arXiv:2303.03307v2 [cs.CV] 3 Dec 2023

使用最大流形容量表征进行自然图像的高效编码

Thomas Yerxa [1] **Yilun Kuang** [2,3] **Eero Simoncelli** [1,2,3] **SueYeon Chung**[1,2]1纽约大学神经科学中心2Flatiron研究所计算神经科学中心3库朗数学科学研究所，tey214@nyu.edu

摘要

高效编码假说认为，感觉系统已经适应其输入的统计特性，在生物学约束下最大化环境信号与其表征之间的互信息。尽管这一假说十分优雅，但信息论量 notoriously 难以测量或优化，因此关于该假说的大多数研究都采用近似、界或替代指标（例如重建误差）。近期提出的一种编码效率度量“流形容量”，量化了能够以线性可分方式表征的物体类别数量，但其计算依赖计算量极大的迭代过程，因而无法用作目标函数。在此，我们将这一度量简化为便于直接优化的形式，用其学习最大流形容量表征（MMCRs），并证明其在当前自监督学习（SSL）识别基准上可与最先进的结果相媲美。实证分析揭示了 MMCR 与其他 SSL 框架所学习表征之间的重要差异，并提出了一种流形压缩产生类别可分性的机制。最后，我们在一组神经预测性基准上评估了多种 SSL 方法，发现 MMCR 作为灵长类腹侧通路的模型具有很强的竞争力。

1 引言

生物视觉系统能够学习关于世界的复杂表征，以支持广泛的认知行为，而无需依赖大量带标签的示例。高效编码假说（8；65）认为，这是通过使感觉表征适应输入信号的统计特性来实现的，从而减少冗余或维数。视觉信号具有若干明显的冗余来源。它们随时间缓慢变化，因为时间上相邻的输入通常对应于同一场景的不同视图，而这些视图彼此之间通常比不同场景的视图更加相似。此外，单个场景内部的变化往往对应于少量参数的变化，例如控制观察条件和照明条件的参数，因此其本质上是低维的。此前的许多研究结果表明，神经回路的计算可以被视为与自然环境中的此类结构相匹配（48；29；65；17）。针对各种模态开展的研究已经在神经数据中发现了与行为任务相关的几何结构（10；25；43；33；56），并探索了用于量化这些结构的度量指标。

近年来，“流形容量理论”的发展为神经表征的几何特性（大小和维数）与其编码容量之间提供了更明确的联系（18）。该理论已用于评估生物和人工神经网络在不同模态下的效率（17；24；66；52）。然而，将其作为构建模型表征的设计原则尚未得到探索。

第37届神经信息处理系统大会（NeurIPS 2023）。

受这些观察的启发，我们希望学习这样的表征：包含同一场景不同视图的流形既紧凑又具有低维数，而对应不同场景的流形则最大程度地彼此分离。具体而言：

- 我们提出了一种流形容量形式，可用作学习的目标函数。
- 我们证明，采用最大流形容量表征（MMCR），能够实现高质量的物体识别（达到自监督学习的最先进水平）；使用标准线性评估范式进行评估时（即对自监督网络的输出应用经过优化的线性分类器）（14）。
- 通过分析内部表征和学习信号，我们探究了无监督目标函数如何促使语义相关特征涌现的潜在机制。
- 我们将其学习到的表征与来自猕猴视觉皮层的神经数据进行比较，以验证 MMCR 作为脑模型的有效性。

因此，我们的工作利用表征效率这一规范性目标，构建了一种新颖的视觉表征模型；该模型既能有效完成识别任务，也与灵长类腹侧通路中的神经反应相一致。

1.1 相关工作

神经表征几何。 以往研究试图描述表征几何如何影响下游任务表现的不同方面，通常通过参与率等谱量进行测量（参与率是协方差矩阵特征值的 l_1 范数与 l_2 范数之比的平方）（32）。Elmoznino 和 Bonner（27）发现，人工神经网络表征的高维性与预测神经数据以及泛化到未见类别的能力相关。Stringer 等人（67）观察到，小鼠皮层对自然图像的表征谱遵循幂律，其衰减系数接近 1；Agrawal 等人（2）报告称，在人工表征中，谱衰减系数接近 1 能有效预测表征泛化到下游任务的能力。

自监督学习。 我们的方法与近期对比式自监督表征学习（SSL）的进展相关，并从中汲取灵感，但其动机和形式有明显不同。许多近期框架设计目标函数，旨在最大化同一对象不同视图表征之间的互信息（58；14；58；68；4））。然而，在高维特征空间中估计互信息十分困难（9）；此外，尚不清楚在目标函数中更精确地逼近互信息是否会带来更好的表征（71）。相比之下，基于自旋玻璃理论发展出的容量度量（35；1）是在“大 N（热力学）极限”下推导的，因此适用于环境维数较大的情形（18；5）。我们考察了其中一种度量——迄今为止仅用于评估表征质量——是否也可作为自监督学习的目标函数。

许多自监督学习方法会最小化同一图像不同增强视图的表征之间的距离，同时采用约束来防止模型坍塌为平凡解（例如，排斥负样本对（14），或进行特征空间白化（77；7；28））。单一成对距离比较的局限性已得到证明，值得一提的是，SwAV（13）中采用的“多裁剪”策略，以及 Tian 等人（68）提出的对比多视图编码方法，都体现了这一点。我们的方法基于这样一种假设：图像的不同视图构成一个连续流形，我们的目标是对其进行压缩。我们通过各视图表征的奇异值谱来刻画每组图像视图，并使用核范数作为流形大小和维数的综合度量。

核范数此前已被用于推断或促使数据表示呈现低秩结构（42；72；49）。具体而言，Wang 等人（72）将其用作正则化项，以补充 InfoNCE 损失。如下文所述，我们的方法与传统的 InfoNCE 损失相比有更根本的不同。我们没有将低秩先验与基于逻辑回归的似然结合起来，而是选择了更对称的做法，即采用高秩似然。这使目标函数能够明确抑制维度坍塌——这是自监督学习中的一个众所周知的问题（45）。

2 最大流形容量表示

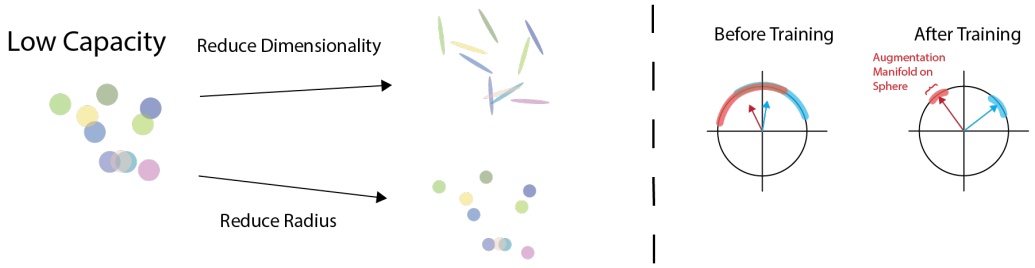

图1：高容量和低容量表征的二维示意图。左：一组随机球形区域的容量（线性可分性）可以通过减小其半径（同时保持其维数不变），或减小其维数（同时保持其平均半径不变）来提高。右：本文提出的目标函数旨在最小化归一化数据向量（即位于单位球面上）的核范数（等于半径与维数平方根的乘积）。训练前，流形具有较大的范围，因此其对应质心向量的矩阵具有较低的核范数。训练后，容量增加：流形被压缩并彼此排斥，从而得到核范数更大、相似性更低的质心矩阵。

2.1 流形容量理论

考虑一组嵌入维数为 D 的特征空间中的 P 个流形，每个流形都分配有一个类别标签。流形容量理论关注的问题是：$\frac{P}{D}$ 的最大值是多少，才能以较高概率存在一个超平面，将随机二分分类（22；34）分隔开来？近期理论研究表明，存在一个临界值，称为流形容量 α_C：当 $\frac{P}{D} < \alpha_C$ 时，找到分隔超平面的概率约为 1。0；当 $\frac{P}{D} > \alpha_C$ 时，该概率约为 0。0（18）。容量 α_C 可根据三个关键量准确预测：（1）流形半径 R_M，衡量流形的大小相对于其质心到原点距离的比例；（2）流形维数 D_M，量化流形沿哪些维度具有显著延展；（3）流形质心之间的相关性。当质心相关性较低时，流形容量可近似表示为 $\phi(R_M\sqrt{D_M})$，其中 $\phi(\cdot)$ 是一个单调递减函数。

对于任意几何形状的流形，可通过一个迭代过程计算其半径和维数：该过程交替确定每个流形上与分类问题相关的“锚点”集合，并计算这些锚点随机投影的统计量（20）。这一过程计算成本高且不可微，因此不适合作为目标函数使用。有关一般理论的更多细节，请参见附录C。然而，若假设流形呈椭圆形，则半径和维数都可用解析形式表示：

$$R_M = \sqrt{\sum_i \lambda_i^2}, \qquad D_M = \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2}, \tag{1}$$

其中，λ_i^2 是该流形协方差矩阵的特征值。作为比较，在计算一组由 100 个 128 维流形组成的数据集的这些值时，每个流形采样 100 个点，使用解析表达式的速度大约比通用迭代过程快 500 倍（按“实际耗时”计算）。

根据流形半径和维数的这些定义，我们可以将容量表示为 $\alpha_C = \phi(\sum_i \sigma_i)$，其中 σ_i 是包含流形上各点的矩阵的奇异值（等价地说，是协方差矩阵特征值的平方根）。在这种形式下，总和就是奇异值的 L_1 范数，称为矩阵的核范数 。将其用作目标函数时，该度量会偏好稀疏解（即非零奇异值数量较少的解），对应于较低的维数。值得将这一目标函数与另一种用于量化大小的自然候选指标进行比较：协方差矩阵的行列式。行列式等于乘积

特征值的乘积（它反映了相应椭球体的平方体积），但不具备核范数对较低维数的偏好。具体而言，由于只要有一个或多个特征值为零，行列式就为零，因此它无法区分维数不同但体积均为零的流形。有损编码率（熵）也被用作紧致性的度量（76）；在高斯模型下，它可简化为协方差矩阵的对数行列式 Ma et al. (50)。在这种情况下，计算行列式之前，会将单位矩阵加到特征协方差矩阵的一个倍数上，从而解决上述维数问题。

