arXiv:2304.00218v1 [cs.CV] 1 Apr 2023

用于自监督学习的掩码层次特征

刘丰港
商汤集团
liufenggang@senseauto.com

Yangguang Li
商汤集团
liyangguang@sensetime.com

Feng Liang
德克萨斯大学奥斯汀分校
jeffliang@utexas.edu

徐吉岚
复旦大学
18210240039@fudan.edu.cn

Bin Huang
商汤集团
huangbin@senseauto.com

Jing Shao
商汤集团
shaojing@senseauto.com

摘要

本文表明，对层次化深度特征进行掩码是一种高效的自监督方法，称为***MaskDeep***。*MaskDeep* 将表征空间中的每个图像块视为一个独立实例。我们对表征空间中的部分图像块进行掩码，然后利用稀疏可见图像块重建高语义图像表征。*MaskDeep* 的直觉在于，模型能够从稀疏可见图像块的语义推理出图像的全局语义。我们进一步在框架中提出了三项设计：*1*）分层深度掩码模块，用于关注图像块表示的层次属性；*2*）多组策略，在不增加编码器额外计算开销的情况下提升效率；*3*）多目标策略，以提供对全局语义的更多描述。我们的 *MaskDeep* 带来了显著改进。在 *ResNet50* 上训练 *200* 个训练轮次后，*MaskDeep* 在 *ImageNet* 上的线性分类任务中取得了 *71.2%* 的 *Top-1*准确率，达到了当前最优结果。在 *COCO* 目标检测任务中，*MaskDeep* 超越了专为目标检测设计的自监督方法 *SoCo*。在训练 *100* 个训练轮次时，*MaskDeep* 达到 *69.6%* 的 *Top-1*准确率，超过了使用 *200* 个训练轮次训练的当前方法（如 *HCSC*）*0.4%*。

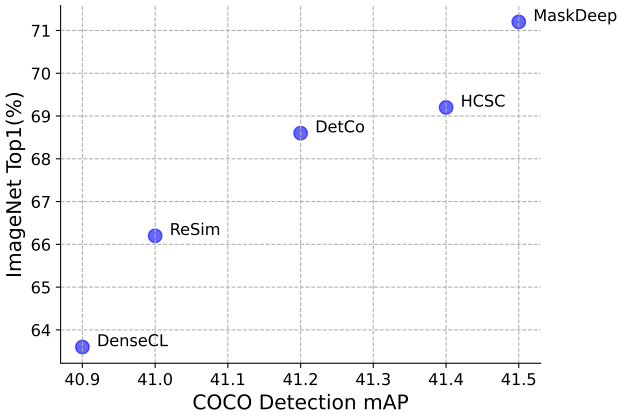


图1：在 ImageNet-1k 分类和 COCO 检测基准上与不同对比学习方法进行比较。所有方法均在 ImageNet-1k 数据集上训练了 200 个训练轮次。

1. 引言

在计算机视觉领域，自监督学习凭借其卓越的表征学习能力，吸引了前所未有的关注。对比学习和掩码图像建模（MIM）是两种最有效的方法。

对比学习方法首先被用于学习不变的图像级全局表征。具体而言，网络通常经过训练，以最大化同一图像的两种数据增强视图之间的相似度，例如 MoCo v1/v2 [13, 6], BYOL [10], 和 SimSiam [7]。这些图像级方法能够学习具有区分性的视觉表征，因此在分类任务上取得了较高的性能。其他对比学习方法 [25, 30, 26] 则尝试通过对齐不同数据增强图像视图中相同位置的局部表征来建模局部信息。例如，SoCo [26] 首先通过选择性搜索生成目标提议 [21] ，然后对齐不同视图的预提议目标特征。这些目标级方法在密集预测任务上取得了较高的性能，但在分类任务上表现出明显的劣势。

近年来，掩码图像建模（MIM）[12, 2, 29, 31,1] 方法在学术界引起了极大关注。MIM 的核心思想是利用稀疏的未掩码图像块重建经过掩码处理的源图像。以 MAE [12] 为例，它会对源图像中的图块进行掩码，将

掩码图像输入编码器模型，然后利用可见图块特征，通过解码器模型预测掩蔽像素。其他方法[2, 29] 则关注预测低级像素时产生的高频噪声。因此，它们不直接预测低级像素，而是预测由分词器嵌入的图块的高级语义。这些掩蔽图像建模方法为视觉变换器模型的预训练带来了突破。我们认为，掩蔽图像建模方法的部分成功归因于对稀疏图块之间关系的探究。这表明，从稀疏的局部图像块中推理不可见的上下文是一项高效的前置任务。

本文遵循这样一种思路：通过推理稀疏局部图像块之间的关系来学习知识。与众不同的是，为避免消耗数千个训练轮次，我们从精心设计的特征空间中多次提取稀疏局部图像块的嵌入，而不是像掩蔽图像建模方法那样反复输入掩码图像。考虑到视觉表征的层次属性，我们引入一种 FPN型模块，使图块嵌入具有层次结构。我们假设，空间特征中的每个点特征都代表源图像中特定区域的高级语义。由此，我们可以直接从隐式表示空间中构建稀疏可见图块嵌入组。我们将这种方法称为 MaskDeep。

