位置所对应的掩码标记进行调节，并获得预预测$\hat{\mathbf{s}}_y(1), \ldots, \hat{\mathbf{s}}_y(M)$。

损失。损失只是预测的斑块级表征$\hat{\boldsymbol{s}}_y(i)$和target-get斑块级表征$\boldsymbol{s}_y(i)$之间的平均$L_2$距离；即、

$$D\left(\hat{\mathbf{s}}_y(i), \boldsymbol{s}_y(i)\right) = \frac{1}{M}\sum_{i=1}^{M}\sum_{j\in B_i}\|\hat{\boldsymbol{s}}_{y\,j} - \boldsymbol{s}_{y\,j}\|_2^2.$$

预测器的参数j和上下文编码器的参数ϑ是通过基于梯度的优化学习的，而目标编码器的参数$\bar{\theta}$ 是通过上下文编码器参数的指数移动平均值来更新的。指数移动平均目标编码器的使用已被证明对用视觉转换器训练JEA至关重要[18, 25, 79]，我们发现I-JEPA也是如此。

4. 相关工作

一系列的工作都是通过预测缺失或损坏的感觉输入的值来探索视觉表征学习。去噪自动编码器使用随机噪声作为输入破坏[67]。上下文编码器根据其周围的情况对整个图像区域进行回归[57]。其他工作将图像着色作为去噪任务[46, 47, 77]。

最近，图像去噪的想法在遮蔽图像建模的背景下被重新审视[9, 36, 71]，其中视觉变换器[29]被用来重建丢失的输入斑块。关于遮蔽自动编码器（MAE）的工作[36]提出了一个高效的架构，只要求编码器处理可见的图像斑块。通过在像素空间中重建缺失的斑块，MAE在大型标记数据集上进行端到端微调时取得了强大的性能，并表现出良好的扩展特性。BEiT[9]预测缺失斑块在标记化空间中的价值；具体而言，使用冻结的离散VAE对图像斑块进行标记化，该标记化是在包含2.5亿张图像的数据集上训练的[58]。然而，像素级的预训练已被证明优于BEiT的微调效果[36]。另一项工作，SimMIM[71]，探索了基于经典的Histogram of Gradients[27]fea- ture空间的重建tar- gets，并证明了比像素空间重建的一些优势。与这些工作不同的是，我们的再现空间是在训练期间通过联合嵌入预测架构学习的。我们的目标是学习语义表征，不需要对下游任务进行广泛的微调。

与我们的工作最接近的是data2vec[8]和Context Autoen- coders[25]。data2vec的方法是通过学习来预测

方法	拱门。	纪元	顶-1
没有视图数据增强的方法没有视图数据增强的			
data2vec [8]	ViT-L/16	1600	77.3
MAE [36]	ViT-B/16	1600	68.0
	ViT-L/16	1600	76.0
	ViT-H/14	1600	77.2
CAE [22]	ViT-B/16	1600	70.4
	ViT-L/16	1600	78.1
SimCLR v2 [21]	RN152（2×）。	800	79.1
DINO[18]	ViT-B/8	300	80.1
iBOT [79]	ViT-L/16	250	**81.0**

表1.**ImageNet**。对ImageNet-1k的线性评价（ViT-H/$_{16448}$是在448×448的分辨率下进行预训练的）。与其他方法相比，I-JEPA提高了线性探测性能。

在预训练期间不依赖手工制作的视图数据增强。此外，I-JEPA展示了良好的可扩展性--较大的I-JEPA模型与视图不变性方法的性能相匹配，而不需要视图数据增强。

通过在线目标编码器计算的缺失斑块的重现；通过避免手工增强，该方法可以应用于不同的模式，并在视觉、文本和语音方面取得可喜的成果。上下文自动编码器使用一个通过重建损失和对齐约束的总和进行优化的编码器/解码器架构，它强制执行表示空间中缺失斑块的可预测性。与这些方法相比，I-JEPA在计算效率方面有明显的改进，并能学习到更多语义上的现成表征。与我们的工作同时，data2vec-v2[7]探索了用于学习各种模式的有效架构。

我们还将 I-JEPA 与基于联合嵌入架构的各种方法进行了比较；例如 DINO [18]、MSN [4] 和 iBOT [79] 。这些方法在预训练期间依靠手工制作的数据增量来学习语义图像表征。关于MSN[4]的工作，在预训练期间使用掩码作为额外的数据增强，而iBOT将数据2vec风格的补丁级重构损失与DINO视图不变性损失相结合。这些方法的共同点是需要处理每个输入图像的多个用户生成的视图，从而阻碍了可扩展性。相比之下，I-

方法	拱门。	纪元	顶-1
方法			
data2vec [8]	ViT-L/16	1600	73.3
MAE [36]	ViT-L/16	1600	67.1
	ViT-H/14	1600	71.5
I-JEPA	ViT-L/16	600	69.4
	ViT-H/14	300	73.3
	ViT-H/16448	300	**77.3**

JEPA只需要处理每个图像的单一视图。我们发现，用I-JEPA 训练的 ViT-Huge/14 比用 iBOT 训练的 ViT-Small/16需要的计算量要少。

I-JEPA	ViT-B/16	600	72.9
	ViT-L/16	600	77.5
	ViT-H/14	300	79.3
	ViT-$_{H}$/16448	300	**81.1**
使用额外视图数据增强的方法			
MSN[4]	ViT-B/4	300	**75.7**
使用额外视图数据增强的方法			
iBOT [79]	ViT-B/16	400	69.7
DINO[18]	ViT-B/8	300	70.0
SimCLR v2 [35]	RN151（2×）。	800	70.2
byol[35]	rn200 (2×)	800	71.2

表2.**ImageNet-1%**。对ImageNet-1K的半监督评估只使用了1%的可用标签。模型通过微调或线性探测进行调整，取决于哪一个

对每一种方法都有最佳效果。ViT-$_{H}$/16448在448×448的分辨率下进行预训练。I-JEPA的预训练优于MAE，MAE也不依赖于手工制作的数据增强。

在预培训期间。此外，I-JEPA还得益于规模。一个ViT-

在448分辨率下训练的H/16超过了以前的方法，包括利用额外手工制作的数据增强的方法。

5. 图像分类

为了证明I-JEPA在不依赖手工制作的数据扩展的情况下学习高级代表，我们报告了使用线性探测和部分微调协议的各种图像分类任务的结果。在本节中，我们考虑在ImageNet-1K数据集[60]上预训练的自监督模型。预训练和评估实施细节在附录A中描述。除非另有说明，所有I-JEPA模型都是在224×224像素的分辨率下训练的。

ImageNet-1K。 表1显示了ImageNet-1K线性评价基准的性能。在自我监督的预训练之后，模型的权重被冻结，在此基础上使用完整的ImageNet- 1K训练集来训练线性分类器。与流行的方法，如屏蔽自动编码器（MAE）[36]、上下文自动编码器（CAE）[22]和data2vec[8]（这些方法在预训练期间也不依赖于大量的手工制作的数据增强）相比，我们看到I-JEPA明显提高了线性探测性能，同时使用更少的计算工作（见第7章）。通过利用I-JEPA提高的效率，我们可以训练更大的模型，在使用一小部分计算量的情况下，超越最好的CAE模型。I-JEPA也受益于规模；特别是，在分辨率为448×448像素的情况下训练的ViT-H/16的性能与视图-H/16的性能相当。