2.2 优化流形容量

现在，我们构建一个基于流形容量的自监督学习目标函数。对于每张输入图像（记为向量 $\mathbf{x}_b \in \mathbb{R}^{\mathbb{D}}$），我们通过应用一组随机增强（每种增强均从同一分布中采样），从对应的流形中生成 k 个样本，得到流形样本矩阵 $\tilde{\boldsymbol{X}}_b \in \mathbb{R}^{D\times k}$。每张增强后的图像都由深度神经网络进行转换，该网络计算由 θ参数化的非线性函数 $f(\mathbf{x}_b;\, \theta)$，并将 d 维响应投影到单位球面上，从而得到流形响应矩阵 $\boldsymbol{Z}_b \in \mathbb{R}^{d\times k}$。质心 $\boldsymbol{c}_b$ 通过对各列（响应向量）取平均值来近似。对于一组图像 $\{\mathbf{x}_1, ..., \mathbf{x}_B\}$，我们计算归一化响应矩阵 $\{\boldsymbol{Z}_1, ..., \boldsymbol{Z}_B\}$，并将相应的质心组合成矩阵 $\boldsymbol{C} \in \mathbb{R}^{d\times B}$。

给定这些响应及其质心，MMCR损失函数可以简单地写为：

$$\mathcal{L} = -||\boldsymbol{C}||_*, \tag{2}$$

其中，$||\cdot||_*$ 表示核范数。该损失函数明确地最大化“质心流形”的范围；有趣的是，仅此就足以学到有用的表征。具体而言，最大化$||\boldsymbol{C}||_*$会隐式地缩小各个对象流形的范围，其衡量方式为 $||\boldsymbol{Z}_b||_*$。我们将在下文阐释这一效应。

仅通过最大化质心核范数进行压缩。 每个质心向量都是单位向量的均值，因此其范数与这些单位向量的平均余弦相似度呈线性关系。具体而言，

$$||\boldsymbol{c}_b||^2 = \frac{1}{K} + \frac{2}{K^2}\sum_{k=1}^{K}\sum_{l=1}^{k-1}\boldsymbol{z}_{b,k}^T\boldsymbol{z}_{b,l} \tag{3}$$

这里 $\boldsymbol{z}_{b,\,i}$ 表示 $\boldsymbol{x}_b$ 的 i^{th} 增强表征。通过考察矩阵的奇异向量分布如何取决于其各列向量的范数和两两相似性，我们可以获得进一步的见解。尽管任意矩阵的奇异值都不存在闭式解，但由两个列向量构成的矩阵可以提供有益的直观理解。若 $\boldsymbol{C} = [\boldsymbol{c}_1, \boldsymbol{c}_2]$，$\boldsymbol{Z}_1 = [\boldsymbol{z}_{1,1}, \boldsymbol{z}_{1,2}]$, $\boldsymbol{Z}_2 = [\boldsymbol{z}_{2,1}, \boldsymbol{z}_{2,2}]$，则$\boldsymbol{C}$ 和 $\boldsymbol{Z}_i$ 的奇异值为：

$$\boldsymbol{\sigma}(\boldsymbol{C}) = \sqrt{\frac{||\boldsymbol{c}_1||^2 + ||\boldsymbol{c}_2||^2 \pm ((||\boldsymbol{c}_1||^2 - ||\boldsymbol{c}_2||^2)^2 + 4(\boldsymbol{c}_1^T\boldsymbol{c}_2)^2)^{1/2}}{2}}, \tag{4}$$

$$\boldsymbol{\sigma}(\boldsymbol{Z}_i) = \sqrt{1 \pm \boldsymbol{z}_{i,1}^T\boldsymbol{z}_{i,2}}. \tag{5}$$

因此，当质心向量的范数达到最大（由于它们是单位向量的质心，其上限为 1），且彼此正交时，$||\boldsymbol{\sigma}(\boldsymbol{C})||_1 = ||\boldsymbol{C}||_*$ 取最大值。如上所述，质心范数是流形内相似性的线性函数。类似地，当流形内相似性达到最大时，$||\boldsymbol{\sigma}(\boldsymbol{Z}_i)||_1 = ||\boldsymbol{Z}_i||_*$ 取最小值。因此，单项 $||\boldsymbol{C}||_*$ 囊括了对比学习框架的两个关键要素。我们将在下文证明，仅最大化 $||\boldsymbol{C}||_*$ 便足以学习到有用的表征。这是因为，对比学习中“正样本”的压缩作用是通过构造质心向量来实现的，因此该目标函数并非不含正样本。例如，如果只使用一个视图，目标函数就缺少压缩成分，无法产生有用的表征。在附录F中，我们通过与显式最小化$||\boldsymbol{Z_b}||_*$ 的情况进行比较，实证表明这种隐式形式能有效降低 $||\boldsymbol{Z_b}||_*$ 。因此，决定流形容量的三个因素（半径、

维数和质心相关性）都可以简洁地用一个仅含单项 $-||\boldsymbol{C}||_*$的目标函数来表示。

计算复杂度。 评估我们的方法的损失需要对 $\boldsymbol{C} \in \mathbb{R}^{d\times B}$进行奇异值分解，其复杂度为 $\mathcal{O}($ $Bd \times \min(B,\ d)$)，其中 B是批量大小， d 是输出的维数。相比之下，在一个批次中计算所有成对距离的对比方法，其复杂度为 $\mathcal{O}(B^2d)$；涉及正则化协方差结构的非对比式方法，其复杂度为 $\mathcal{O}(Bd^2)$。此外，我们的方法的复杂度与所用视图的数量无关（尽管特征提取阶段的复杂度随视图数量线性增长），而成对相似度度量的复杂度则随视图数量呈二次增长。还值得注意的是，通过最大化 $||\boldsymbol{C}||_*$进行隐式压缩，相较于显式压缩，在计算复杂度方面更有优势。这是因为评估诸如 $\sum_{b=1}^{B}||\boldsymbol{Z}_b||_*$ 这样的项，其计算复杂度为 $\mathcal{O}(B^2d \times \min(B,\ d))$。

2.3 最优嵌入的条件

近期，HaoChen 等人（39）基于“总体增强图”的谱分解提出了一个框架，为线性探测下自监督学习在下游任务上的表现提供了理论保证。Balestriero 和 LeCun（6）将这项工作扩展到对其他各种自监督学习目标的研究。如下文所示，利用这一方法也能为我们提出的目标函数下表征的最优性导出明确条件。

给定一个数据集 $\boldsymbol{X'} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_N]^T \in \mathbb{R}^{N\times D'}$，我们通过为原始数据随机创建 k 个增强视图，构建一个新数据集： $\boldsymbol{X}$ = [视图$_1$（$\boldsymbol{X'}$）.....，视图$_k$（$\boldsymbol{X'}$）] $\in \mathbb{R}^{Nk\times D}$。这样做的好处是，我们现在可以利用这样一个认知：同一底层数据点的不同视图在语义上相关。我们可以用对称矩阵 $\boldsymbol{G} \in \{0,1\}^{Nk\times Nk}$ 表达这一相似性概念：如果增强后的数据点 i 和 j 在语义上相关，则 $\boldsymbol{G}_{ij} = 1$；并且由于任何数据点都与自身相关，因此 $\boldsymbol{G}_{ii} = 1$。我们可以对 $\boldsymbol{G}$ 进行归一化，使其行和列的总和为 1（因此 $\boldsymbol{G}$ 的各行是 k 稀疏的，非零元素均为 $1/k$）。
现在设 $\boldsymbol{Z} \in \mathbb{R}^{Nk\times d}$为增强数据集的一个嵌入。那么我们有 $\boldsymbol{GZ} = [\boldsymbol{C}, ..., \boldsymbol{C}]^T$，其中 $\boldsymbol{C}$ 是上文引入的质心向量矩阵，而 $\boldsymbol{C}$ 的重复次数为 k。由于 $\boldsymbol{\sigma}([\boldsymbol{C}, ..., \boldsymbol{C}]) = \sqrt{k}\boldsymbol{\sigma}(\boldsymbol{C})$，因此我们可以将 MMCR损失函数写为：

$$\mathcal{L} = -||\boldsymbol{GZ}||_* \tag{6}$$

这一联系使我们能够对该损失下的最优嵌入 $\boldsymbol{Z}$ 作出以下论述，我们将在附录A中证明这些论述：

定理： 在所提出的损失下，最优嵌入 $\boldsymbol{Z}^*$的左奇异向量是 $\boldsymbol{G}$的特征向量，而 $\boldsymbol{Z}^*$ 的奇异值与 $\boldsymbol{G}$的前 d 个特征值成正比。

3 结果

架构。 在所有实验中，我们均采用 ResNet-50(40) 作为骨干架构（对于在 CIFAR 上训练的变体，我们移除了最大池化层）。按照 Chen 等人（14）的做法，我们在 ResNet 的平均池化层输出后添加一个小型感知机，使得 $z_i = g(h(x_i))$，其中 h是 ResNet， g 是多层感知机。对于 ImageNet-1k/100，我们使用维度为$[8192, 8192, 512]$ 的多层感知机；对于较小的数据集，我们使用 $[512, 128]$。

优化。 我们采用了文献（38）提出的数据增强组合。对于 ImageNet，我们使用 LARS 优化器，学习率设为 4.8，前 10 个训练轮数进行线性预热，之后采用余弦衰减；批量大小为 2048，预训练 100 个训练轮数。请注意，尽管我们报告的结果使用默认批量大小 2048，但批量大小小至 256 也能获得合理结果（与批量大小 2048 相比，性能下降 1.2%——有关批量大小扫描，请参见附录 J）。此外，在 ImageNet 预训练中，我们还使用了动量编码器（所有视图均输入在线网络和一个额外网络；该额外网络的参数是在线网络参数的缓慢移动平均值，并将每张图像的所有嵌入取平均以形成质心）。我们发现，使用动量编码器在下游分类方面略有优势