为了从稀疏局部图像块组中学习知识，我们从稀疏图块嵌入出发，推理并预-测图像的全局语义。其直觉在于，局部对象包含细粒度的目标语义，而完整图像则展现了覆盖所有空间位置的高级语义信息。整合局部目标特征以表征完整图像的语义，可以成为一种有效的自监督目标。如图 2 所示，稀疏图块特征（包括“人头”“狗腿”和“树”等）被整合起来，用于预测图像内容：“一个人在公园里遛狗。”如此一来，局部目标理解和全局上下文信息推理都得到了兼顾。这一特性使 MaskDeep 成为统一的自监督学习器，适用于图像级和目标级下游任务。具体而言，在 MaskDeep 中，在线分支用于构建图块特征，动量分支用于提取完整图像的信息。仅使用约十个稀疏图块来预测完整图像的信息。预测通过最大化联合图像块表示与全局表示之间的余弦相似度进行监督。与先前的对比学习方法不同， [10],MaskDeep 在动量编码器之后无需投影器。

在分类和密集预测下游任务上的大量实验表明，我们的 MaskDeep 具有良好的有效性和通用性。如图1所示，在 ResNet50 上训练 200 个训练轮次后，MaskDeep 达到了 71.2 % 在 ImageNet 上的 Top1准确率线性分类，以及在 COCO 目标检测任务上 41.5% 的边界框平均精度均值，超过了当前的自监督方法。在表7中，MaskDeep 的表现优于专为目标检测设计的 SoCo，平均精度均值高出 0.1。

我们的贡献可概括为以下三点。

- 我们提出了一种新颖的自监督学习方法 MaskDeep，通过在精心设计的隐式表示空间中屏蔽视图来进行学习。

- 我们直接对齐联合图像块表示与全局表示，无需投影器。同时，该方法兼顾局部目标理解和全局上下文信息推理。

- MaskDeep 在分类和密集预测任务上均取得了最先进性能：在训练 200 个训练轮次后，分类 Top1 准确率为 71.2%，目标检测边界框平均精度为 41.5%，实例分割掩码平均精度为 36.1%。

2. 相关工作

2.1. 对比学习

基于对比的自监督学习方法 [13, 6, 10, 5, 27, 28, 26, 25, 30, 24, 16, 23]旨在避免对大量标注的依赖。这些方法的总体目标是通过区分正图像对和负图像对来学习有用的图像语义。MoCo [13]提出维护一个负样本记忆队列，用于计算对比损失，从而减轻了训练过程中对大批量大小的需求。另一类方法只使用正样本对。例如，BYOL [10] 提出了一种非对称架构：在在线分支后附加一个预测器，并通过动量策略更新目标分支。

除图像级自监督研究外，目标级自监督研究 [28, 26, 25, 30] 在目标级学习图像表征，主要针对目标检测和语义分割等下游任务。举几个例子，InsLoc [30] 通过将前景实例复制并粘贴到不同的背景图像中，在实例级对比损失中构造正样本对。为了在分类任务和实例级密集预测任务上都保持竞争力，DetCo [28]将 MoCo [13] 推广为实例级监督。DetCo 通过聚合非重叠图像块的特征，获得联合局部表示。这些工作通过实验验证了在自监督预训练中引入检测特性的有效性。

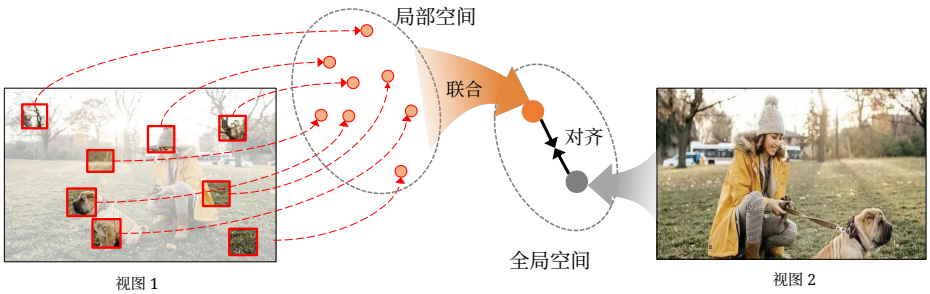


图 2：MaskDeep 的监督方式。局部图像块嵌入局部特征空间，而全局图像嵌入全局特征空间。随后，整合局部特征，使其与全局特征对齐。

2.2. 掩码图像建模

掩码图像建模在面向 Transformer 模型的自监督学习领域取得了出色表现。BEiT [2]对图像进行分块掩码，并预测掩码图像块对应的离散视觉标记；这些视觉标记的目标直接取自预训练的视觉标记器。MAE [12]利用自编码器架构重建掩码图像块的原始归一化 RGB 像素，无需将掩码标记输入编码器。Data2Vec [1] 和 IBOT [31] 证明了将掩码图像块的特征与完整图像的特征对齐是有效的。与这些方法不同，我们的 MaskDeep 策略对深层特征而非输入图像进行掩码，目标是将多个深层稀疏特征与完整图像特征对齐。

3. 方法

3.1. 概述

我们提出的自监督学习方法的整体框架如图 3 所示。MaskDeep 遵循对比学习的主要范式。我们使用一对编码器和动量编码器来处理两个不同的视图。这两个视图经过 MoCov2[6] 的数据增强方法预处理。MaskDeep 着重探索局部语义与全局语义之间的关系，并降低构建局部语义处理过程的计算开销。我们从隐式表示空间构建局部语义，并整合这些局部语义以表达完整图像的语义。MaskDeep 的核心设计包含三个部分：层次化深度掩码模块、预测器与目标，以下章节将详细介绍这些内容。