方法	拱门。	CIFAR100	地点205	iNat18
没有视图数据增强的方法				
data2vec [8]	ViT-L/16	81.6	54.6	28.1
MAE [36]	薇塔-H/14	77.3	55.0	32.9
I-JEPA	薇塔-H/14	**87.5**	**58.4**	**47.6**
使用额外视图数据增强的方法使用额外数据增强的方法				
DINO [18]	ViT-B/8	84.9	57.9	55.9
iBOT [79]	ViT-L/16	**88.3**	**60.4**	**57.3**

方法	拱门。	悬臂/计数	缆线/距离
没有视图数据增强的方法			
data2vec [8]	ViT-L/16	85.3	71.3
MAE [36]	ViT-H/14	**90.5**	**72.4**
I-JEPA	ViT-H/14	86.7	**72.4**
方法			
DINO [18]	ViT-B/8	86.6	53.4
iBOT [79]	ViT-L/16	85.7	62.8

表3.**用于图像分类的线性探针转移**。对下游图像分类任务的线性评估。I-JEPA明显优于以前不使用增量的方法（MAE和data2vec），并缩小了与在预训练期间利用手工制作的数据增量的最佳基于视图不变性的方法的差距。

诸如iBOT[79]等不变的方法，尽管避免了使用手工制作的数据扩展。

Low-Shot ImageNet-1K。表2显示了1%的ImageNet基准的性能。这里的想法是使预训练的模型适用于ImageNet分类，只使用1%的可用ImageNet标签，大约相当于每类12或13张图像。模型通过微调或线性探测进行调整，取决于哪种方法最适合各自的工作。在使用类似的编码器架构时，I-JEPA优于MAE，同时需要更少的预训练epochs。I-JEPA使用ViT- H/14架构，与使用data2vec[8]预训练的ViT-L/16的性能相匹配，而使用的计算量却大大减少（见第7节）。通过提高图像输入分辨率，I-JEPA优于以前的方法，包括在预训练期间利用额外的手工制作的数据增强的联合嵌入方法，如MSN[4]、DINO[17]和iBOT[79]。

转移学习。表3显示了使用线性探针的各种下游图像分类任务的性能。I-JEPA明显优于以前不使用增强的方法（MAE和data2vec），并缩小了与基于视图不变性的最佳方法的差距，这些方法在预训练期间利用手工制作的数据增强，甚至在CIFAR100和Place205上超过了流行的DINO[18]的线性探头。

6. 本地预测任务

如第5节所示，I-JEPA学习了语义图像表征，大大改善了以前的方法，如MAE和data2vec的下游图像分类性能。此外，I-JEPA从规模中获益，可以缩小差距，甚至超过、

表 4. **低水平任务的线性探针转移**。对由物体计数（Clevr/Count）和深度预测（Clevr/Dist）组成的下游低层次任务进行线性评估。I-JEPA方法在预训练中有效地捕捉了低层次的图像特征，并在物体计数和深度预测等任务上优于基于视图不变性的方法。

基于视图不变性的方法，利用额外的手工制作的数据增强。在这一节中，我们发现I-JEPA也能学习局部图像特征，并在低层次和密集的预测任务（如物体计数和深度预测）上超过了基于视图不变性的方法。

表4显示了使用线性探针的各种低层次任务的性能。在预训练后，编码器的权重被冻结，并在上面训练一个线性模型，在Clevr数据集上进行物体计数和深度预测[43]。与DINO和iBOT等视图不变的方法相比，I-JEPA方法在预训练期间有效地捕捉了低层次的图像特征，并在物体计数（Clevr/Count）和（以很大的幅度）深度预测（Clevr/Dist）方面胜过它们。

7. 可扩展性

模型效率。与以前的方法相比，I-JEPA 具有高度的可扩展性。图 5 显示了在 1% 的 ImageNet-1K 上进行的半监督评估与 GPU 时间的关系。I-JEPA 比以前的方法需要更少的计算量，并在不依赖手工制作的数据扩展的情况下实现了强大的性能。与直接使用像素作为目标的基于重建的方法（如MAE）相比，I-JEPA通过在表示空间中计算焦油而引入了额外的开销（每次迭代大约慢7%的时间）。然而，由于I-JEPA收敛的迭代次数大约少了5倍，我们在实践中仍然看到显著的计算节省。与基于视图不变性的方法相比，如iBOT，它依靠手工制作的数据增强来创建和处理每个图像的多个视图，I-JEPA的运行速度也明显加快。特别是，一个巨大的I-JEPA模型（ViT-H/14）需要的计算量比一个小的iBOT模型（ViT-S/16）少。

扩大数据规模。我们还发现 I-JEPA 从较大的数据集的预训练中受益。 表5显示了转移

预训	拱门。	CIFAR100	地点205	INat18	缆绳/计数	缆线/距离
IN1k	薇塔-H/14	87.5	58.4	47.6	86.7	72.4
IN22k	薇塔-H/14	**89.5**	57.8	50.5	**88.6**	**75.0**
IN22k	ViT-G/16	**89.5**	**59.1**	**55.3**	86.7	73.0

表5.**消融数据集和模型大小**。评估预训练数据集大小和模型大小对转移任务的影响。I-JEPA从更大更多样化的数据集中获益。当增加预训练数据集的大小（IN1k与IN22k）时，我们看到ViT-H/14模型的性能提高。通过在ImageNet-22k上训练一个更大的模型ViT- G/16模型，我们观察到在语义任务上有进一步的性能改进。ViT-H/14在IN1k上训练了300个历时，在IN22k上相当于900个IN1K历时。ViT-H/16的训练量相当于600个IN1k epochs。

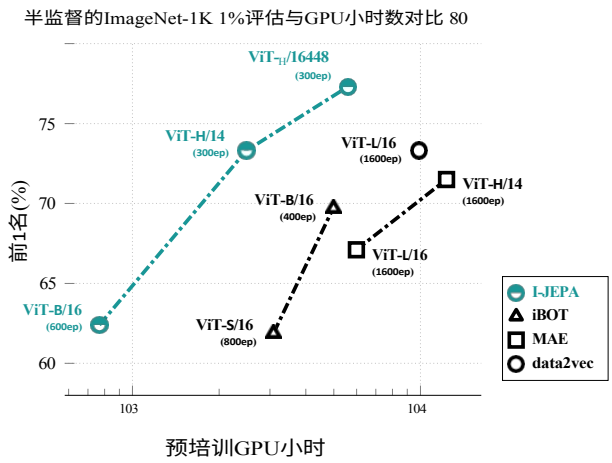


图5.**缩放**。对 ImageNet-1K 1% 的半监督评估是预训练 GPU 时间的函数。I-JEPA 需要比以前的方法更少的计算量来实现强大的性能。与 MAE 和 data2vec 相比，I-JEPA 通过需要更少的预训练 epochs 而获得了显著的速度提升。与依赖于手工制作的数据增强的iBOT相比，一个巨大的I-JEPA模型（ViT-H/14）比他们最小的模型（ViT-S/16）需要更少的计算。