表 1：学习得到的特征在下游分类任务上的评估。最左侧几列展示了在 ImageNet（IN）上采用标准冻结线性评估流程的结果。此设置下大多数方法的结果取自 Ozsoy 等人（59），SwAV 的结果则取自原始论文（13）。第 2 列和第 3 列展示了在 ImageNet 上的半监督评估结果（分别使用 1% 和 10 % 的标签进行微调）。VICReg 和 CorInfo Max 在此设置下的结果摘自 Ozsoy 等人（59）。最后三列展示了在其他数据集上的冻结线性评估结果。我们评估了在 100 个训练轮数设置下可在线获取预训练权重的模型；MoCo、Barlow Twins 和 BYOL 的结果取自 solo-learn da Costa 等人（23）（https://github. com/vturrisi/solo-learn），而 SwAV 和 SimCLR 的结果取自 VISSL Goyal 等人（37）（https://github.com/facebookresearch/vissl/blob/main/MODEL_ZOO.md）。对于我们进行的所有评估，报告的均为 3 次评估运行的均值和标准差）。

方法	IN	1%	10%	Food-101	Flowers-102	DTD
W-MSE (28)	69.4	-	-	-	-	-
NNCLR (26)	69.4	-	-	-	-	-
SwAV (13)	64.6	-	-	-	-	-
SimSiam (15)	68.1	-	-	-	-	-
CorInfoMax (59)	69.1	44.9	64.3	-	-	-
VICReg (7)	68.7	44.8	62.2	-	-	-
BarlowTwins (77)	68.7	45.1 ±.12	61.7 ±.03	69.8 ±.03	86.2 ±.22	67.7 ±.20
SimCLR（14）	66.5	42.6 ±.02	61.6 ±.09	67.2 ±.24	84.0 ±.19	64.8 ±.07
BYOL (38)	69.3	49.8 ±.05	65.0 ±.05	70.6 ±.1	84.8 ±.43	67.6 ±.14
MoCo-V2 (16)	67.4	43.4 ±.07	63.2 ±.07	68.6 ±.03	82.4 ±.27	66.6 ±.16
SwAV（13）	**72.1**	49.8 ±.09	66.9 ±.05	72.1 ±.08	89.3 ±.1	68.2 ±.18
MMCR（2 个视图）	69.5 ±.02	46.6 ±.02	63.9 ±.02	72.0 ±.02	90.0 ±.24	68.5 ±.07
MMCR（4 个视图）	71.5 ±.04	49.4 ±.05	66.0 ±.05	73.2 ±.07	91.0 ±.04	**70.4** ±.46
MMCR（8 个视图）	**72.1** ±.04	**51.0** ±.02	**67.7** ±.11	**73.6** ±.04	**91.4** ±.07	**70.0** ±.24

有关性能，请参见附录 O 中的这项消融实验。对于较小的 CIFAR-10，我们使用了较小的批量大小、更多的视图（40 个），以及采用固定学习率的 Adam 优化器。具体细节请参见附录 D。

3.1 迁移到下游任务

我们采用标准的线性评估技术：冻结编码网络的参数，并通过监督训练线性分类器 (14)，以验证我们的方法能否从数据中提取语义相关特征。我们还进行了半监督评估，即使用少量带标签样本对所有模型参数进行微调；此外，我们还检验了所学特征能否泛化到三个分布外数据集：Flowers-102、Food-101 和可描述纹理数据集 (57; 11; 19)。结果汇总于表 1，线性分类器的训练细节见附录 G。最后，我们还评估了表现最佳的模型以及各个基线在 VOC07 数据集上的目标检测性能。我们遵循 (41; 77) 的做法，使用 Faster R-CNN 检测头和 C-4 主干网络，按照 1x 训练计划对表征网络进行微调，以用于检测。MMCR 的平均精度均值（mAP）达到 54.6，基线方法的表现则介于 53.1 至 56.0 之间。这表明，尽管我们的框架受到分类理论的启发，所学特征仍能泛化到其他视觉任务。详细结果见附录 M。

请注意，所有纳入的模型均采用相同的骨干架构（ResNet-50）、数据集（ImageNet-1k）和预训练周期数（100）进行训练；我们在附录 P 中简要探讨了更长时间预训练的影响。关于在较小数据集上预训练的网络的结果，请参见附录 I。

3.2 学得表征的分析

接下来，我们进行一系列实验，以阐明通过不同自监督学习流程学习到的表征之间的差异，并提出一种增强流形压缩产生类别可分性的机制。

为降低计算需求，这些分析针对在 CIFAR-10 数据集上训练的模型开展。

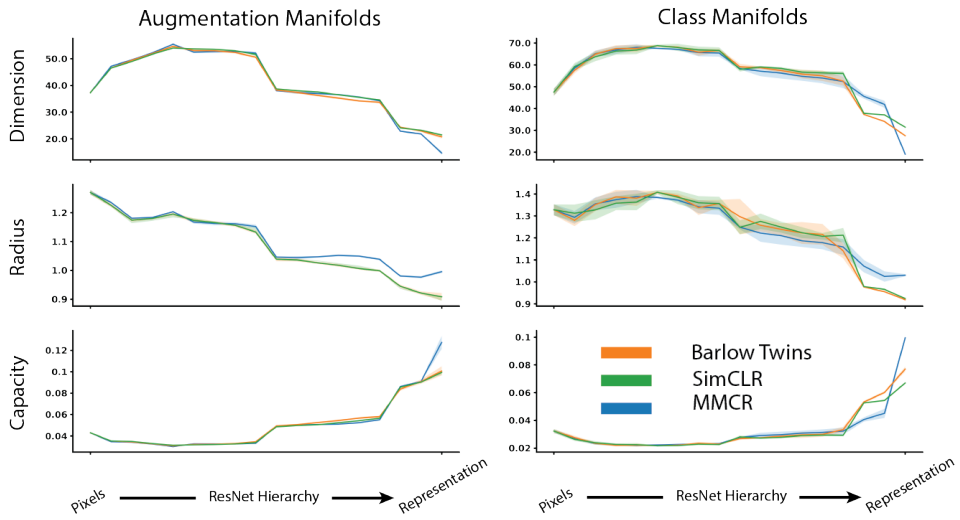

图 2：平均场流形容量分析。流形维数（上排）、半径（中排）和容量（下排）随表征层级中的阶段变化（从像素输入到编码器输出／学习到的表征）。阴影区域表示均值周围的 95% 置信区间（分析使用了数据集中的 5 个不同随机样本，见附录 E）。

平均场理论流形容量。图 2 显示，尽管我们的表征是通过假设椭圆流形几何的目标函数进行优化的，但它仍具有较高的、更一般的平均场流形容量（相较于基线方法）。为求完整，我们还分析了类别流形的几何结构；其上的点是同类别不同样本的表征。这项分析进一步表明，学习最大化增强流形容量能够压缩并分离类别流形，从而形成有用的表征。有趣的是，MMCR 似乎采用了不同于基线方法的策略来提高容量，产生半径更大、但维数更低的类别／增强流形（图 2）。这些几何差异直到层级的末端才显现，这表明各网络的早期层执行的是模式化变换，而损失函数诱导的专门化要到后续层才会出现。

梯度相干性促成神经流形的涌现。我们假设，MMCR 中的类别可分性源于：同类别样本所对应的增强流形，比不同类别样本所对应的增强流形，能够通过更相似的变换得到最优压缩。为通过实证检验这一假设，我们评估属于同一类别的输入对应的目标函数梯度。随后，我们可以检查同一类别的不同批次所得到的梯度是否比不同类别的梯度彼此更相似；若是，则说明压缩同类别增强流形的策略彼此相对相似。图3表明事实确实如此：无论是在不同训练轮次还是模型层级中，以余弦相似度衡量的类内梯度相干性始终高于类间梯度相干性。

流形子空间对齐。 类内梯度相干性为类别可分性的产生提供了一种合理的机制解释，但无法解释为什么同一类别的样本会采用相似的压缩策略。为探究这一问题，我们考察像素域中增强流形的几何特性。在这里，我们观察到类内相似性分布与跨类别相似性分布之间存在细微但可测量的差异，如图4顶行所示。像素域中增强流形几何特性的这一细微差异，进而导致了前文观察到的梯度相干性增强；这使得表征能够以相似的方式重新排列和重塑同一类别的增强流形（图4底行），从而实现更好的类别线性分离。不仅质心

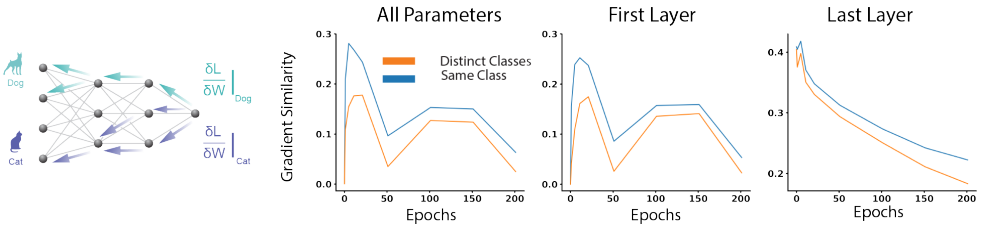

图3：单类别批次对的梯度余弦相似度。我们绘制了训练过程中，来自同一类别或不同类别的单类别批次所产生的梯度对，在模型参数的不同子集（所有参数，以及第一个和最后一个线性算子）上的平均成对相似度。由于使用了大量批次，这些均值的 95% 置信区间小到无法显示。左侧的可视化展示了单类别梯度以更相似的方向反向流经模型这一事实。

同类别流形的质心比来自不同类别的流形更集中于表征空间中相似的区域（图4第三列底行），而且同类别流形彼此之间的形状也更为相似（图4底行第1列和第2列显示，同类别流形占据相对夹角较小的子空间，并共享更多方差）。

接下来，我们探究 MMCR 目标函数学得的表征与通过其他自监督损失函数优化得到的表征有何不同。MMCR 鼓励质心彼此正交，而 InfoNCE 损失 (14) 则鼓励负样本对尽可能不相似；当它们位于同一 子空间的相反区域时，这一目标便得以实现。Barlow Twins (77) 损失并非特征向量相似性的显式函数，而是鼓励批次维度上的单个特征彼此相关、不同特征彼此不相关。图5表明，这些直觉得到了实证支持：与两种基线方法相比，MMCR 表征产生的增强流形质心彼此显著更为正交。