3.2. 层次化深度掩码模块

分层深度掩码模块生成一个层次化表征空间，使图块嵌入具备层次属性。它基于特征金字塔网络模块[17]构建。对于 ResNet，给定一幅图像，编码器可从不同阶段生成图像特征，例如 $\{C_2, C_3, C_4, C_5\}$，其中，特征 C_i 对应于 i_{th} 阶段。我们将图像特征 C_i 输入一个 FPN型模块。FPN型模块的输出，例如$\{P_2, P_3, P_4, P_5\}$，被构建为层次化表征空间。通过这种方式，同时考虑低层图像特征（例如， C_2）和高层图像特征（例如， C_5），以形成多尺度表示空间。

随后，在空间维度上，对层次化深度掩码模块的输出 $\{P_2, P_3, P_4, P_5\}$ 进行点级随机掩码。在层次化深度掩码模块每个尺度的输出中，我们保留相同数量 k 的点特征，以确保每个尺度受到同等关注。从保留下来的 $4 \times k$ 未掩码点特征中，我们在不同尺度级别随机采样 k 个点特征，作为最终的图块特征组。具体而言，从层次化深度掩码模块的每个尺度中以相同概率选取一个点。这样，图块特征组便能保留视觉表征的层次化特性。

此外，我们提出了一种专用于本方法的多组策略。得益于深度掩码模块，图块特征组可以重复采样多次，而无需额外消耗编码器的计算资源。每个图块特征组都会被独立用于预测图像的全局语义。多组策略变相增大了批量大小，同时消耗更少的计算资源。

3.3. 预测器

预测器通过两个简单的线性层聚合图块特征组。第一个线性层后接批量归一化[15], 和修正线性单元（ReLU）[20]。该预测器的架构与其他对比学习方法中的预测器相同。将一组图块特征表示为 $\mathcal{H} = \{h_i\}_{i=1}^{k}$，其中 $h_i \in \mathrm{R}^d$是一个空间位置处的图块特征， k 是每组中的图块特征数量

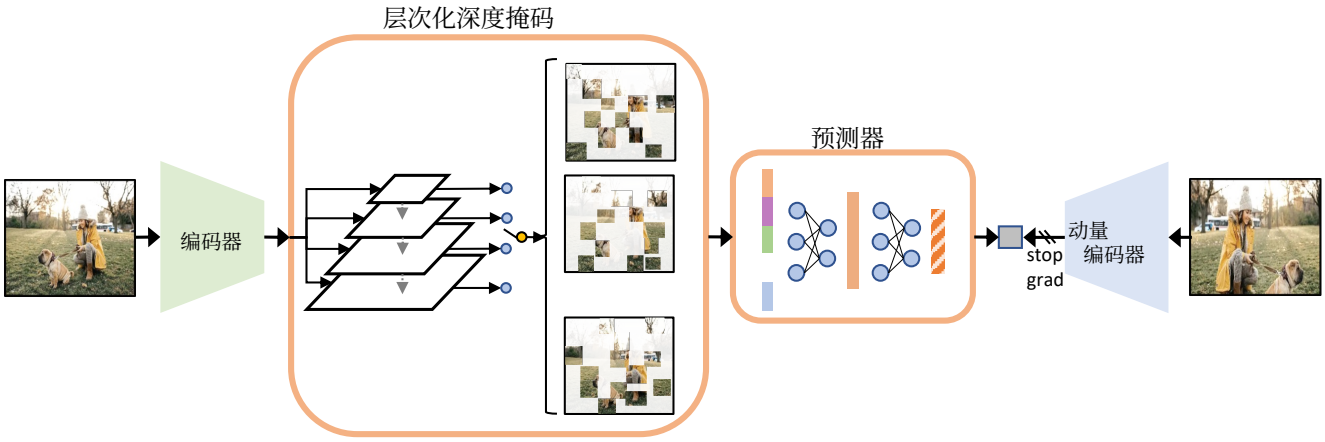


图 3：MaskDeep 的框架。MaskDeep 使用来自层次化表示空间的稀疏图像块特征，对齐完整图像的全局语义。将图像的一个视图输入编码器后，我们通过层次化深度掩码模块生成层次化表示空间。我们从表示空间中采样多组图像块。随后，每组图像块都由一个简单的线性预测器解码。完整图像的全局语义由动量编码器根据图像的另一个视图生成。

。在 MaskDeep 中，我们采样K 组图块特征，记为 $\{\mathcal{H}_i\}_{i=1}^{K}$。每组图块特征都会拼接后输入预测器。预测器 q 将图像块表示组 $\mathcal{H}$ 聚合为联合特征 $q\left(\mathcal{H}_{i=1}^{K}\right)$，该特征能够描述图像的主要内容。

3.4. 目标

目标是动量编码器f_ϕ最后一层的直接输出，不包含投影器等任何额外的参数化模块。对于 ResNet，目标经过全局池化后成为一个向量。所有目标都从计算图中分离。目标记为 $\mathcal{G}$，预期用于表示图像的全局语义。动量编码器的参数根据在线编码器的参数通过指数移动平均（EMA）进行更新。EMA 对防止全局分支发生模型坍塌至关重要 [13, 10]。EMA 的计算方式如下：

$$\phi := \lambda\phi + (1-\lambda)\theta, \tag{1}$$

其中，λ 是动量系数，θ 是在线编码器的参数，ϕ 是动量编码器的参数。训练过程中，我们将该系数从 0.99 逐步增大到 1。

此外，我们提出了一种多目标策略。与多裁剪策略不同，后者仅将多个视图输入在线编码器；多目标策略则只将额外视图输入动量编码器。动量编码器采用停止梯度策略，因此不会创建计算图。内存使用量的增幅很小。在我们的多目标策略中，联合图块表示需要预测多个目标。多目标可表示为 $\{\mathcal{G}_j\}_{j=1}^{N}$，其中$N$ 是目标的数量。多目标策略提供了更多关于全局语义的描述，从而能够维持训练稳定性。