当增加预训练数据集的大小（IN1K与IN22K）时，在语义和低水平任务上的学习性能。当在一个更大的、更多样化的数据集上进行预训练时，这些概念上不同的任务的迁移学习性能会得到改善。

缩放模型大小。表5还显示，在IN22K上进行预训练时，I-JEPA从较大的模型规模中获益。与ViT-H/14模型相比，ViT-G/16的预训练明显改善了Place205和INat18等图像分类任务的下游性能，但没有改善低级下游任务的性能--ViT-G/16使用更大的输入斑块，这对局部预测任务是不利的。

8. 预测器的可视化

预测器在I-JEPA中的作用是接受上下文编码器的输出，并以位置掩码为条件，预测目标黑色在掩码指定的位置的表现。一个自然的问题是，以位置掩码标记为条件的预测器是否正在学习正确捕捉目标的位置不确定性。为了定性地研究这个问题，我们将预测器的输出可视化。我们使用以下的可视化方法，使研究界能够独立地重现我们的发现。在预训练之后，我们冻结了上下文编码器和预测器的权重，并按照RCDM框架[13]训练一个解码器，将预测器输出的平均池映射到像素空间。图6显示了各种随机种子的解码器输出。各个样本的共同特征代表了包含在平均池的预测器表现中的信息。I-JEPA预测器正确地捕捉了位置上的不确定性，并产生了具有正确姿势的高级物体部分（例如，鸟的背面和汽车的顶部）。

9. 消融

在表示空间进行预测。表7比较了当损失在像素空间与表示空间计算时，使用线性探针在1%的ImageNet-1K上的低拍性能。我们猜想，I-JEPA的一个重要组成部分是损失完全在表示空间中计算，从而使目标编码器有能力产生抽象的预测目标，对于这些目标，不相关的像素级细节被消除了。从表7可以看出，在像素空间进行预测会导致线性探测性能的明显下降。

屏蔽策略。表6将我们的多块遮蔽与其他遮蔽策略进行了比较，如栅格化遮蔽，其中图像被分割成四个大象限，目标是使用一个象限作为背景来预测其他三个象限，以及基于重建的方法中通常使用的传统块和随机遮蔽。在块状遮蔽中，目标是一个单一的图像块，背景是

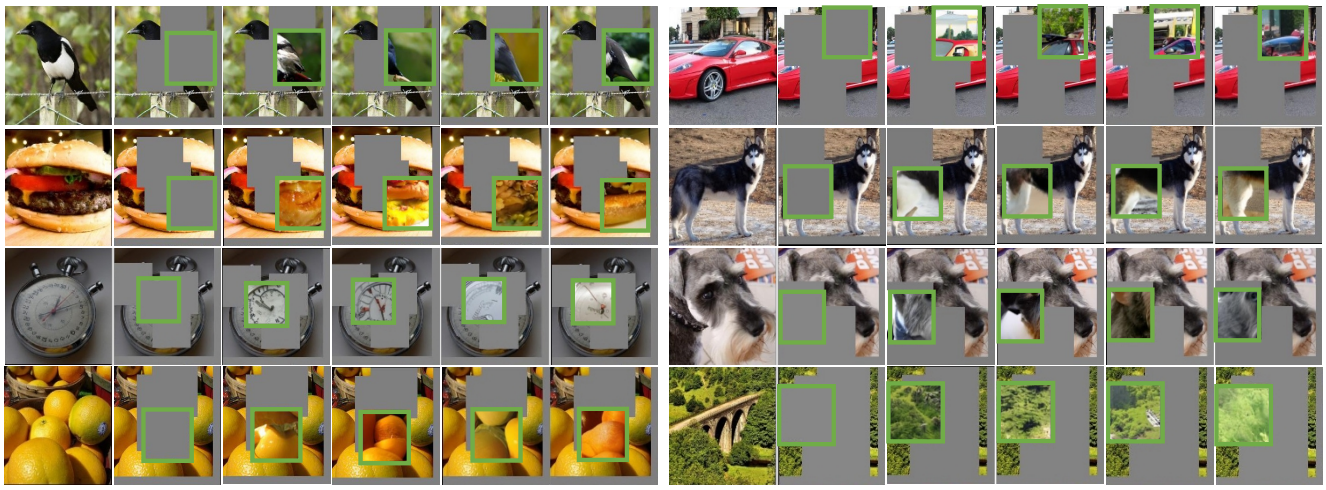


图6.**I-JEPA预测器表征的可视化**。对于每幅图像：第一列包含原始图像；第二列包含背景图像，由预训练的 I-JEPA ViT-H/14编码器处理。后面几列的绿色边框包含生成模型对预训练的I-JEPA预测器的输出进行解码的样本，其条件是对应于绿色边框位置的位置掩码标记。各个样本的共同特征代表了I-JEPA预测中包含的信息。I-JEPA预测器正确地捕捉了位置的不确定性，并产生了具有正确姿势的高级物体部分（例如，鸟的背部和汽车的顶部）。在不同样本之间变化的特性代表了不包含在表示中的信息。在这种情况下，I-JEPA预测器放弃了精确的低层次细节以及背景信息。

	目标		**背景介绍**		
面罩	类型	频率。	类型	平均。比率*	**顶-1**
多区块	块(0.15, 0.2)	4	区块(0.85, 1.0) × 补数	0.25	**54.2**
栅格化	四象限	3	补充	0.25	15.5
块	块(0.6)	1	补充	0.4	20.2
随机	随机(0.6)	1	补充	0.4	17.6

*Avg. *比率是指相对于图像中的补丁总数而言，上下文区块中的补丁的平均数量。*

表6.**消融掩蔽策略**。在对ViT-B/16进行I-JEPA预训练300次后，仅使用1%的可用标签对ImageNet-1K进行线性评估。拟议的多块屏蔽策略的比较。在光栅化遮蔽中，图像被分割成四个大象限；一个象限被用作预测其他三个象限的背景。在块状遮蔽中，目标是一个单一的图像块，背景是图像的补充。在随机遮蔽中，目标是一组随机图像斑块，上下文是图像补充。所提出的多块遮蔽策略有助于指导I-JEPA学习语义表征。

目标	**拱门。**	**纪元**	**顶点-1**
目标-编码器输出	ViT-L/16	500	**66.9**
像素	ViT-L/16	800	40.7

表7.**消融的目标**。在ImageNet-1K上仅使用1%的可用标签进行线性评估。当损失应用于像素空间而非表示空间时，I-JEPA表示的语义水平明显下降，突出了预训练期间目标编码器的重要性。