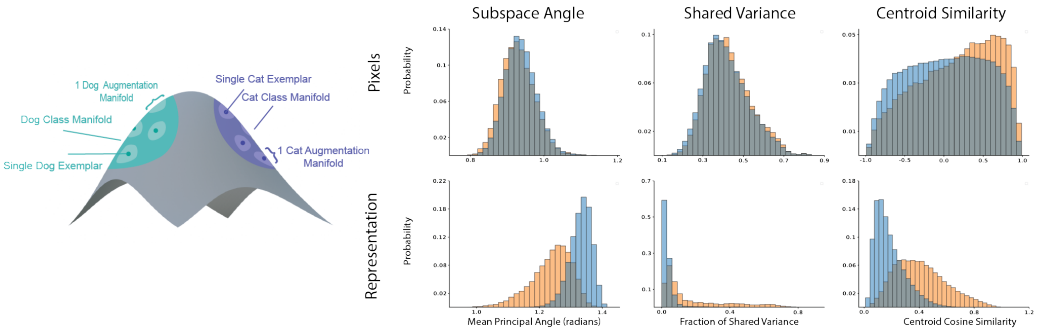

图 4：来自同一类别（橙色）和不同类别（蓝色）的增强流形的各种相似度指标分布。图中分别展示了输入图像（上排）和学习到的表征（下排）的结果。左：学习到的表征中样本—增强流形—类别流形结构的示意图。

3.3 生物学相关性

神经科学为人工神经网络的许多发展提供了启发，因此值得探究自监督学习网络能否刻画生物视觉系统中神经元的测量行为。作为一项简单的测试，表2展示了我们的模型与另外五种自监督学习模型在*Brain-Score* 数据库（64；30；53；51）上的表现。我们发现，MMCR 在解释灵长类视觉区 V2 和 V4 的神经数据方面表现最佳，在 V1 和 IT 方面排名第二，并且对 11 个数据集中的 8 个具有最高或第二高的预测能力（见附录 K）。

此外，我们还考察了近期描述的神经群体的一般响应谱特征。具体而言，Stringer 等人（67）报告称，……的特征谱

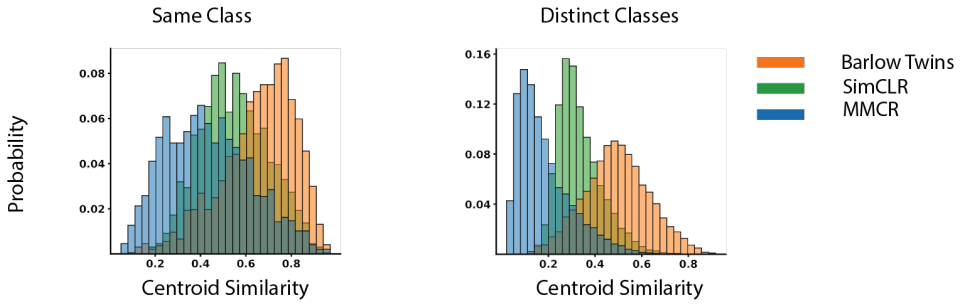

图5：根据不同自监督学习目标训练的模型的质心余弦相似度。左图显示同一类别样本的增强流形质心余弦相似度分布，右图则显示不同类别样本的相同分布。请注意，由于我们分析的是 ResNet 骨干网络的输出（经过了整流），因此余弦相似度的最小可能值为 0。

表2：自监督模型的神经预测和谱特性。前四列给出了在灵长类腹侧通路四个区域记录的神经元的平均 Brain-Score 相关值（64）。最后两列给出了参与率（PR）和谱衰减系数（α），这些数值是根据 ImageNet 验证集特征的 10 组自助法样本估算得出的。估算谱衰减系数时，我们省略了尾部（功率衰减更快的部分）；确切细节见 L。表中数值表示均值和均值标准误，粗体表示在标准误范围内的最佳表现。MMCR 的结果来自使用 8 个视图训练的模型。

模型	V1	V2	V4	IT	PR	α
MMCR	。**494** ±.006	。**311** ± .005	。**481** ±.005	.416 ± .003	279.2 ±.3	1.04 ± 1e-4
SimCLR	。**500** ±.007	.288 ± .007	.475 ±.004	。**420** ± .003	124.6 ±.1	1.34 ± 3e-4
BYOL	。**500** ±.006	.291 ± .007	。**477** ±.005	.404 ± .003	248.0 ±.4	1.35 ± 1e-4
MoCo	。**499** ±.007	.293 ± .006	。**477** ±.005	**.417** ± .003	147.7 ±.2	1.44 ± 3e-4
Barlow	。**498** ±.007	.293 ± .008	。**477** ±.005	.404 ± .003	252.2 ±.2	1.21 ± 3e-4
SwAV	.488 ±.007	.296 ± .009	.463 ±.004	.398 ± .003	163.9 ±.2	1.15 ± 3e-4

V1视觉区群体活动的协方差矩阵呈幂律衰减，衰减系数约为 1（$\lambda_n \propto n^{-\alpha}$ with $\alpha \approx 1$ where λ_n is the nth 特征值）。后续对人工网络的研究发现，这种衰减谱与更强的抵御对抗扰动的鲁棒性以及更优的泛化特性相关（55；2）。此外，近期多项研究探讨了表征维度与神经可预测性之间的联系（70；63）。具体而言，Elmoznino 和 Bonner（27）报告称，较高的内在维度（以表征协方差的参与率衡量）与更强的神经活动预测能力相关。表2列出了各表征在 ImageNet 验证集上的参与率，以及协方差谱的衰减系数（有关各项实验的更多详情，请参见附录L）。我们发现，MMCR 的参与率（维数）最高——请注意，这与目标函数中优化的量不同，后者位于嵌入空间中。此外，MMCR 得到的特征的衰减系数也最接近 1。

我们发现，在这一受控设置下，每个模型通过线性回归所解释的神经方差比例都非常相近。这与近期及同期的研究（21；62）一致；这些研究指出，无法区分替代模型是当前用于模型与大脑比较的主流范式的一项弱点。然而，我们的结果表明，不同的自监督学习算法会产生具有显著不同几何结构的表征（谱特性中的参与率和衰减系数等指标存在较大差异便证明了这一点）。这表明，有必要开发用于比较模型与数据的新度量指标，例如能够捕捉候选模型之间这些重要差异的几何度量。

4 讨论

我们提出了一种受流形容量理论启发的新型自监督学习算法。许多现有的自监督学习方法可根据其通过分别对嵌入格拉姆矩阵或协方差矩阵施加约束来避免坍塌，划分为“对比式”或“非对比式”。我们的方法在两者之间取得了折中，直接优化嵌入矩阵本身的奇异值。通过直接优化这一群体层面特征（谱），我们能够利用单项目标函数同时促进对齐与均匀性（71）。此外，这种形式无需在实例之间或维度之间进行大量成对比较。因此，学习 MMCR 十分高效，既不需要较大的批量大小，也不需要较高的嵌入维度。最后，我们的方法可以自然地扩展到多视图情形，同时以极小的计算成本增幅带来更优的性能。

我们的表述将流形几何近似为椭圆形，在降低计算需求的同时，仍能产生有用的学习信号以及具有高流形容量的网络。具体而言，我们能够充分利用流形容量分析的普适性，深入了解训练后 MMCR 网络的几何结构。未来研究可探索基于对平均场流形容量进行更温和降低的目标函数，以捕捉非椭圆形结构。耐人寻味的是，与 Barlow Twins 或 SimCLR 相比，我们的方法产生的增强流形和类别流形维数更低，但半径更大（图 2）。我们尚不清楚其原因，但这些差异表明，容量分析能够为阐明不同 SSL 范式所鼓励的不同编码策略提供有用工具。

最后，我们考察了近期提出的两种关于神经表征理想谱特性的理论。对于我们所考察的一组模型，接近 1 的谱衰减系数与分布内任务上的更佳表现以及对未见数据集的泛化相关（在每项分布外分类任务中，MMCR、SwAV 和 Barlow 均位列前三），这一发现总体上与 Agrawal 等人（2）的实证和理论发现一致。然而，我们还发现，高维性并不总是对应较强的神经可预测性：尽管具有最低的维数，SimCLR 在神经可预测性方面表现强劲。这或许并不令人意外，意味着仅凭全局维数不足以解释灵长类腹侧通路中神经元的响应特性。事实上，我们的实验进一步支持了日益增多的研究工作，即需要采用与线性可预测性互补的方法来区分视觉系统的候选模型（21）。令人欣慰的是，该领域已经开始着手解决这一局限；例如，一项同期工作（12）发现，将神经可预测性误差分解为不同模式，能够揭示不同模型如何拟合神经数据的不同方面（即使它们产生相似的总体可预测性）。

提升人工网络作为神经计算模型质量的一个有前景的方向，是纳入与生物学上合理的学习相关的约束（即需要采用局部学习规则；Illing 等人（44））。另一个互补方向是让训练数据分布更好地贴合生态输入。例如，相较于增强不变性，时间不变性似乎更适合作为神经系统优化的目标函数（73；3；61）。我们推测，若将 MMCR 目标函数扩展为在时间维度上运行，可能会非常适合神经回路实现，因为其计算只需要一些机制来追踪（1）短时间尺度上的时间均值（以形成质心），以及（2）群体活动在更长时间尺度上的奇异值。

神经科学领域不断涌现出新近采集的数据集，这些数据集记录了规模日益庞大的神经元群体活动；与此同时，利用规范性视角理解这些测量结果的方法也由来已久。本文证明，一种已被证实有助于从神经数据中获取洞见的技术（60；75；31），可以重新表述为目标函数，用于学习图像的有效抽象表示。未来的工作应利用这些学习模型生成可通过实验检验的预测，并为新的实验设计提供约束，从而在建模与分析之间形成闭环。