3.5. 损失函数

损失函数就是余弦相似度损失，定义为 $<\cdot,\cdot>$。该损失函数对齐联合图块特征表示 $q(\mathcal{H})$与全局表示$\mathcal{G}$。通过多组策略，我们可以得到 K 联合图块特征。通过多目标策略，我们可以得到 N 全局特征。每一对联合图块特征和全局特征都会进行对齐，如公式 2 所述

$$\mathcal{L} = -4 \cdot \sum_{i=1}^{K}\sum_{j=1}^{N}\left(< p(\mathcal{H}_i^1), \mathcal{G}_j^2 > + < p(\mathcal{H}_i^2), \mathcal{G}_j^1 >\right), \tag{2}$$

其中，第一项表示视图 1 的掩码表示与视图 2 的全局表示之间的对齐。第二项交换了在线分支和动量分支所使用的视图。

4. 实验

为评估 MaskDeep 的通用性，我们在分类和密集预测任务上对其进行评估，包括 ImageNet 线性分类[8], PASCAL VOC

表1：不同下游任务的表现，包括线性分类（CLS）、目标检测（DET）和实例分割（SEG）。

方法	训练轮次	CLS		VOC目标检测			COCO目标检测与实例分割 2x					
		Top1	Top5	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
监督学习	90	76.5	-	53.5	81.3	58.8	40.0	59.9	43.1	34.7	56.5	36.9
MOCOv2	200	67.5	-	57.0	82.4	63.6	38.9	58.4	42.0	34.2	55.2	36.5
BYOL	100	66.5	-	-	-	-	-	-	-	-	-	-
BYOL	300	72.5	-	51.9	81.0	56.5	40.3	60.5	43.9	35.1	56.8	37.3
DenseCL	200	63.6	85.9	58.6	83.0	65.6	40.9	60.7	44.1	35.6	57.3	38.1
ReSim-C4	200	66.2	86.9	**58.6**	83.0	65.6	41.0	60.7	44.5	35.7	57.6	38.0
DetCo	200	68.6	88.5	57.8	82.6	64.2	41.2	61.0	44.7	35.8	57.8	38.2
SoCo-C4	100	59.7	82.8	58.5	**83.4**	**65.7**	41.1	61.0	44.4	35.6	57.5	38.0
HCSC	200	69.2	-	-	82.5	-	41.4	-	-	-	-	-
MaskDeep	100	69.6	88.6	57.9	82.9	64.8	41.2	61.1	44.7	35.8	57.7	38.2
MaskDeep	200	**71.2**	**89.5**	58.0	82.8	64.7	**41.5**	**61.1**	**45.3**	**36.1**	**58.0**	**38.8**

目标检测[9], COCO 目标检测和实例分割[18]。

4.1. 实现细节

我们的 MaskDeep 在 ImageNet1k 数据集上进行预训练，该数据集包含 128 万张图像。数据增强流程遵循 MoCov2。我们使用 SGD优化器，并采用余弦衰减学习率调度[19] 以及 10 个训练轮次的预热期。学习率设为 0.02。当训练 100 个训练轮次时，批量大小为 1024。权重衰减设为 10^{-5}。余弦衰减动量系数 λ 从 0.99 开始，并逐渐增加至 1。当训练 200 个训练轮次时，我们将批量大小增加至 2048，并将学习率加倍至 0.04。

我们采用标准的 ResNet-50 作为骨干网络，并启用同步批归一化[15]。在分层深度掩码模块中，默认使用阶段 $\{P_3, P_4, P_5\}$ 的输出。分层深度掩码模块的输出维度为 512，与每个图像块表示的维度相同。对于多组策略，我们采样 16 组图块特征。对于多目标策略，我们额外使用 4 个全局目标。因此，在公式 2 中，K 为 16，N 为 4。默认情况下，我们在每组中采样 12 个图块特征。预测器的输入维度等于一组图块特征维度之和。预测器的隐藏维度为 4096。预测器的输出维度与目标的输出维度相同，均为 2048。

4.2. 结果

我们在一系列下游任务上，将监督学习和自监督学习方法与 MaskDeep 进行了比较。

ImageNet线性分类。 遵循MoCo [13], , 我们冻结预训练骨干网络，并训练一个线性分类头，训练轮次为100轮。初始学习率为30，并在第60和80个训练轮次时除以10。训练图像经过随机裁剪并调整大小至 224×224，然后以0.5的概率进行水平翻转。如表7所示，在训练200个训练轮次时，MaskDeep达到71.2%的Top-1 准确率和89.5%的Top-5 准确率，显著优于所有自监督学习方法。在训练100个训练轮次时，MaskDeep达到69.6%的Top-1 准确率，比训练200个训练轮次的HCSC高出0.4%。这表明MaskDeep是一种更加高效的表征学习方法。

PASCAL VOC 目标检测。 我们采用 PASCAL VOC trainval07+12 进行微调，并报告 test07 集上的 AP^{bb}、AP^{bb}_{50} 和 AP^{bb}_{75} 结果。遵循[25, 28], 中的协议，我们使用 C4 骨干网络对 Faster R-CNN 目标检测器进行微调。如表7所示，MaskDeep 略逊于专为目标检测任务设计的最先进方法。MaskDeep 展现出了相当的结果。