景是图像的补充。请注意，在所有考虑的策略中，背景和目标块之间没有重叠。我们

图像补充。在随机遮蔽中，目标是一组随机斑块，背

发现多区块屏蔽有助于指导I-JEPA学习语义表征。关于多区块遮蔽的其他消融可以在附录C中找到。

10. 总结

我们提出了I-JEPA，这是一种简单而有效的学习语义图像表征的方法，不依赖于手工制作的数据增强。我们表明，通过预测表征空间，I-JEPA比像素重建方法收敛得更快，并能学习高语义水平的表征。与基于视图不变性的方法相比，I-JEPA 突出了一条学习具有联合嵌入架构的一般表征的道路，而无需重新依赖手工制作的视图增强。

参考文献

[1] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. 通过同步聚类和表征学习的自我标签。*Internatinoal Conference on Learning Representations*, 2020.1

[2] Mahmoud Assran, Randall Balestriero, Quentin Duval, Florian Bordes, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, and Nicolas Ballas. 自监督学习中的隐藏统一集群先验。*International Con- ference on Learning Representations*, 2023.1, 13

[3] Mahmoud Assran, Nicolas Ballas, Lluis Castrejon, and Michael Rabbat.监督加速了视觉表征的对抗性半监督学习的预训练。*自我监督学习的NeurIPS研讨会*, 2020.13

[4] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Michael Rabbat, and Nicolas Ballas.用于标签有效学习的遮蔽连体网。*欧洲会议 计算机视觉*，2022。1, 2, 3, 5, 6, 12, 13, 16, 17

[5] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Armand Joulin, Nicolas Ballas, and Michael Rabbat.通过支持样本的非参数化预测视图分配来进行视觉特征的半监督学习。*IEEE/CVF国际计算机会议 Vision*, 2021.3, 13

[6] Philip Bachman, R Devon Hjelm, and William Buchwalter. 通过最大化跨视图的相互信息来学习表征。*神经信息处理的进展 系统*，32，2019。13

[7] Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli.高效的自我监督学习与视觉、语音和语言的情境化目标表征。*arXiv preprint arXiv:2212.07525*, 2022。5

[8] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli.Data2vec： *arXiv preprint arXiv:2202.03555*, 2022.1, 3, 4, 5, 6, 13

[9] Hangbo Bao, Li Dong, and Furu Wei.Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.1, 3, 4, 13

[10] Adrien Bardes, Jean Ponce, and Yann LeCun. Vi- creg: Variance-invariance-covariance regularization for self supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.1, 3, 13

[11] Adrien Bardes, Jean Ponce, and Yann LeCun.Vicregl：局部视觉特征的自我监督学习。*arXiv 预印本 arXiv:2210.01571*, 2022.1, 13

[12] Florian Bordes, Randall Balestriero, Quentin Garrido, Adrien Bardes, and Pascal Vincent.Guillotine regularization：通过删除，改善深度网络的泛化。*arXiv preprint arXiv:2206.13378*, 2022.13

[13] Florian Bordes, Randall Balestriero, and Pascal Vincent.高保真可视化你的自我监督代表知道什么。*Transactions on Machine Learning Research*, 2022.7, 16

[14] John Bridle, Anthony Heading, and David MacKay.无监督的分类器、相互信息和 "幻影焦油"。

得到。*神经信息处理系统的进展*，4， 1991。13

[15] Jane Bromley, James W Bentz, Le´on Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard Sa¨ckinger, and Roopak Shah.使用 "连体 "时间延迟神经网络的签名验证。*国际模式识别杂志 和人工智能*，7（04）：669-688，1993。1, 3

[16] Zhaowei Cai, Avinash Ravichandran, Paolo Favaro, Manchen Wang, Davide Modolo, Rahul Bhotika, Zhuowen Tu, and Stefano Soatto.半监督视觉变换器 at scale. *arXiv preprint arXiv:2208.05688*, 2022.13

[17] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Pi- otr Bojanowski, and Armand Joulin.*ArXiv preprint arXiv:2006.09882*, 2020.1, 6

[18] Mathilde Caron, Hugo Touvron, Ishan Misra, Herve´ Je´gou, Julien Mairal, Piotr Bojanowski, and Armand Joulin.*arXiv 预印本 arXiv:2104.14294*, 2021.Emerg- ing properties in self-supervised vision transformers.1, 3, 4, 5, 6, 12, 13

[19] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Hee- woo Jun, David Luan, and Ilya Sutskever.来自像素的生成性预训练。在*Ma- chine Learning国际会议上*，第1691-1703页。PMLR, 2020.13

[20] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Ge- offrey Hinton.*预印本arXiv:2002.05709*，2020年，视觉表征对比学习的简单框架。1, 2, 13

[21] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton.Big self-supervised mod- els are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020.5

[22] 陈小康，丁明宇，王小迪，辛颖，莫申彤，王云浩，韩淑敏，罗平，曾刚，王敬东。自监督表征学习的上下文自动编码器。*arXiv 预印本 arXiv:2202.03026*, 2022.5

[23] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. *arXiv preprint arXiv:2003.04297*, 2020.12, 13

[24] Xinlei Chen and Kaiming He.探索简单的连体表示学习。*arXiv预印本arXiv:2011.10566*， 2020。1, 3, 13

[25] Xinlei Chen, Saining Xie, and Kaiming He.*arXiv preprint arXiv:2104.02057*, 2021.关于训练自监督视觉变换器的经验研究。4

[26] Yubei Chen, Adrien Bardes, Zengyi Li, and Yann LeCun.Intra-instance vicreg: Bag of self-supervised image patch embedding. *arXiv preprint arXiv:2206.08954*, 2022.13

[27] Navneet Dalal and Bill Triggs.用于人类检测的定向重力柱状图.In *2005 IEEE computer soci-ety conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886-893.Ieee, 2005.4

[28] 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。 *arXiv预印本arXiv:1810.04805*，2018。1

[29] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words：跨越尺度的图像识别。*arXiv预印本 arXiv:2010.11929*, 2020。3, 4, 12, 13

[30] Alaaeldin El-Nouby, Gautier Izacard, Hugo Touvron, Ivan Laptev, Herve´ Jegou, and Edouard Grave.大规模数据集对自我监督的预训练是必要的吗？*arXiv 预印本 arXiv:2112.10740*, 2021.13

[31] 卡尔-弗里斯顿皮质反应的理论。*皇家学会的哲学交易B：生物科学*，360(1456)：815-836，2005。1

[32] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pe´rez, and Matthieu Cord.通过预测视觉词包学习表征。在*IEEE/CVF计算机视觉和模式会议论文集 识别*，第6928-6938页，2020。13

[33] Ian Goodfellow, Yoshua Bengio, and Aaron Courville.*深度学习*。MIT press, 2016.13

[34] Priya Goyal, Quentin Duval, Jeremy Reizenstein, Matthew Leavitt, Min Xu, Benjamin Lefaudeux, Mannat Singh, Vinicius Reis, Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Ishan Misra 。Vissl. https://github.com/facebookresearch/vissl, 2021.12

[35] Jean-Bastien Grill, Florian Strub, Florent Altche´, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent：a new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.1, 3, 5, 12, 13