参考文献

[1] Alia Abbaras、Benjamin Aubin、Florent Krzakala 和 Lenka Zdeborová。Rademacher复杂度与自旋玻璃：复制理论与学习统计理论之间的联系。载于数学与科学机器学习，第27–54页。PMLR，2020年。[2] Kumar K Agrawal、Arnab Kumar Mondal、Arna Ghosh 和 Blake Richards。通过测量特征谱衰减评估自监督学习中的表征质量。神经信息处理系统进展，35：17626–17638，2022年。[3] Arthur Aubret、Markus R Ernst、Céline Teulière 和 Jochen Triesch。是时候增强自监督视觉表征学习了。载于第十一届国际学习表征会议，2022年。[4] Philip Bachman、R Devon Hjelm 和 William Buchwalter。通过最大化不同视图之间的互信息来学习表征。神经信息处理系统进展，32，2019年。[5] Yasaman Bahri、Jonathan Kadmon、Jeffrey Pennington、Sam S Schoenholz、Jascha Sohl-Dickstein 和 Surya Ganguli。深度学习的统计力学。《凝聚态物理年度评论》，11(1)，2020年。[6] Randall Balestriero 和 Yann LeCun。对比式与非对比式自监督学习能够恢复全局和局部谱嵌入方法。*arXiv*预印本 *arXiv:2205.11508*，2022年。[7] Adrien Bardes、Jean Ponce 和 Yann LeCun。VICReg：用于自监督学习的方差—不变性—协方差正则化。*arXiv*预印本 *arXiv:2105.04906*，2021年。[8] Horace B Barlow 等。感觉信息转换可能遵循的原则。《感觉通信》，1(01)，1961年。[9] Mohamed Ishmael Belghazi、Aristide Baratin、Sai Rajeshwar、Sherjil Ozair、Yoshua Bengio、Aaron Courville 和 Devon Hjelm。互信息神经估计。载于国际机器学习会议，第531–540页。PMLR，2018年。[10] Silvia Bernardi、Marcus K Benna、Mattia Rigotti、Jérôme Munuera、Stefano Fusi 和 C Daniel Salzman。海马体和前额叶皮层中抽象的几何结构。《细胞》，183(4)：954–967，2020年。[11] Lukas Bossard、Matthieu Guillaumin 和 Luc Van Gool。Food-101——利用随机森林挖掘判别性成分。载于欧洲计算机视觉会议，2014年。[12] Abdulkadir Canatar、Jenelle Feather、Albert Wakhloo 和 SueYeon Chung。神经预测与对齐的谱理论。*arXiv*预印本 *arXiv:2309.12821*，2023年。[13] 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。神经信息处理系统进展，33：9912–9924，2020年。[14] Ting Chen、Simon Kornblith、Mohammad Norouzi 和 Geoffrey Hinton。用于视觉表征对比学习的简单框架。载于国际机器学习会议，第1597–1607页。PMLR，2020年。[15] Xinlei Chen 和 Kaiming He。探索简单的孪生表征学习。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第15750–15758页，2021年。[16] Xinlei Chen、Haoqi Fan、Ross Girshick 和 Kaiming He。采用动量对比学习改进基线。*arXiv*预印本 *arXiv:2003.04297*，2020年。[17] SueYeon Chung 和 LF Abbott。神经群体几何：理解生物与人工神经网络的一种方法。《神经生物学当前观点》，70：137–144，2021年。[18] SueYeon Chung、Daniel D Lee 和 Haim Sompolinsky。一般感知流形的分类与几何结构。《物理评论*X*》，8(3)：031003，2018年。

[19] Mircea Cimpoi、Subhransu Maji、Iasonas Kokkinos、Sammy Mohamed 和 Andrea Vedaldi。在自然环境中描述纹理。载于*IEEE*计算机视觉与模式识别会议论文集，第 3606–3613 页，2014年。[20] Uri Cohen、SueYeon Chung、Daniel D Lee 和 Haim Sompolinsky。深度神经网络中对象流形的可分性与几何。《自然-通讯》，11(1):1–13，2020年。[21] Colin Conwell、Jacob S. Prince、Kendrick N. Kay、George A. Alvarez 和 Talia Konkle。18亿次回归能告诉我们，大脑和机器中塑造高级视觉表征的压力是什么？*bioRxiv*，2023年。doi: 10.1101/2022.03.28.485868。URL https://www.biorxiv.org/content/early/2023/07/01/2022.03.28.485868。[22] Thomas M Cover。线性不等式系统的几何与统计性质及其在模式识别中的应用。《*IEEE*电子计算机汇刊》，(3):326–334，1965年。[23] Victor Guilherme Turrisi da Costa、Enrico Fini、Moin Nabi、Nicu Sebe 和 Elisa Ricci。solo-learn：一个用于视觉表征学习的自监督方法库。*J. Mach. Learn.Res.*，23:56–1，2022年。[24] Joel Dapello、Jenelle Feather、Hang Le、Tiago Marques、David Cox、Josh McDermott、James J DiCarlo 和 SueYeon Chung。神经群体几何揭示随机性在鲁棒感知中的作用。神经信息处理系统进展，34:15595–15607，2021年。[25] James J DiCarlo 和 David D Cox。解析不变物体识别。《认知科学趋势》，11(8):333–341，2007年。[26] Debidatta Dwibedi、Yusuf Aytar、Jonathan Tompson、Pierre Sermanet 和 Andrew Zisserman。有朋友相助：通过最近邻对比学习获得视觉表征。载于*IEEE/CVF*国际计算机视觉大会论文集，第 9588–9597 页，2021年。[27] Eric Elmoznino 和 Michael F Bonner。高性能视觉皮层神经网络模型受益于高潜在维度。*bioRxiv*，2022年。[28] Aleksandr Ermolov、Aliaksandr Siarohin、Enver Sangineto 和 Nicu Sebe。用于自监督表征学习的白化。载于国际机器学习会议，第 3015–3024 页。PMLR，2021年。[29] Adrienne L Fairhall、Geoffrey D Lewen、William Bialek 和 Robert R de Ruyter van Steveninck。自适应神经编码的效率与歧义。《自然》，412(6849):787–792，2001年。[30] Jeremy Freeman、Corey M Ziemba、David J Heeger、Eero P Simoncelli 和 J Anthony Movshon。灵长类动物第二视觉区的功能与知觉特征。《自然-神经科学》，16(7):974–981，2013年。[31] Emmanouil Froudarakis、Uri Cohen、Maria Diamantaki、Edgar Y Walker、Jacob Reimer、Philipp Berens、Haim Sompolinsky 和 Andreas S Tolias。小鼠皮层视觉层级中的对象流形几何。*BioRxiv*，第 2020–08 页，2020年。[32] Stefano Fusi、Earl K Miller 和 Mattia Rigotti。神经元为何混合信息：高级认知所需的高维性。《神经生物学当前观点》，37:66–74，2016年。[33] Juan A Gallego、Matthew G Perich、Lee E Miller 和 Sara A Solla。用于运动控制的神经流形。《神经元》，94(5):978–984，2017年。[34] Elizabeth Gardner。神经网络模型中的相互作用空间。《物理学杂志*A*：数学与一般》，21(1):257，1988年。[35] Elizabeth Gardner 和 Bernard Derrida。神经网络模型的最优存储特性。《物理学杂志*A*：数学与一般》，21(1):271，1988年。[36] F Gerl 和 U Krey。通过腔室方法研究 Potts模型感知机的存储容量与最优学习。《物理学杂志*A*：数学与一般》，27(22):7353，1994年。

[37] Priya Goyal、Quentin Duval、Jeremy Reizenstein、Matthew Leavitt、Min Xu、Benjamin Lefaudeux、Mannat Singh、Vinicius Reis、Mathilde Caron、Piotr Bojanowski、Armand Joulin 和 Ishan Misra。Vissl。https://github.com/facebookresearch/vissl，2021年。[38] Jean-Bastien Grill、Florian Strub、Florent Altché、Corentin Tallec、Pierre Richemond、Elena Buchatskaya、Carl Doersch、Bernardo Avila Pires、Zhaohan Guo、Mohammad Gheshlaghi Azar 等。Bootstrap your own latent——一种新的自监督学习方法。神经信息处理系统进展，33:21271–21284，2020年。[39] Jeff Z HaoChen、Colin Wei、Adrien Gaidon 和 Tengyu Ma。谱对比损失为自监督深度学习提供可证明的保证。神经信息处理系统进展，34:5000–5011，2021年。[40] Kaiming He、张祥雨、任少卿和孙剑。用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议论文集，第 770–778 页，2016年。[41] 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。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 9729–9738 页，2020年。[42] O J Hénaff、N Rabinowitz、J Ballé 和 E P Simoncelli。自然图像的局部低维性。载于国际表征学习会议（*ICLR*），美国加利福尼亚州圣迭戈，2015年5月。URL http://arxiv.org/abs/1412.6626。[43] Olivier J Hénaff、Yoon Bai、Julie A Charlton、Ian Nauhaus、Eero P Simoncelli 和 Robbe LT Goris。初级视觉皮层使自然视频轨迹变得平直。《自然-通讯》，12 (1):1–12，2021年。[44] Bernd Illing、Jean Ventura、Guillaume Bellec 和 Wulfram Gerstner。局部可塑性规则可通过自监督对比预测学习深度表征。神经信息处理系统进展，34:30365–30379，2021年。[45] Li Jing、Pascal Vincent、Yann LeCun 和 Yuandong Tian。理解对比自监督学习中的维度坍塌。*arXiv*预印本 *arXiv:2110.09348*，2021年。[46] Diederik P Kingma 和 Jimmy Ba。Adam：一种随机优化方法。*arXiv*预印本 *arXiv:1412.6980*，2014年。[47] Andrew V Knyazev 和 Merico E Argentati。基于A的标量积中子空间之间的主角：算法与扰动估计。*SIAM*科学计算杂志，23(6):2008–2040，2002年。[48] Simon Laughlin。简单的编码过程可提升神经元的信息容量。*Zeitschrift für Naturforschung c*，36(9-10):910–912，1981年。[49] José Lezama、Qiang Qiu、Pablo Musé 和 Guillermo Sapiro。OLE：正交低秩嵌入——一种即插即用的深度学习几何损失。载于*IEEE*计算机视觉与模式识别会议论文集，第 8109–8118 页，2018年。[50] Yi Ma、Harm Derksen、Wei Hong 和 John Wright。通过有损数据编码与压缩对多变量混合数据进行分割。*IEEE*模式分析与机器智能汇刊，29(9):1546–1562，2007年。[51] Najib J Majaj、Ha Hong、Ethan A Solomon 和 James J DiCarlo。简单学习得到的加权下颞神经元放电率之和，可准确预测人类核心物体识别表现。神经科学杂志，35(39):13402–13418，2015年。[52] Jonathan Mamou、Hang Le、Miguel Del Rio、Cory Stephenson、Hanlin Tang、Yoon Kim 和 Sueyeon Chung。深度语言表征中可分流形的涌现。载于国际机器学习会议，第 6713–6723 页。PMLR，2020年。[53] Tiago Marques、Martin Schrimpf 和 James J DiCarlo。多尺度分层神经网络模型：连接灵长类初级视觉皮层中的单个神经元与物体识别行为。*bioRxiv*，第 2021–03 页，2021年。