COCO 目标检测和实例分割。 在迁移到 COCO 2017 数据集上的目标检测和实例分割任务时，我们将 MaskDeep 与最先进的方法进行了比较。我们将预训练的骨干网络与 Mask R-CNN 检测器（C4 骨干网络）结合使用[14]。遵循 [28, 26], 中的设置，采用标准的 $2\times$ 训练轮次方案。如表7所示，在 200 个训练轮次下，MaskDeep 达到 41.5% 的边界框平均精度和 36.1% 的掩码平均精度，分别比 SoCo 高出 0.1% 和 0.2%。在训练 200 个轮次时，MaskDeep 的结果也略优于其他目标级自监督学习方法。

总的来说，我们的 MaskDeep 在分类、目标检测和分割任务上都取得了不错的提升。MaskDeep 展现出卓越的综合能力。如图1所示，MaskDeep 超越了当前的

方法，包括图像级和目标级自监督学习方法。

5. 消融实验

我们逐步开展了一系列消融实验，以验证 MaskDeep 中各个组件的有效性。除非另有说明，所有消融实验均采用默认批量大小 1024 和 100 个预训练轮次。下游任务的配置保持不变。我们在 ImageNet 线性探测基准上评估 ResNet。

表 2：ImageNet 线性探测基准上的消融实验结果。

模型	分层模块	多组	多目标	Top1
BYOL				66.5
朴素				64.9
MaskDeep	✓			67.9
MaskDeep	✓	✓		68.8
MaskDeep	✓	✓	✓	**69.6**

由于我们遵循对比学习方法的主要范式，并使用余弦相似度损失函数，因此将 BYOL 视为基线。消融结果如表2所示。第一行表示 BYOL 训练100个训练轮次时的性能。第二行表示一种在表征空间中对图块进行掩码的朴素思路。在这种朴素思路中，直接从编码器的最后一层采样一组点特征。如表2所示，这种朴素思路的 top-1 准确率为64.9%，而在相同的100个训练轮次下，SimCLR 达到64.8%，BYOL 达到66.5%。基线的表现略差，但该方法有效，并能从朴素的前置任务中学习表征。它落后于图像级自监督方法，但优于目标级自监督方法，例如 DenseCL，其准确率仅为 63.6%，如表7所示。

通过层次深度掩码模块关注层次属性后，我们的 MaskDeep 获得了 +3.0% 的提升。其 top-1 准确率达到 67.9%，比 BYOL 高出 1.4%。借助层次深度掩码模块，其性能可与当前的自监督学习方法相媲美。此外，我们引入了专用于本方法的多组策略。直观而言，多组策略增加了样本数量，从而带来了 +0.9% 的提升。最后，我们引入了多目标策略，以提供更多监督信号。该策略使训练过程更加稳定。如表 2 所示，ResNet 的性能提升了 +0.8%。

6. 分析

本节将进一步考察 MaskDeep 中各组件的不同超参数。

层次化深度掩码模块的尺度级别和维度。 表 3 展示了尺度级别数量和维度对层次化深度掩码模块的影响。为了解尺度的影响，我们对层次化深度掩码模块从最小尺度到最大尺度的层数进行了消融实验。

在表 3 (a) 中，级别 1 表示仅使用最后一个阶段的输出 $\{P_5\}$，级别 2 表示使用最后两个阶段的输出 $\{P_4, P_5\}$，级别 3 表示使用各阶段的输出 $\{P_3, P_4, P_5\}$，级别 4 表示使用所有阶段的输出 $\{P_2, P_3, P_4, P_5\}$。结果表明，仅使用最后一个阶段的输出 $\{P_5\}$ 时，性能会下降约 1.4%。这说明层次化深度掩码模块发挥着不可或缺的作用。除仅使用 $\{P_5\}$ 外，我们的方法对其他级别都具有稳健性。结果表明，级别 4 略有负面影响。这可能是由于 $\{P_2\}$ 的分辨率较高，导致每个点的语义信息不足。使用各阶段的输出 $\{P_3, P_4, P_5\}$ 时，我们的方法取得了最佳性能。

表 3：层次化深度掩码模块的尺度级别和维度分析。

（a）尺度级别。

级别	1	2	3	4
Top1	67.4	68.7	**68.8**	68.6

（b）维度。

Dim	256	512	1024
Top1	68.5	**68.8**	68.6

表 3（b）展示了层次化深度掩码模块维度的影响。当 ResNet50 的输出维度为 $[256, 512, 1024, 2048]$, 时，维度为 512 时的表现最佳。我们总结认为，该维度应接近编码器的输出维度。较小的维度会限制每个图块特征的表征能力，而较大的维度会降低训练速度并增加学习难度。

一组中的图块特征数量。我们分析了可见图块特征的数量。不同图块特征数量的结果如表4所示。研究发现，MaskDeep 对从8到20的补丁数量不敏感。当数量为 12时，MaskDeep 达到最佳性能。图块过少无法充分覆盖图像内容，而图块过多会引入重复的图像块表示和噪声，从而增加学习