[36] 何开明，陈新磊，谢赛宁，李阳浩，Piotr Dolla´r，和Ross Girshick。掩码自动编码器是可扩展的视觉学习器。*IEEE/CVF计算机视觉会议 和模式识别*，2022年。1, 2, 3, 4, 5, 6, 12, 13, 15, 16

[37] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick.Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.1, 3, 12, 13

[38] 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 *Proceed- ings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770-778, 2016.3

[39] Olivier Henaff.数据效率高的图像识别与自相矛盾的预测编码。在 *机器学习国际会议上*，第4182-4192页。PMLR, 2020.13

[40] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio.通过相互形成估计和最大化学习深度表征。*arXiv预印本 arXiv:1808.06670*, 2018.13

[41] Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama.通过信息最大化的自我增强训练学习离散表征。In- *ternational conference on machine learning*, pages 1558- 1567.PMLR, 2017.13

[42] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross

吉尔希克.Clevr：一个用于组合语言和基本视觉推理的诊断数据集。In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901-2910, 2017.12

[43] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick.Clevr：用于组合语言和元素的诊断数据集 tary 视觉推理。In *CVPR*, 2017.6

[44] Andreas Krause, Pietro Perona, and Ryan Gomes.通过正则化的信息最大化进行的犯罪性聚类。*Advances in neural information processing systems*, 23, 2010.13

[45] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.2009.12

[46] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich.为自动 着色学习表征。2016.4

[47] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich.着色作为视觉的代理任务 理解。2017.4

[48] Yann LeCun.通往自主机器智能的道路 gence 0.9版。2, 2022-06-27.2022.2, 3

[49] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fujie Huang.基于能量的学习教程。*预测 结构化数据*，1（0），2006。2

[50] Ralph Linsker.感知网络中的自组织。*计算机*，21（3）：105-117，1988。13

[51] Ilya Loshchilov and Frank Hutter.解耦权重衰减 正则化。*arXiv预印本arXiv:1711.05101*, 2017。12

[52] Yi Ma, Doris Tsao, and Heung-Yeung Shum.On the principles of parsimony and self-consistency for the emergence of intelligence.*Frontiers of Information Technology & Elec- tronic Engineering*, pages 1-26, 2022.13

[53] Ishan Misra and Laurens van der Maaten.幌子不变量表征的自我监督学习。In *Proceed- ings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6707-6717, 2020.13

[54] Jovana Mitrovic, Brian McWilliams, Jacob Walker, Lars Buesing, and Charles Blundell.通过不变的因果机制进行表征学习。*学习表征的国际会议*，2021年。13

[55] Aaron van den Oord, Yazhe Li, and Oriol Vinyals.*ArXiv preprint arXiv:1807.03748*, 2018.13

[56] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch：一个具有代表性的、高性能的深度学习库。*Ad- vances in neural information processing systems*, 32, 2019.12

[57] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros.语境编码器：通过画图进行特征学习。In *Proceedings of the IEEE con- ference on computer vision and pattern recognition*, pages 2536-2544, 2016.1, 4

[58] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever.

零距离文本到图像的生成。In *International Confer- ence on Machine Learning*, pages 8821-8831.PMLR, 2021.4

[59] Rajesh PN Rao和Dana H Ballard.视觉皮层中的预测编码：对一些类外接受场效应的功能解释。*Nature neuroscience*, 2(1):79-87, 1999.1

[60] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, San- jeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, 和李飞飞。Imagenet大规模视觉识别挑战。*国际计算机视觉杂志*，115（3）：211-252， 2015。5, 12

[61] Antti Tarvainen和Harri Valpola.Mean teachers are bet- ter role models：加权平均一致性目标im- prove半监督式深度学习结果。*arXiv预印本 arXiv:1703.01780*, 2017。12

[62] Yuandong Tian, Xinlei Chen, and Surya Ganguli.在没有连贯性对的情况下理解自我监督学习的动态。在*机器 学习国际会议上*，第10268-10278页。PMLR, 2021.13

[63] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve´ Je´gou.通过注意力训练数据有效的图像变换器和提炼。在*国际机器学习会议上*， 第10347-10357页。PMLR，2021年。3

[64] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic.On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, 2019.13

[65] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie.Inaturalist物种分类和保护数据集。在*IEEE计算机视觉和模式识别会议论文集中*，第8769-8778页， 2018。12

[66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszko- reit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin.注意力是你所需要的一切。In *Advances in neural information processing systems*, pages 5998-6008, 2017.3

[67] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Le´on Bottou.叠加去噪自动编码器：在一个具有局部去噪标准的深度网络中学习有用的代表。*机器学习研究杂志*, 11(12), 2010.1, 4, 13

[68] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer.*arXiv 预印本 arXiv:2112.09133*, 2021.Masked feature predic- tion for self-supervised visual pre-training.1, 13

[69] 吴志荣, 熊元军, 余思达, 和林大华.通过非参数实例判别进行无监督的特征学习。在*IEEE计算机视觉和模式识别会议论文集中*，第3733-3742页， 2018。13

[70] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le.*arXiv preprint arXiv:1904.12848*, 2019年.无监督数据增强.13

[71] 谢振达，张正，曹越，林雨桐，鲍建民，姚竹良，戴琦，胡瀚。Simmim：A sim-ple framework for masked image modeling. *arXiv preprint arXiv:2111.09886*, 2021.1, 4

[72] Yang You, Igor Gitman, and Boris Ginsburg.卷积网络的大批量 训练，2017。12

[73] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo.Cutmix：用正则化策略训练具有可定位特征的强分类器。In *Proceedings of the IEEE/CVF International Con- ference on Computer Vision*, pages 6023-6032, 2019.16

[74] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Ste´phane Deny.巴洛双胞胎：通过减少冗余进行自我监督学习。*arXiv预印本arXiv:2103.03230*， 2021。1, 3, 13

[75] 翟晓华， Joan Puigcerver， Alexander Kolesnikov， Pierre Ruyssen，Carlos Riquelme，Mario Lucic，Josip Djo- longa， Andre Susano Pinto， Maxim Neumann， Alexey Doso- vitskiy，Lucas Beyer，Olivier Bachem， Michael Tschannen，Marcin Michalski，Olivier Bousquet ，Sylvain Gelly，和Neil Houlsby。表征学习的大规模研究与 视觉任务适应基准，2019年。12

[76] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup ： Beyond empirical risk minimiza- tion.*Internatinoal Conference on Learning Representations*, 2018.16

[77] Richard Zhang, Phillip Isola, and Alexei A Efros.彩色的图像着色。2016.4

[78] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Tor- ralba, and Aude Oliva.使用地方数据库学习场景识别的深度特征。*Advances in neural in- formation processing systems*, 27, 2014.12

[79] 周景浩，陈伟，王辉宇，沈伟，谢慈航，Alan Yuille ，和孔涛。Ibot：带有在线标记器的图像贝尔特预训练。*International Conference on Learning Representations*, 2022.2, 4, 5, 6, 12, 13