[54] Albert W Marshall、Ingram Olkin 和 Barry C Arnold。《不等式：主序理论及其应用》，第143卷。施普林格，1979年。[55] Josue Nassar、Piotr Sokol、SueYeon Chung、Kenneth D Harris 和 Il Memming Park。关于 1/n 神经表征与鲁棒性。《神经信息处理系统进展》，33：6211–6222，2020年。[56] Edward H Nieh、Manuel Schottdorf、Nicolas W Freeman、Ryan J Low、Sam Lewallen、Sue Ann Koay、Lucas Pinto、Jeffrey L Gauthier、Carlos D Brody 和 David W Tank。海马体中抽象习得知识的几何结构。《自然》，595(7865)：80–84，2021年。[57] Maria-Elena Nilsback 和 Andrew Zisserman。大规模类别花卉的自动分类。载于《2008年第六届印度计算机视觉、图形学与图像处理会议》，第722–729页。IEEE，2008年。[58] Aaron van den Oord、Yazhe Li 和 Oriol Vinyals。基于对比预测编码的表征学习。arXiv 预印本 arXiv:1807.03748，2018年。[59]Serdar Ozsoy、Shadi Hamdan、Sercan Ö Arik、Deniz Yuret 和 Alper T Erdogan。基于信息最大化准则的自监督学习。arXiv 预印本 arXiv:2209.07999，2022年。[60] Nihaad Paraouty、Justin D Yao、Léo Varnet、Chi-Ning Chou、SueYeon Chung 和 Dan H Sanes。感觉皮层可塑性支持听觉社会学习。《自然-通讯》，14(1)：5828，2023年。[61] Nikhil Parthasarathy、SM Eslami、João Carreira 和 Olivier J Hénaff。自监督视频预训练可获得强大的图像表征。arXiv 预印本 arXiv:2210.06433，2022年。[62] Andrew Saxe、Stephanie Nelli 和 Christopher Summerfield。如果深度学习是答案，那么问题是什么？《自然综述：神经科学》，22(1)：55–67，2021年。[63] Rylan Schaeffer、Mikail Khona 和 Ila Fiete。神经科学中的深度学习并非免费的午餐：以内嗅皮层-海马环路模型为例。《神经信息处理系统进展》，35：16052–16067，2022年。[64] Martin Schrimpf、Jonas Kubilius、Ha Hong、Najib J. Majaj、Rishi Rajalingham、Elias B. Issa、Kohitij Kar、Pouya Bashivan、Jonathan Prescott-Roy、Franziska Geiger、Kailyn Schmidt、Daniel L. K. Yamins 和 James J. DiCarlo。Brain-score：哪种人工神经网络在物体识别方面最像大脑？bioRxiv 预印本，2018年。网址 https://www.biorxiv.org/ content/10.1101/407007v2。[65] Eero P Simoncelli 和 Bruno A Olshausen。自然图像统计与神经表征。《神经科学年度评论》，24(1)：1193–1216，2001年。[66] Cory Stephenson、Jenelle Feather、Suchismita Padhy、Oguz Elibol、Hanlin Tang、Josh McDermott 和 SueYeon Chung。剖析不变语音识别。《神经信息处理系统进展》，32，2019年。[67] Carsen Stringer、Marius Pachitariu、Nicholas Steinmetz、Matteo Carandini 和 Kenneth D Harris。视觉皮层群体反应的高维几何结构。《自然》，571(7765)：361–365，2019年。[68] Yonglong Tian、Dilip Krishnan 和 Phillip Isola。对比多视图编码。载于《欧洲计算机视觉会议》，第776–794页。施普林格，2020年。[69] Nenad Tomasev、Ioana Bica、Brian McWilliams、Lars Buesing、Razvan Pascanu、Charles Blundell 和 Jovana Mitrovic。突破自监督残差网络的极限：在 ImageNet 上，无标签时我们能否超越监督学习？arXiv 预印本 arXiv:2201.05119，2022年。[70] Greta Tuckute、Jenelle Feather、Dana Boebinger 和 Josh H McDermott。许多（但并非所有）深度神经网络音频模型都能捕捉脑反应，并呈现模型阶段与脑区之间的对应关系。bioRxiv，第2022–09页，2022年。

[71] Tongzhou Wang 和 Phillip Isola。通过超球面上的对齐与均匀性理解对比表征学习。载于《国际机器学习会议》，第9929–9939页。PMLR，2020年。[72] Yu Wang、Jingyang Lin、Qi Cai、Yingwei Pan、Ting Yao、Hongyang Chao 和 Tao Mei。用于无监督对比学习的低秩促进先验。*IEEE*模式分析与机器智能汇刊，2022年。[73] Laurenz Wiskott 和 Terrence J Sejnowski。慢特征分析：不变性的无监督学习。《神经计算》，14(4)：715–770，2002年。[74] Zhirong Wu、Yuanjun Xiong、Stella X Yu 和 Dahua Lin。通过非参数实例判别进行无监督特征学习。载于《IEEE计算机视觉与模式识别会议论文集》，第3733–3742页，2018年。[75] Justin D Yao、Klavdia O Zemlianova、David L Hocker、Cristina Savin、Christine M Constantino- ple、SueYeon Chung 和 Dan H Sanes。将声学信息转化为顶叶皮层中的感觉决策变量。《美国国家科学院院刊》，120 (2)：e2212120120，2023年。[76] Yaodong Yu、Kwan Ho Ryan Chan、Chong You、Chaobing Song 和 Yi Ma。通过最大编码率降低原理学习多样且具有判别性的表征。《神经信息处理系统进展》，33：9422–9434，2020年。[77] Jure Zbontar、Li Jing、Ishan Misra、Yann LeCun 和 Stéphane Deny。Barlow Twins：通过冗余降低进行自监督学习。载于《国际机器学习会议》，第12310–12320页。PMLR，2021年。

一个最优嵌入

回顾 Balestriero 和 LeCun（6）所描述的自监督学习设定：给定一个数据集 $\boldsymbol{X'} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_N]^T \in \mathbb{R}^{N \times D'}$，我们通过从原始数据中随机创建 k 个增强视图来构建一个新数据集，$\boldsymbol{X} = [\text{视图}_1(\boldsymbol{X'}) ..., \text{视图}_k(\boldsymbol{X'})] \in \mathbb{R}^{Nk \times D}$。这样做的优点是，我们现在可以利用同一底层数据点的不同视图在语义上相关这一知识。我们可以用对称矩阵 $\boldsymbol{G} \in \{0,1\}^{Nk \times Nk}$ 表示这种相似性概念：如果增强后的数据点 i 和 j 在语义上相关，则令 $\boldsymbol{G}_{ij} = 1$（并且$\boldsymbol{G}_{ii} = 1$，因为任意数据点都与自身相关）。我们可以对 $\boldsymbol{G}$ 进行归一化，使其行和列的和均为 1（因此 $\boldsymbol{G}$ 的各行是 k 稀疏的，非零项等于 $1/k$）。

现在令 $\boldsymbol{Z} \in \mathbb{R}^{Nk \times d}$ 为增强数据集的一个嵌入。那么我们有 $\boldsymbol{GZ} = [\boldsymbol{C}, ..., \boldsymbol{C}]^T$，其中 $\boldsymbol{C}$ 是上文引入的质心向量矩阵，而 $\boldsymbol{C}$ 的重复次数为 k。由于 $\boldsymbol{\sigma}([\boldsymbol{C}, ..., \boldsymbol{C}]) = \sqrt{k}\boldsymbol{\sigma}(\boldsymbol{C})$，因此我们可以将 MMCR 损失函数写为：

$$\begin{aligned} \mathcal{L} &= -||\boldsymbol{GZ}||_* \\ &= -||\boldsymbol{Q\Lambda Q}^T\boldsymbol{USV}^T||_* \\ &= -||\boldsymbol{\Lambda Q}^T\boldsymbol{US}||_* \end{aligned} \tag{7}$$

其中，我们对实对称矩阵 $\boldsymbol{G}$ 进行了特征值分解，并对$\boldsymbol{Z}$进行了奇异值分解，随后利用了奇异值谱在左乘或右乘正交变换下保持不变这一事实。现在我们将证明，当 $\boldsymbol{Z}$ 的左奇异向量是 $\boldsymbol{G}$ 的特征向量，且 $\boldsymbol{Z}$ 的奇异值与 $\boldsymbol{G}$的特征值成比例时，该目标函数可达到全局最优解。在下文中，我们始终假设数据集的大小大于嵌入的维数，即 $N > d$，实际应用中通常如此。首先，我们证明一个关于通过补零扩展的矩阵（即嵌入更高维空间中的矩阵）的谱的简单引理。

引理 A.1 对于 $\mathbb{R}^{N \times N}$，其中，且 $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ $\boldsymbol{B} \in \mathbb{R}^{N \times d}$ $d < N$ $||\boldsymbol{AB}||_* = ||\boldsymbol{A}\tilde{\boldsymbol{B}}||_*$ $\tilde{\boldsymbol{B}} = [\boldsymbol{B}, \boldsymbol{0}] \in$：

证明：首先注意到 $\boldsymbol{AB} = [\boldsymbol{AB}, \boldsymbol{0}]$，因此只需证明，对于任意 $\boldsymbol{X}$，有 $\boldsymbol{\sigma}(X) = \boldsymbol{\sigma}([\boldsymbol{X}, \boldsymbol{0}])$。对 $\boldsymbol{X}$ 进行奇异值分解，

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{U} & \tilde{\boldsymbol{U}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma} \\ \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{V}^T \end{bmatrix} = \boldsymbol{U\Sigma V}^T$$

因此，X 的一个有效奇异值分解为

$$\tilde{\boldsymbol{X}} = \begin{bmatrix} \boldsymbol{U} & \tilde{\boldsymbol{U}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{V}^T & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix}$$