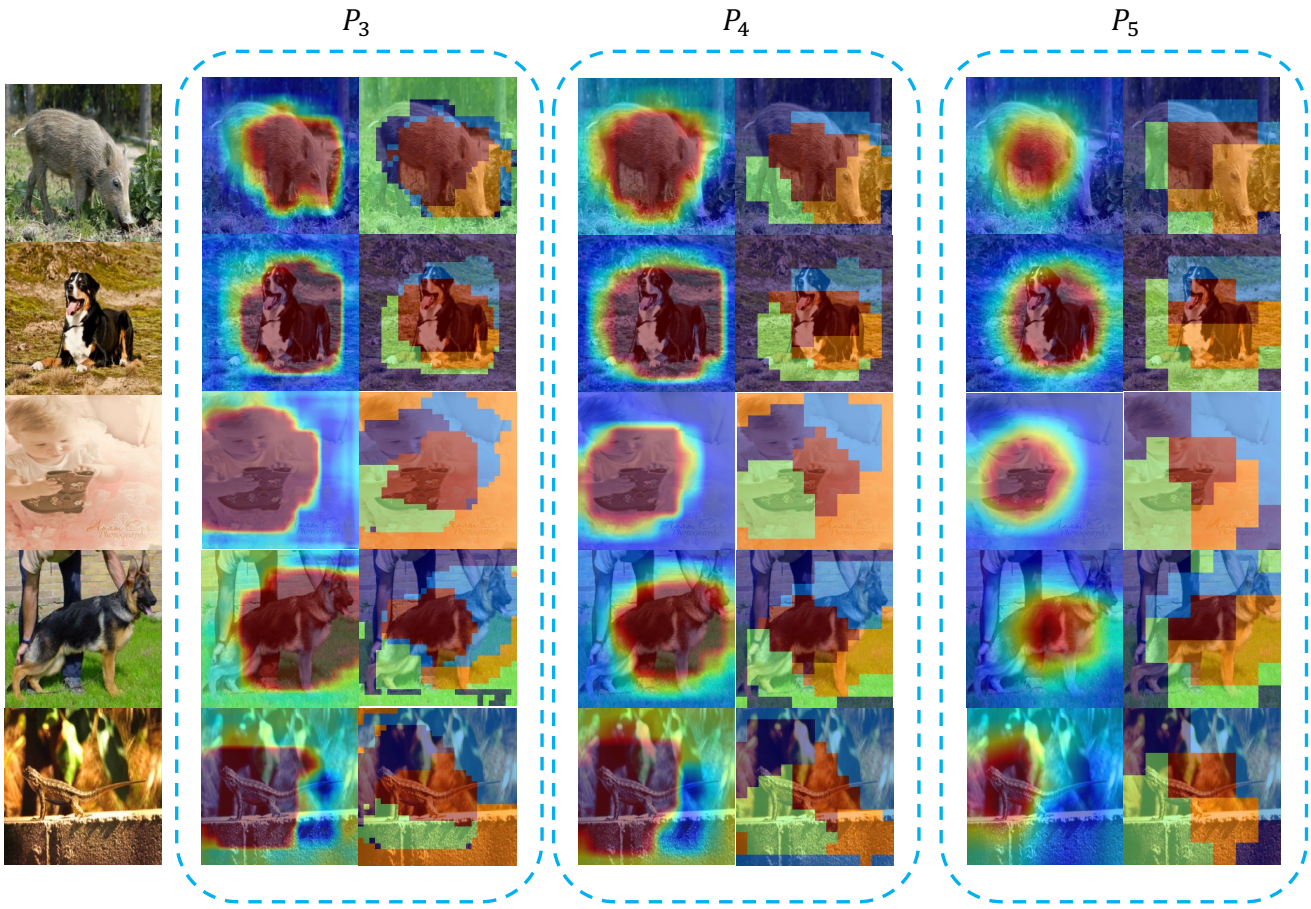


图4：隐式表示空间的可视化。我们绘制了层次化深度掩码模块中每个尺度级别 $\{P_3, P_4, P_5\}$ 的结果。对于每个尺度级别，第一列为 Grad-CAM 结果，第二列为聚类图结果。

难度。

表4：对每组图块特征数量的分析。

图块数量	4	8	12	16	20	24
Top1	68.4	68.7	**68.8**	68.7	68.8	68.6

多组策略中的组数量。表5展示了组数量的影响。当组数量为1时，表示仅采样一组图块，此时未使用多组策略。当组数量从1增加到8时，我们的方法取得了 +0.6%的显著提升。当组数量从8增加到16时，性能提升趋于饱和，仅增加了0.3%。当我们继续将组数量增加到24时，性能不再提升，甚至略有下降。

多目标策略中的目标数量。该

表5：对多组策略中组数量的分析。

组数量	1	8	16	24
Top1	67.9	68.5	**68.8**	68.7

表6显示，目标数量从0增加到4时，性能会有所提升。与不使用多目标策略相比，额外增加4个目标可使性能提升0.8%。但增加更多目标带来的收益很小。我们认为，多目标策略能够提供稳定的监督信号。

预测器的隐藏维度。预测器包含两个线性层。正如以往研究所示，预测器的隐藏维度是一个重要的超参数。在本部分中，我们对表7中的超参数进行了研究。结果表明，当维度为4096、图像块数量为12且图像块

表6：多目标数量中目标数量的分析。

目标数量	0	2	4	6
Top1	68.8	69.0	**69.6**	69.5

维度为512。

表7：预测器隐藏维度的分析。

Dim	2048	4096	8192
Top1	68.5	**68.8**	68.7

采用多裁剪策略。多裁剪策略 [4]是自监督学习方法中的一种有效技巧。以往研究证实，更多视图可以提升自监督学习模型的效果。我们将多裁剪策略应用于MaskDeep。我们采用多尺度多裁剪策略作为HSCS [11]。具体而言，使用三种不同尺度的视图 $\{192, 160, 128\}$ ，每种尺度采样两次。结果如表8所示。采用多裁剪策略后，MaskDeep在训练100个训练轮次时性能提升了2.5%，在训练200个训练轮次时性能提升了2.8%。加入多裁剪策略后，MaskDeep的表现也优于以往的方法。