A. 实施细节

A.1. 预培训

架构。对于I-JEPA预训练，我们使用Vision Transformer[29](ViT)架构来设计上下文编码器、目标编码器和预测器。虽然上下文编码器和目标编码器对应于标准的ViT架构，但预测器被设计为一个轻量级（窄）的ViT架构。具体来说，我们将预测器的嵌入维度固定为384，同时保持自我关注头的数量与骨干语境编码器的数量相等。对于较小的ViT-B/16语境编码器，我们将预测器的深度设定为6。对于ViT-L/16、ViT-H/16和ViT-H/14上下文编码器，我们将预测器的深度设置为12。最后，ViT-G/16使用深度为16的预测器。I-JEPA在没有`[cls]`标记的情况下进行了预训练。我们使用目标编码器进行评估，并平均汇集其输出，以产生一个全局图像表示。

优化。我们使用AdamW[51]来优化上下文编码器和预测器的权重。我们的默认批处理量为2048，在预训练的前15个历时中，学习率从10^{-4}线性增加到10^{-3}，此后按照余弦计划衰减到10^{-6}。按照文献[4, 18]，在整个训练过程中，权重衰减从0.04线性增加到0.4。
预训练。目标编码器的权重在初始化时与上下文编码器的权重相同，此后通过指数移动平均来更新[4, 18, 23, 35, 37, 61]。我们使用0.996的动量值，并线性地增加这个动量值。
在整个预训练过程中，按照[4，18]的规定，将数值调至1.0。

屏蔽。默认情况下，我们取样4个可能重叠的目标块掩码，其随机比例范围为（015, 0.2），长宽比范围为（0.75, 1.5）。我们取样1个上下文块掩码，其随机比例范围为（0.85，1.0），长宽比为单位。随后，我们消除上下文块遮罩中与4个目标块遮罩中的任何一个重叠的区域。上下文区块掩码和目标区块掩码对迷你批次中的每张图像都是独立采样的。为了确保高效的批量处理，我们限制了同处一地的GPU上的所有上下文掩码的大小是相同的。同样地，我们也限制在同一地点的GPU上的所有目标掩码的大小是相同的。在PyTorch[56]中，仅用几行代码就可以有效地实现掩码取样器，该函数在数据加载器进程中运行。简而言之，在每次迭代中，数据加载器都会为每张图像返回一个小型批次的图像以及一组上下文和目标掩码，确定上下文和目标视图所需的补丁索引。

A.2. 下游任务

线性评估。当评估iBOT[79]、DINO[18]或MAE[36]等方法时，这些方法利用Vision Trans- formers[29]的额外`[cls]`标记，我们使用VISSL[34]的默认配置来评估iNaturalist18[65]、CIFAR100[45]、Clevr/Count[42, 75]、Clevr/Dist[42, 75]和Places205[78]上所有的模型。我们冻结编码器，并在以下表示法中返回最佳数字：1）最后一层的`[cls]`标记表示，2）最后4层的`[cls]`标记的连接。对于每个表示法，我们尝试两种不同的头：1）一个线性头，或者2）一个线性头，前面有一个批量规范化，并返回最佳数字。我们使用VISSL[34]的默认数据增强：随机调整大小裁剪和水平翻转，但Clevr/Count和Clevr/Dist除外，我们只使用中心裁剪和水平翻转，因为随机裁剪会干扰计数物体和估计距离的能力，从场景中删除关键物体。对于CIFAR100，我们将图像的大小调整为224×224像素，以便保持与预训练时使用的斑块数量相等。

因为我们的I-JEPA实现使用了没有`[cls]`标记的Vision Transformer架构，所以我们调整了默认的VISSL评估配方，以利用平均集合补丁表示而不是`[cls]`标记。因此，我们报告了以下表示法中的最佳线性评估数：1）最后一层的平均集合补丁表示，2）最后4层的平均集合补丁表示的连接。在其他方面，我们保持线性探测配方的一致性

。

ImageNet评估。为了在ImageNet[60]上评估I-JEPA，我们调整了VISSL的配方，以使用平均集合代表，而不是`[cls]`标记。按照MAE[36]，我们使用LARS[72]优化器，批量大小为16384，并训练50个epochs的线性探测。我们使用一个具有阶梯式衰减的学习率，每15个历时除以10的系数，并扫描三种不同的参考学习率[0.01, 0.05, 0.001]，以及两个权重衰减值[0.0005, 0.0]。

低拍评估。为了评估我们的模型在ImageNet-1%低照度任务上的表现，我们采用了MAE[36]的微调协议。我们使用AdamW优化器和余弦学习率调度器在ImageNet-1%上对我们的ViT-L/H模型进行了50次的微调。我们使用512的批次大小，0.75的学习率层衰减和0.1的标签平滑。我们使用MAE中默认的randaugment数据增量。与MAE所做的微调相比，我们不使用mixup、cutmix、
随机擦除或放弃路径。对于I-JEPA，我们对ViT-L/16使用3e /5e^{-5-2} ，对ViT-H/14使用3e /4e^{-5-1} ，对ViT-H/16使用3e /4e$^{-5-1}{}_{448}$ 的学习速率/重量衰减。Semi-VIT在半监督学习的背景下也探索了类似的微调策略，用于低照度学习[16]。

B. 更广泛的相关工作

使用联合嵌入架构的视觉表征的自我监督学习是一个活跃的研究方向[3, 10, 12, 18, 23, 24, 35, 37, 54, 69, 79]。这些方法训练一对编码器为同一图像的两个或多个视图输出相似的嵌入。为了避免病态解，许多流行的联合嵌入方法使用明确的正则化[5, 10, 18, 20]或架构约束[24, 35]。基于架构约束的崩溃预防利用特定的网络设计选择来避免崩溃，例如，在联合嵌入的一个分支中停止梯度流[20]，在联合嵌入的一个分支中使用动量编码器[35]，或使用非对称预测头[8, 20, 35]。最近的工作[62]试图从理论上理解（在某些简化的设置中）具有架构约束的联合嵌入方法如何在没有明确的正则化的情况下避免表示崩溃。

在联合嵌入结构中，典型的基于正则化的防止塌陷的方法试图使表征所占的空间体积最大化。这通常是通过InfoMax[52]原则来激励的。事实上，在无监督的表征学习中，一个长期存在的信念是，所产生的表征应该既能最大限度地了解输入信息，又能满足某些简单性约束[33,50]。前一个目标通常被称为信息最大化原则（InfoMax），而后者有时被称为简约原则[52]。这种表征学习的方法已经被提出了几十年（例如，[14]），在历史上，简单性约束是通过鼓励所学的表征是稀疏的、低维的或分离的，也就是说，表征向量的各个维度应该是统计学上独立的[33]。现代方法通过自我监督的损失项强制执行简单性约束和InfoMax正则化[6, 40, 41, 44, 55, 64]。一个例子是广泛的视图不变性惩罚[53]，通常与独立性[10,74]或低维度约束相结合，例如，通过在单位超球上投影表示[20, 35, 37]。然而，尽管InfoMax原则已被广泛使用，但也有许多批评，特别是它不能区分不同类型的信息（例如，噪音和语义）[2]。事实上，我们希望模型捕捉的特征集并不总是具有最高的边际熵（最大的信息含量）。