显然，$||\boldsymbol{X}||_* = ||\boldsymbol{X}||_*$

定理： 当 $\boldsymbol{Z}$ 的左奇异向量为 $\boldsymbol{G}$的特征向量，且 $\boldsymbol{Z}$ 的奇异值与 $\boldsymbol{G}$的前 d 个特征值成正比时，所提出的损失达到全局最小值。

证明： 令 $\tilde{\boldsymbol{Z}} = [\boldsymbol{Z}, \boldsymbol{0}] \in \mathbb{R}^{N \times N}$。根据引理 A.1，我们有 $||\boldsymbol{GZ}||_* = ||\boldsymbol{G}\tilde{\boldsymbol{Z}}||_*$。可以使用冯·诺依曼迹不等式证明 $||\boldsymbol{G}\tilde{\boldsymbol{Z}}||_* \leq \sum_{i=1}^{Nk} \boldsymbol{\sigma}_i(\boldsymbol{G})\boldsymbol{\sigma}_i(\tilde{\boldsymbol{Z}})$（证明见 Marshall 等人 (54)）。考察 (4) 可知，当 $\boldsymbol{U} = \boldsymbol{Q}$ 时，该界可以达到。因此，该问题可归结为约束优化问题：

$$\begin{aligned} &\min_{\boldsymbol{\sigma}_i(\tilde{\boldsymbol{Z}})} \sum_{i=1}^{Nk} \boldsymbol{\sigma}_i(\boldsymbol{G})\boldsymbol{\sigma}_i(\tilde{\boldsymbol{Z}}) \\ &\text{subject to} \sum_{i=1}^{Nk} \boldsymbol{\sigma}_i(\tilde{\boldsymbol{Z}})^2 = Nk \end{aligned}$$

其中的约束源于 $\boldsymbol{Z}$ 的列是单位向量这一事实。直观地说，我们是在最大化固定向量 $\boldsymbol{\sigma}(\boldsymbol{G})$ 与一个 L2 范数固定的向量之间的内积。

˜当然，解法是尽可能使这两个向量对齐，也就是说，当$i = 1, ..., d$时，$\boldsymbol{\sigma}_i(\boldsymbol{Z}) \propto \boldsymbol{\sigma}_i(\boldsymbol{G})$。值得注意的是，根据构造，对于 $i > d$，$\boldsymbol{\sigma}_i(\boldsymbol{Z}) = 0$，而与这些零奇异值对应的 $\boldsymbol{U}$ 的列不受约束。

B 用于MMCR的PyTorch风格伪代码

```
# h: 编码器
# g: 投影头
# T: 动量温度
# B: 批量大小
# K: 数据增强的数量
# D: 投影器输出维数 #
# lmbda: 权衡参数
f_o, g_o = ResNet50(), MLP() # 在线网络
f_t, g_t = ResNet50(), MLP() # 使用相同参数初始化动量网络
```

C 平均场理论流形容量背景信息

为完整起见，我们总结 Chung等人 (18) 提出的一些核心论证，他们发展了流形容量理论的一般形式。

平均场理论 回顾流形容量分析的问题设定：给定一组嵌入在维数为 D 的特征空间中的 P 个流形，每个流形都被赋予一个随机二元类别标签 (18)。流形容量理论关注的问题是：$\frac{P}{D}$ 的最大值是多少，才能使一个超平面以高概率将两个类别分开？在热力学极限下，其中 P, $D \to \infty$，但 $\frac{P}{D}$ 保持有限，逆容量可精确写为，

$$\alpha_M^{-1} = \mathbb{E}_{\vec{T}}[F(\vec{T})] \tag{8}$$

其中，$F(\vec{T}) = \min_{\vec{V}} \left\{ \|\vec{V} - \vec{T}\|^2 \mid g_{\mathcal{S}}(\vec{V}) \geq 0 \right\}$，$\mathcal{S}$是定义流形几何的集合（即$\vec{T}$构成单个流形上各点的向量集合），$T$是从白色多元高斯分布中抽取的随机向量，而 $g_{\mathcal{S}}(\vec{V}) = \min_{\vec{S}} \{\vec{V} \cdot \vec{S} \mid \vec{S} \in \mathcal{S}\}$ 是凹支撑函数。

该凸优化问题的 KKT 方程为：

$$\begin{aligned} \vec{V} - \vec{T} - \lambda \tilde{S}(\vec{T}) &= 0 \\ \lambda &\geq 0 \\ g_{\mathcal{S}}(\vec{V}) - \kappa &\geq 0 \\ \lambda \left[g_{\mathcal{S}}(\vec{V}) - \kappa \right] &= 0. \end{aligned} \tag{9}$$

其中，$\tilde{S}(\vec{T})$ 是支撑函数的次梯度。当支撑函数可微时，次梯度是唯一的，并且等于梯度，

$$\tilde{S}(\vec{T}) = \nabla g_{\mathcal{S}}(\vec{V}) = \arg\min_{\vec{S} \in \mathcal{S}} \vec{V} \cdot \vec{S} \tag{10}$$

$\tilde{S}(\vec{T})$ 是满足第一个 KKT 方程的 $\mathcal{S}$ 凸包中的唯一点，称为由随机向量 T 诱导的 $\mathcal{S}$ 的“锚点”（$\vec{}$the “anchor point”）。

锚点的等价解释对于给定的二分分类（随机二元类别标记），最大间隔分离超平面的权重向量可分解为至多 P 个向量之和，每个流形贡献一个向量，该向量位于该流形的凸包内。该点的位置是流形相对于空间中所有其他流形的位置的函数，并取决于具体的随机标签集合。因此，对于每个独立流形，都存在一个由分离超平面决定的点的分布。利用腔室方法可以证明，这些点正是参与求解上述优化问题（36）的锚点。

数值求解 为了数值求解平均场方程，先抽取若干随机$\vec{}$ $\vec{}$ $\vec{}$ $\vec{}$高斯向量 T，然后针对每个 T，通过求解上文给出的二次规划程序，确定 V 和 S。随后，将 F 或样本 $\vec{}T$ 的均值作为容量的估计值。

流形几何 通过引入两个关键量——流形半径 R_M 和流形维数R_M——可以简化对容量如何随锚点统计特性变化的描述：

$$\begin{aligned} R_M^2 &= \mathbb{E}_{\vec{T}}[||\tilde{S}(\vec{T})||^2] \\ D_M &= \mathbb{E}_{\vec{T}}[\vec{T} \cdot \hat{S}(\vec{T})] \end{aligned} \tag{11}$$

其中，$\hat{}S(\vec{T})$ 是沿锚点 $\tilde{S}$ 方向的单位向量。特别是，如正文所述，流形容量可以近似表示为 $\phi(R_M\sqrt{D_M})$，其中 ϕ 是一个单调递减函数。

椭圆几何 当流形呈现椭圆对称性时，流形半径和维数可以用锚点协方差矩阵的特征值表示：

$$
\begin{aligned}
R_M^2 &= \sum_i \lambda_i^2 \\
D_M &= \frac{\left(\sum_i \lambda_i\right)^2}{\sum_i \lambda_i^2}
\end{aligned}
\tag{12}
$$

因此，在这种情况下，R_M 是锚点的总变异性，而 D_M 是锚点协方差的广义参与率，是一种广为人知的维数软度量。

D 额外预训练信息

CIFAR/STL-10 的设置我们直接采用 Zbontar 等人（77）给出的各项数据增强参数，但对于这些较低分辨率的图像，我们省略了高斯模糊和日晒增强。所有模型均使用 Adam 优化器（46）训练 500 轮，学习率为 1e -3，权重衰减为 1e -6。对于这三种方法，我们在投影头中使用了一个隐藏层的多层感知机，隐藏维度为 512，输出维度为 128 g。我们针对每种方法遍历不同的批量大小，并选择下游任务表现最佳的设置。对于 SimCLR 和 Barlow Twins，在所有 3 个数据集上，我们发现批量大小为 128（候选值为 32、64、128、256 和 512）时效果最佳。对于 MMCR，批量大小与所用数据增强数量之间存在权衡，而最佳权衡值高度依赖于数据集。对于 CIFAR-10 和 CIFAR-100，我们使用批量大小 32 和 40 个视图；对于 STL-10，我们使用批量大小 64 和 20 个视图。对于 Barlow Twins，我们使用了 $\lambda = \frac{1}{128}$，它会根据损失中的对角项和非对角项元素数量进行归一化。对于 SimCLR，我们采用了推荐设置 $\tau = 0.5$。通过更全面的超参数搜索，并采用更贴近原始工作的方案，两个基线方法（MMCR 可能也一样）的整体表现都能得到提升。例如，两种方法可能都会受益于以下做法：增大批量大小、使用 LARS 优化器（专为大批量优化而设计）、采用先线性预热再进行余弦退火的学习率调度器、延长训练时间，以及使用更多样的数据增强（例如加入日晒增强和高斯模糊）。此外，Barlow Twins 指出，使用远大于我们所用规模的投影网络可能有助于提升表征。我们的主要目标是证明 MMCR 能够生成可与这些基线相媲美的表征，而不是在小规模数据集上取得最先进的结果，因此我们尽可能采用了简化方案（使用开箱即用的 Adam 优化器进行优化并固定学习率，同时固定投影维数等架构超参数）。

ImageNet-100 的设置 对于 ImageNet，我们更 closely 遵循以往工作的预训练流程。我们使用 2048 的批量大小，并为 MMCR 使用较少的视图数量（4），同时采用 Zbontar 等人（77）提出的完整数据增强方案。为提高效率，我们减少了训练轮数（200）。对于 MMCR 和 SimCLR，我们按照原始工作（14）的做法，将投影头的投影器隐藏维度设为 4096。对于 Barlow Twins，我们使用了推荐的具有 8192 隐藏维度和输出维度的 2 层多层感知机，并设置了$\lambda = 5e-3$，但这些超参数针对完整的 ImageNet 数据集进行了优化，不一定适用于 ImageNet-100。与文献中此前针对该数据集使用 ResNet-18 骨干网络报告的结果相比，我们无法通过使用 ResNet-50 骨干网络获得更好的下游性能，因此报告（23）中给出的 ResNet-18 性能。对于 SimCLR，我们使用 $\tau = 0.1$，这是较大批量大小下的推荐设置。