表8：采用多裁剪策略时的性能。所有方法均采用多裁剪策略进行训练。

方法	训练轮次	Top1
UniVIP	200	73.1
HCSC	200	73.3
VICRegL[3]	300	71.2
MaskDeep	100	73.1
MaskDeep	200	**74.0**

7. 可视化

层次化深度掩码模块生成的隐式表征空间是我们方法的重要组成部分。我们从隐式表征空间中采样所有图块特征。我们假设，每个图块特征都包含与特定图像区域相对应的局部语义，因此可以从若干图块特征推断全局图像内容。因此，我们使用 Grad-CAM [22] 和聚类图对隐式表征空间进行可视化，以验证我们的想法。在可视化过程中，我们仅对视图 1 应用调整大小预处理方法。当应用 Grad-CAM 时，我们仅使用一个局部表示 q（$\mathcal{H}^1$）和一个全局表示 $\mathcal{G}^2$ 来计算余弦相似度损失。我们将每个尺度级别的梯度作为层次化深度掩码模块输出的权重。对于聚类图，我们选择 5 个点特征作为初始特征。然后，我们计算空间维度上所有点特征与这 5 个初始特征之间的相似度。最后，我们将相似度最高的点进行聚类，并用相同的颜色显示同一聚类。

图 4 展示了可视化结果。Grad-CAM 结果表明，Hierarchical Deep-Masking 模块的所有尺度级别都能够成功关注图像中的主体。覆盖主体的区域具有更高的激活值，而背景区域（例如地面或天空）获得的注意力较低。对于不同的尺度级别，注意力的粒度也有所不同。 P_3具有最大的分辨率，展示了细粒度注意力。在第一张图像中，猪头的轮廓得到了准确感知。聚类图显示，具有相似上下文的图块特征可以聚类在一起。背景和主体是可区分的。在细粒度的 P_3 级别中，这种区别表现得更加清晰。以第一张猪的图片为例，背景被聚类到同一个聚类中，而大部分猪身体现出较高的语义相似度。此外，在不同的尺度级别之间，聚类结果是一致的。可视化结果支持了我们的假设，即每个图块特征都包含特定的局部语义。对齐隐式局部图像块表示和全局图像表示的过程是有意义的。

8. 结论

本文提出了一种简单而有效的自监督学习方法，即 MaskDeep。我们在单一在线动量框架中对齐局部和全局表征。我们证明，从稀疏图像块特征推断全局图像内容是一项有效的自监督学习任务。此外，我们首次从隐式表征空间中构建图块特征，而不是对源图像进行掩码处理。我们对隐式表征空间进行了可视化。结果证明，每个点都包含相应上下文的语义信息。我们提出了三项创新设计：层次化深度掩码模块、多组策略和多目标策略。这些设计提升了我们方法的效率。与当前的自监督方法相比，我们的 MaskDeep 在分类、目标检测和分割等下游任务上表现更优。在 ResNet50 上训练 200 轮后，MaskDeep 在 ImageNet 数据集上的线性分类 Top-1 准确率达到 71.2%，取得了当前最先进的结果。在 COCO 目标检测任务上，MaskDeep 的边界框平均精度均值达到 41.5%，掩码平均精度均值达到 36.1，取得了当前最先进的结果，优于大多数目标级自监督方法。