与基于不变性的预训练的贡献正交，另一项工作试图通过人为地掩盖部分输入并训练网络重建隐藏内容来学习表征[67]。自回归模型，特别是去噪自动编码器，预测来自噪声视图的干净视觉输入[8, 9, 19, 36, 67]。通常，目标是在像素级[29, 36, 70]，或在补丁标记级，使用标记器[9, 68]预测缺失的输入。虽然这些工作表现出令人印象深刻的可扩展性，但与联合嵌入方法相比，它们通常在低层次的语义抽象中学习特征[4]。

最近，有一组方法试图将联合嵌入架构和基于重建的方法结合起来[30]，其中他们将不变性预训练损失与补丁级重建损失结合起来，如iBOT方法[79]中。由于基于视图不变性的方法通常偏向于学习全局图像表征，从而限制了它们对其他计算机视觉任务的适用性，所以我们的想法是，增加局部损失项可以提高计算机视觉中其他流行任务的性能[11, 26, 32]。对比预测编码的框架[55]也与这个关于局部损失项的工作思路密切相关。在图像的背景下[39]，这里的想法是使用一个对比性目标与卷积网络相结合来区分重叠的图像斑块表示。具体来说，目标是鼓励一个图像斑块的表征对其正下方的图像斑块进行预测，同时推开其他斑块的表征。与这项工作相比，所提出的I-JEPA方法是非对比性的，不寻求对图像斑块进行区分。相反，我们的目标是要从一个单一的背景块中预测各种目

标块的表征。这是通过联合嵌入预测结构实现的，使用的预测器网络是以目标块在图像中的位置对应的位置嵌入为条件的。第8节中的定性实验表明，我们架构中的预测器网络学会了正确地执行这种局部到局部的区域特征映射，并学会了正确地捕捉图像中的位置不确定性。

目标		背景情况	
规模	频率。	规模	**顶部-1**
(0.075, 0.2)	4	(0.85, 1.0)	19.2
(0.1, 0.2)	4	(0.85, 1.0)	39.2
(0.125, 0.2)	4	(0.85, 1.0)	42.4
(0.15, 0.2)	4	(0.85, 1.0)	**54.2**
(0.2, 0.25)	4	(0.85, 1.0)	38.9
(0.2, 0.3)	4	(0.85, 1.0)	33.6

表8.**多区块遮蔽的目标区块大小的消减**。在1%的ImageNet-1K上进行线性评估（仅使用1%的可用标签）；在ViT-B/16的I-JEPA预训练中消减多块目标大小，持续300个历时。只要上下文有足够的信息量，预测更大的（语义）块就能提高低照度的准确性。

目标		背景情况	
规模	频率。	规模	**顶部-1**
(0.15, 0.2)	4	(0.40, 1.0)	31.2
(0.15, 0.2)	4	(0.65, 1.0)	47.1
(0.15, 0.2)	4	(0.75, 1.0)	49.3
(0.15, 0.2)	4	(0.85, 1.0)	**54.2**

表9.**消减多区块遮蔽的上下文大小**。在1%的ImageNet-1K上进行线性评估（只使用1%的可用标签）；在ViT-B/16的I-JEPA预训练中消减多区块目标大小300个历时。减少多区块上下文大小会降低低照度性能。

目标		背景情况	
规模	频率。	规模	**顶部-1**
(0.15, 0.2)	1	(0.85, 1.0)	9.0
(0.15, 0.2)	2	(0.85, 1.0)	22.0
(0.15, 0.2)	3	(0.85, 1.0)	48.5
(0.15, 0.2)	4	(0.85, 1.0)	**54.2**

表10.**消减多区块遮蔽的目标数**。对1%的ImageNet-1K进行线性评估（仅使用1%的可用标签）；在ViT-B/16的I-JEPA预训练中，消减多块目标的数量为300次。增加目标块的数量提高了低照度的准确性。

C. 额外的消融

本节遵循与第9节相同的实验方案。我们报告了一个具有冷冻骨架的线性探针的结果，它是在低照度的1% ImageNet-1K基准上训练出来的。

多区块掩蔽策略。我们提出了一个扩展的多块屏蔽策略，我们改变了目标块的比例（表8）、背景比例（表9）和目标块的数量（表10）。我们使用I-JEPA在各种多块设置下对ViT-B/16进行了300个历时的训练，并使用线性探针对1%的ImageNet-1K基准进行了性能比较。简而言之，我们发现预测几个相对较大的（语义）目标块，并使用信息量足够大的（空间分布）背景块是很重要的。

在目标编码器的输出端进行屏蔽。I-JEPA中一个重要的设计选择是，目标块是通过屏蔽目标编码器的*输出*而不是

输入来获得的。表11显示了在使用I-JEPA对ViT-H/16进行预训练300次时，这一设计选择对所学表征的语义水平的影响。在对输入进行屏蔽的情况下，我们对每个目标区域通过目标编码器向前传播一次。在预训练期间对目标编码器的输出进行屏蔽，会产生更多的语义预测目标，并提高线性探测性能。

目标屏蔽	拱门。	纪元	顶-1
输出	ViT-H/16	300	**67.3**
输入	ViT-H/16	300	56.1

表11.**消除目标编码器的屏蔽输出**。在ImageNet-1K上仅使用1%的可用标签进行线性评估；在对ViT-H/16进行I-JEPA预训练的过程中，消除了对目标编码器输出的屏蔽效果，时间为300个历时。在预训练过程中屏蔽目标编码器的输出，大大改善了预训练表征的线性探测性能。

预测器深度。我们在表12中考察了预测器深度对下游低射性能的影响。我们使用6层预测器网络或12层预测器网络对ViT-L/16进行500个历时的预训练。与使用较浅的预测器预训练的模型相比，使用较深的预测器预训练的模型在下游低射性能方面有明显的改善。

预测器深度	拱门。	纪元	顶-1
6	ViT-L/16	500	64.0
12	ViT-L/16	500	**66.9**

表12.**消减预测器的深度**。在ImageNet-1K上仅使用1%的可用标签进行线性评估；消减ViT-L/16预训练500次的预测器深度的影响。增加预测器深度会使预训练表征的线性探测性能得到明显改善。

权重衰减。在表13中，我们评估了预训练期间权重衰减的影响。我们探索了两种权重衰减策略：将权重衰减从0.04线性增加到0.4，或者使用0.05的固定权重衰减。在预训练过程中使用较小的权重衰减，在微调时能提高ImageNet的下游性能-1%。然而，这也导致了线性评估中的性能下降。在本文中，我们使用第一种权重衰减策略，因为它提高了线性评估下游任务的性能。

重量衰减	拱门。	纪元	ImageNet-1%	图像网线性评价
0.04 → 0.4	ViT-L/16	600	69.4	**77.8**
0.05	ViT-L/16	600	**70.7**	76.4

表13.**消除训练前的权重衰减**。我们比较了我们默认的训练前权重衰减策略，我们将权重衰减从0.04线性增加到0.4，与使用0.05的固定权重衰减。在预训练期间使用较小的权重衰减可以提高
在ImageNet上的微调性能-1%，然而，这也导致了线性评估中的性能下降。

预测器宽度。我们在表14中探讨了预测器宽度的影响。我们将使用ViT-L编码器和386通道的预测器的I-JEPA与使用1024通道的预测器的类似模型进行比较。请注意，ViT-L编码器有1024个通道。使用预测器宽度的瓶颈可以提高ImageNet上的下游性能1%。

预测器宽度	拱门。	纪元	顶-1
384	ViT-L/16	600	**70.7**
1024	ViT-L/16	600	68.4

表14.**消减预测器的宽度**。我们报告了在ImageNet-1K 1%上使用微调的结果。我们比较了两个宽度为384或1024的预测器。注意I-JEPA编码器是一个具有1024个通道的ViT-L。在预测器中设置宽度瓶颈可以改善
下游性能。

D. 在完整的ImageNet上进行微调

在本节中，我们报告了I-JEPA在整个ImageNet数据集上进行微调时的性能。我们专注于ViT- $H/16_{448}$ ，因为这个架构在MAE方面取得了最先进的性能[36]。

我们使用类似于MAE的微调协议。具体来说，我们使用AdamW和余弦学习率计划对我们的模型进行50次微调。基础学习率被设置为10^{-4}，批次大小为528。我们使用mixup[76]设置为0.8，cutmix[73]设置为1.0，下降路径概率为0.25，权重衰减设置为0.04来训练。我们还使用了一个层衰减为0.75.最后，我们使用了与MAE相同的rand-augment数据-augmentations、

表15报告了微调的结果。I-JEPA达到了87.1的最高精确度。尽管 I-JEPA 训练的历时比 MAE 少 5.3 倍，但其性能与最佳 MAE 模型相差不到 1%。这一结果表明，在对整个ImageNet数据集进行微调时，I-JEPA具有竞争力。

方法	拱门。	历时	顶点-1
MAE [36]	ViT-H/14448	1600	**87.8**
I-JEPA	ViT-H/16448	300	87.1

表15.**在完整的 ImageNet 数据集上进行微调**。I-JEPA 实现了有竞争力的性能。尽管 I-JEPA 的训练时间比 MAE 少 5.3 倍，但 I-JEPA 接近 MAE 方法。

E. RCDM可视化

为了使预训练的神经网络在像素空间的表征可视化，我们使用RCDM框架[13]。RCDM框架训练一个解码器网络h_ω，包括一个生成扩散模型，从图像的表示向量$\boldsymbol{s}_x$和该图像的噪声版本$\tilde{x} = \boldsymbol{x} + \epsilon$，其中$\epsilon$是一个加性噪声向量，重建一个图像$\boldsymbol{x}$。具体来说，解码器的目标是最小化损失函数$\|h_\omega(\tilde{\boldsymbol{x}}, \boldsymbol{s}_x) - \epsilon\|$。我们使用默认的超参数[13]对每个RCDM网络进行300,000次迭代训练。在训练完解码器后，随后可以将未见过的测试图像$\boldsymbol{s}_y$的表征向量与各种随机噪声向量一起送入解码器，以生成几个像素级的可视化表征，从而深入了解预训练网络的表征中所捕获的特征。各个样本的共同特征代表了表征中包含的信息。另一方面，不同样本之间的品质代表了不包含在表征中的信息。

在图6中，通过将预测器的平均集合输出（以特定的目标区域为条件）与各种随机噪声向量一起送入解码器网络，获得了可视化效果。在图7和图8中，通过将目标编码器的平均集合输出与各种随机噪声向量一起输入到解码器网络中，可以得到可视化的结果。

E.1. 编码器的可视化

在图7中，我们将ViT-H/14目标编码器输出的平均集合I-JEPA表示可视化。第一列包含原始图像，而随后的几列包含合成样本，这些样本是通过将图像的平均集合表示与各种随机噪声向量一起送入解码器得到的。图7表明，I-JEPA目标编码器能够正确捕捉有关物体及其姿势的高层次信息，而放弃低层次的图像细节和背景信息。

图8显示了类似的可视化，但是当使用MSN[4]预训练的ViT-L/7目标编码器来计算图像表示时。MSN方法使用联合嵌入架构来训练上下文和目标编码器，以强制执行全局图像表征对各种手工制作的数据增强和缺失斑块的不变性。虽然MSN预训练的网络能够捕捉到第一列图像的高层次语义信息，但它在生成的样本中也表现出较高的变

异性，例如物体姿势、物体比例和实例数量的变异。简而言之，MSN预训练抛弃了图像中的大部分局部结构，这与I-JEPA形成鲜明对比，后者保留了输入图像中大部分局部结构的信息。

图7.**I-JEPA目标编码器的可视化表示**。对于每幅图像：第一列包含原始图像；随后各列包含生成模型解码预训练的 I-JEPA 目标编码器的平均输出的样本。各个样本的共同特征代表了I-JEPA表示中包含的信息。I-JEPA能够正确捕捉关于物体和它们的姿势的高级信息。不同样本的特征代表不包含在表示中的信息。I-JEPA 编码器放弃了精确的低层次细节以及背景信息。

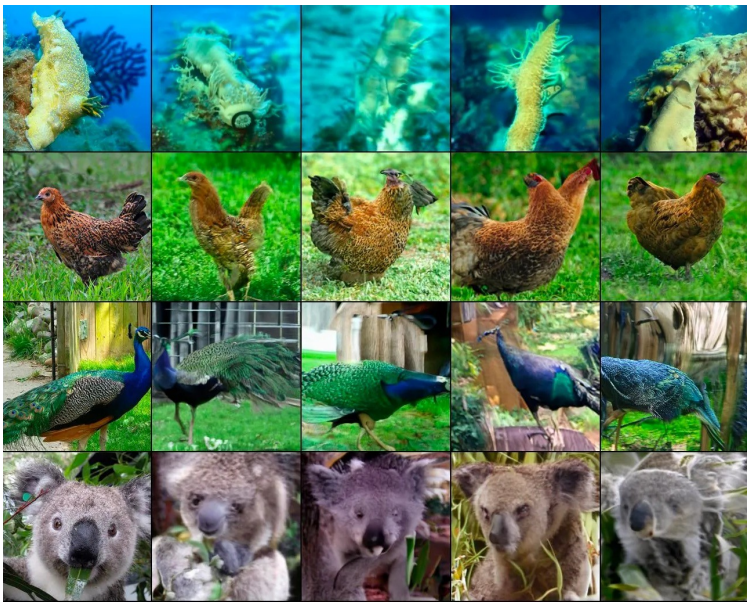


图8.**MSN目标-编码器的可视化表示**。对于每幅图像：第一列包含原始图像；随后的几列包含生成模型对冻结的MSN编码器的输出进行解码的样本[4]。各个样本的共性代表了表示中包含的信息。不同样本之间的特性代表了MSN没有捕捉到的信息。与I-JEPA相比，MSN样本显示出更高的可变性。MSN从输入中保留的信息较少。特别是，它抛弃了全局结构信息，如物体的姿势或甚至实例的数量。