参考文献

[1] Alexei Baevski、Wei-Ning Hsu、Qiantong Xu、Arun Babu、Jiatao Gu 和 Michael Auli。Data2Vec：一种用于语音、视觉和语言自监督学习的通用框架。发表于国际机器学习大会，第 1298–1312 页。PMLR，2022年。[2] Hangbo Bao、Li Dong、Songhao Piao 和 Furu Wei。BEiT：图像变换器的 BERT 预训练。*arXiv* 预印本 *arXiv:2106.08254*，2021年。[3] Adrien Bardes、Jean Ponce 和 Yann LeCun。VICRegL：局部视觉特征的自监督学习。*arXiv* 预印本 *arXiv:2210.01571*，2022年。[4] Mathilde Caron、Ishan Misra、Julien Mairal、Priya Goyal、Piotr Bojanowski 和 Armand Joulin。通过对比聚类分配进行视觉特征的无监督学习。发表于神经信息处理系统进展，33：9912–9924，2020年。[5] Ting Chen、Simon Kornblith、Mohammad Norouzi 和 Geoffrey Hinton。用于视觉表征对比学习的简单框架。发表于国际机器学习大会，第 1597–1607 页。PMLR，2020年。[6] Xinlei Chen、Haoqi Fan、Ross Girshick 和 Kaiming He。通过动量对比学习改进基线。*arXiv* 预印本 *arXiv:2003.04297*，2020年。[7] Xinlei Chen 和 Kaiming He。探索简单的孪生表征学习。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 15750–15758 页，2021年。[8] Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。ImageNet：大型层次化图像数据库。发表于*2009*年*IEEE*计算机视觉与模式识别会议，第 248–255 页。IEEE，2009年。[9] Mark Everingham、Luc Van Gool、Christopher KI Williams、John Winn 和 Andrew Zisserman。PASCAL 视觉对象类别（VOC）挑战赛。国际计算机视觉期刊，88(2)：303–338，2010年。[10] Jean-Bastien Grill、Florian Strub、Florent Altch´e、Corentin Tallec、Pierre Richemond、Elena Buchatskaya、Carl Doersch、Bernardo Avila Pires、Zhaohan Guo、Mohammad Ghesh-laghi Azar 等。Bootstrap Your Own Latent——一种自监督学习新方法。发表于神经信息处理系统进展，33：21271–21284，2020年。[11] Yuanfan Guo、Minghao Xu、Jiawen Li、Bingbing Ni、Xuanyu Zhu、Zhenbang Sun 和 Yi Xu。HCSC：层次化对比选择性编码。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 9706–9715 页，2022年。[12] Kaiming He、Xinlei Chen、Saining Xie、Yanghao Li、Piotr Doll´ar 和 Ross Girshick。掩码自编码器是可扩展的视觉学习器。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 16000–16009 页，2022年。[13] Kaiming He、Haoqi Fan、Yuxin Wu、Saining Xie 和 Ross Girshick。用于无监督视觉表征学习的动量对比。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 9729–9738 页，2020年。[14] Kaiming He、Georgia Gkioxari、Piotr Doll´ar 和 Ross Girshick。Mask R-CNN。在*IEEE*国际计算机视觉会议论文集中，第2961–2969页，2017年。[15] Sergey Ioffe 和 Christian Szegedy。批量归一化：通过减少内部协变量偏移加速深度网络训练。在国际机器学习会议中，第448–456页。PMLR，2015年。[16] Yangguang Li、Feng Liang、Lichen Zhao、Yufeng Cui、Wanli Ouyang、Jing Shao、Fengwei Yu 和 Junjie Yan。监督无处不在：一种数据高效的对比语言-图像预训练范式。*arXiv*预印本 *arXiv:2110.05208*，2021年。[17] Tsung-Yi Lin、Piotr Doll´ar、Ross Girshick、Kaiming He、Bharath Hariharan 和 Serge Belongie。用于目标检测的特征金字塔网络。在*IEEE*计算机视觉与模式识别会议论文集中，第2117–2125页，2017年。[18] Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Doll´ar 和 C Lawrence Zitnick。Microsoft COCO：上下文中的常见物体。在欧洲计算机视觉会议中，第740–755页。Springer，2014年。[19] Ilya Loshchilov 和 Frank Hutter。SGDR：具有热重启的随机梯度下降。*arXiv*预印本 *arXiv:1608.03983*，2016年。[20] Vinod Nair 和 Geoffrey E Hinton。修正线性单元改进受限玻尔兹曼机。在*Icml*中，2010年。[21] Shaoqing Ren、Kaiming He、Ross Girshick 和 Jian Sun。Faster R-CNN：利用区域提议网络实现实时目标检测。神经信息处理系统进展，28：91–99，2015年。[22] Ramprasaath R Selvaraju、Michael Cogswell、Abhishek Das、Ramakrishna Vedantam、Devi Parikh 和 Dhruv Batra。Grad-CAM：通过基于梯度的定位解释深度网络。在*IEEE*国际计算机视觉会议论文集中，第618 626页，2017 – 。[23] Jing Shao、Siyu Chen、Yangguang Li、Kun Wang、Zhenfei Yin、Yinan He、Jianing Teng、Qinghong Sun、Mengya Gao、Jihao Liu 等。Intern：迈向通用视觉的新型学习范式。*arXiv*预印本 *arXiv:2111.08687*，2021年。[24] Luya Wang、Feng Liang、Yangguang Li、Wanli Ouyang、Honggang Zhang 和 Jing Shao。RePre：通过重构式预训练改进自监督视觉Transformer。*arXiv*预印本 *arXiv:2201.06857*，2022年。[25] Xinlong Wang、Rufeng Zhang、Chunhua Shen、Tao Kong 和 Lei Li。用于自监督视觉预训练的密集对比学习。在*IEEE/CVF*计算机视觉与模式识别会议论文集中，第3024–3033页，2021年。[26] Fangyun Wei、Yue Gao、Zhirong Wu、Han Hu 和 Stephen Lin。通过目标级对比学习对齐检测预训练。神经信息处理系统进展，34，2021年。[27] Zhirong Wu、Yuanjun Xiong、Stella X Yu 和 Dahua Lin。通过非参数实例辨别进行无监督特征学习。在*IEEE*计算机视觉会议论文集中，

计算机视觉与模式识别，第3733 3742页，2018 – 。[28] Enze Xie、Jian Ding、Wenhai Wang、Xiaohang Zhan、Hang Xu、Peize Sun、Zhenguo Li 和 Ping Luo。DetCo：用于目标检测的无监督对比学习。在*IEEE/CVF*国际计算机视觉会议论文集中，第8392–8401页，2021年。[29] Zhenda Xie、Zheng Zhang、Yue Cao、Yutong Lin、Jianmin Bao、Zhuliang Yao、Qi Dai 和 Han Hu。Simmim：用于掩码图像建模的简单框架。在*IEEE/CVF*计算机视觉与模式识别会议论文集中，第9653–9663页，2022年。[30] Ceyuan Yang、Zhirong Wu、Bolei Zhou 和 Stephen Lin。用于自监督检测预训练的实例定位。在*IEEE/CVF*计算机视觉与模式识别会议论文集中，第3987–3996页，2021年。[31] Jinghao Zhou、Chen Wei、Huiyu Wang、Wei Shen、Cihang Xie、Alan Yuille 和 Tao Kong。IBOT：利用在线分词器进行图像BERT预训练。*arXiv*预印本 *arXiv:2111 07832*，2021年。。