ImageNet-1k 的设置：对于 ImageNet-1k，我们采用了与 ImageNet-100 基本相同的设置，但提高了投影网络的容量（使用一个具有 2 个隐藏层、隐藏维度为 8192、输出维度为 512 的多层感知机）。我们根据批量大小线性调整学习率：$\mathrm{lr} = 0.6 \times \frac{\text{批量大小}}{256}$。此外，我们将预训练周期减少到 100。最后，对于 ImageNet-1k，我们发现使用动量编码器可以略微提升下游性能（在 ImageNet 上进行冻结线性评估时约提升 +1%）。具体而言，我们使用与初始“在线”网络相同的参数初始化一个结构完全相同的编码网络和投影器架构；训练期间，该“动量”网络的权重会跟随在线网络参数缓慢移动的平均值变化（只有在线网络参数会通过梯度

下降进行更新）。每个增强视图都会输入在线网络和动量网络，随后将得到的嵌入取平均，形成批次中相应图像的质心向量。我们使用了 0.99 的动量系数。

预训练使用 16 个 A100 GPU 和 8 个视图（我们计算量最大的设置）大约需要 32 小时。

E 表征分析详情

E.1 流形容量分析

对于每个预训练模型，我们对一组图像执行前向传播后，提取 ResNet 层级结构中的层激活。对于类别流形分析，这组图像包含10个类别，每个类别有100个样本。增强流形则包含100个样本，每个样本有100个示例。遵循（20），我们提取 ResNet-50 中所有卷积层经过 ReLU 非线性变换后的激活。以粗体标出的具体提取层见表3。最终分析结果是在5次使用不同随机种子的数据采样，以及将中间特征随机投影到低维空间（默认5000维）后取平均得到的。

E.2 梯度相干性分析

在图3中，对于 CIFAR-10 的每个类别，我们生成100个批次，每个批次包含来自某一特定类别的32个样本增强流形（每个样本包含40次数据增强）。随后，我们在训练的不同阶段测量每个批次的损失函数梯度，并计算每一对梯度之间的余弦相似度。在训练的所有阶段，由同类别批次生成的梯度之间的平均余弦相似度，都大于由不同类别批次生成的梯度之间的平均余弦相似度（左列）。当分别分析 ResNet-50 层级结构中不同阶段参数的梯度时，这一观察结果仍然成立（分别见中列和右列）。

E.3 流形子空间对齐

对于图4，我们从 CIFAR-10 数据集中 500 张图像的增强流形中生成了 100 个样本。随后，我们测量每对流形之间的平均子空间角（左列）、共享方差比例（中列）以及质心余弦相似度。生成图5数据时也采用了相同的步骤。

子空间角。 除了测量各个对象流形的大小和维数外，我们还希望刻画流形对之间的重叠程度。为此，我们测量它们的子空间之间的夹角（47）；这是角度概念的一种推广，适用于任意维数的子空间。

共享方差。 对象流形的内在维度通常低于其嵌入空间的维数。因此，用于计算子空间角的一组主向量中，沿若干向量方向的数据方差较低，导致许多主角缺乏实际意义。为了解决这一局限，我们还计算了包含对象流形的线性子空间之间的共享方差。

F 隐式 MMCR 可有效降低增强流形核范数

为了测试隐式流形压缩是否确实降低了平均增强流形核范数，我们可以改变 λ 的值。下面展示了在 CIFAR-10 上训练期间，取若干不同 lambda 值时损失中两项的变化过程。在这些实验中，批量大小为 64，每张图像的增强次数为 4.0。如图 6 所示，在所有参数值下，单个流形的压缩程度几乎相同。

表 3：用于 MFTMA 分析的 18 个 ResNet-50 提取层（**粗体**）

层	Type		二维卷积尺寸（H × W × C）
像素	**输入**		None
conv1	$\begin{bmatrix} [l]\text{二维卷积 批归} \\ \text{一化 } \mathbf{ReLU} \end{bmatrix} \times 1$		$[[l]7 \times 7 \times 64] \times 1$
二维卷积 2 x_	$\begin{bmatrix} [l] \begin{bmatrix} [l]\text{二维卷积 批归一} \\ \text{化} \end{bmatrix} \times 3 \ \mathbf{ReLU} \end{bmatrix}$	$\times 3$	$\begin{bmatrix} [l]1 \times 1 \times 64 \\ 3 \times 3 \times 64 \\ 1 \times 1 \times 256 \end{bmatrix} \times 3$
二维卷积 3 x_	$\begin{bmatrix} [l] \begin{bmatrix} [l]\text{二维卷积 批归一} \\ \text{化} \end{bmatrix} \times 3 \ \mathbf{ReLU} \end{bmatrix}$	$\times 4$	$\begin{bmatrix} [l]1 \times 1 \times 128 \\ 3 \times 3 \times 128 \\ 1 \times 1 \times 512 \end{bmatrix} \times 4$
二维卷积 4 x_	$\begin{bmatrix} [l] \begin{bmatrix} [l]\text{二维卷积 批归一} \\ \text{化} \end{bmatrix} \times 3 \ \mathbf{ReLU} \end{bmatrix}$	$\times 6$	$\begin{bmatrix} [l]1 \times 1 \times 256 \\ 3 \times 3 \times 256 \\ 1 \times 1 \times 1024 \end{bmatrix} \times 6$
第 5 个卷积层 x_	$\begin{bmatrix} [l] \begin{bmatrix} [l]\text{二维卷积 批归} \\ \text{一化} \end{bmatrix} \times 3 \ \mathbf{ReLU} \end{bmatrix}$	$\times 3$	$\begin{bmatrix} [l]1 \times 1 \times 512 \\ 3 \times 3 \times 512 \\ 1 \times 1 \times 2048 \end{bmatrix} \times 3$

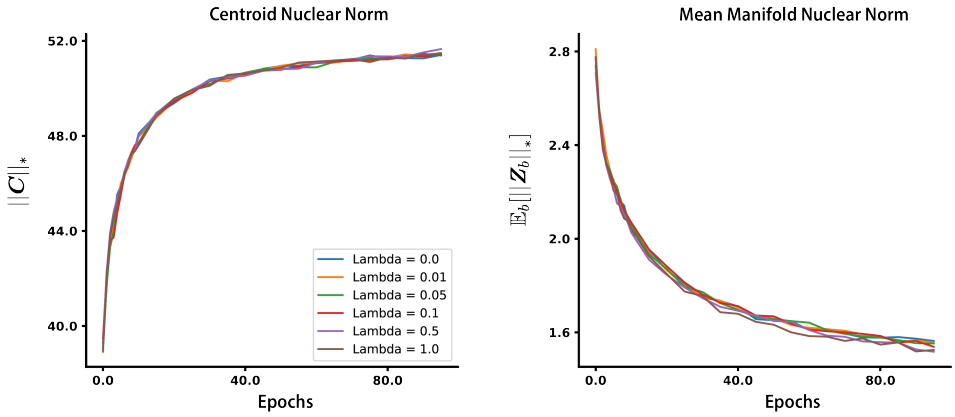

图 6：不同 λ取值对应的验证损失值

G 分类评估流程

CIFAR 和 STL-10：在预训练期间，我们使用 k近邻分类器（k=200）监控所有模型，并每 5 个训练轮数保存一次检查点。预训练结束后，我们在所有监控准确率处于观测到的最高准确率 1% 范围内的检查点上训练线性分类器，并选择线性分类准确率最高的模型。线性分类器使用 Adam 优化器进行训练，批量大小为 1024，初始学习率为 0.1，并在 50 个训练轮数内按照余弦调度器进行衰减。在线性分类器训练期间，训练时使用与无监督预训练期间相同的数据增强集合；测试时仅使用中心裁剪和随机水平翻转。

ImageNet-1k/100：对于 ImageNet 数据集，我们严格遵循最广泛采用的评估流程。预训练后，我们冻结编码器权重，并使用随机梯度下降，以监督方式训练一个线性层；批量大小为 2048，学习率为 1.6，权重衰减为 1e-6，共训练 50 轮。训练线性分类器期间，唯一使用的数据增强是随机裁剪和随机水平翻转；评估时则对输入进行中心裁剪。
半监督：对于半监督评估，我们主要遵循 Bardes 等人（7）所述的流程。我们使用随机梯度下降优化器，动量为 0.9，权重衰减为 1e-6，并采用标准交叉熵损失。数据增强流程与上述相同。由于线性分类器从头开始训练，且表征需要微调，因此骨干网络参数的学习率缩小为原来的十分之一；骨干网络和分类器的学习率均采用余弦衰减计划，训练 20 轮。我们使用 256 的批量大小，并针对每个模型在 [0.1、0.3、1.0] 范围内遍历学习率。

其他下游分类任务：这些数据集上的分类器采用与 CIFAR 和 STL-10 数据集上相似的方式进行训练。唯一的区别在于，对于每种方法和数据集，我们将批量大小遍历 [128, 256, 512, 1024] ，并将初始学习率遍历 [3e-2、3e-3、3e-4], , 且数据增强流程与标准的 ImageNet 训练设置一致（训练时仅进行随机裁剪和水平翻转，评估时进行缩放和中心裁剪）。

H 训练指标

在下方的图 7 中，我们监测训练过程中目标函数（第二个面板）、平均增强流形核范数、质心范数以及在测试集上评估的平均质心相似性的变化。

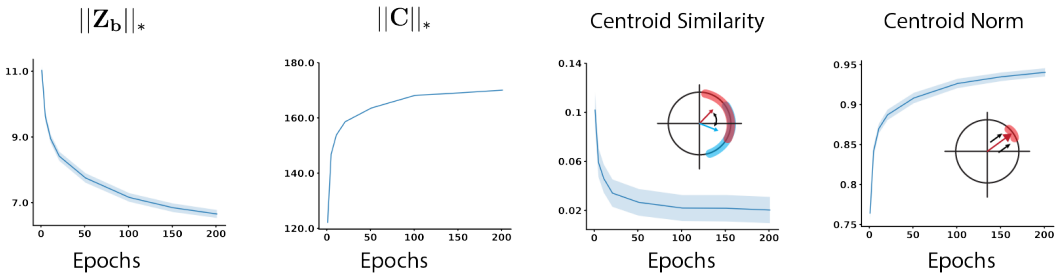

图 7：训练期间各种度量指标的变化。几何度量指标在由 CIFAR-10 数据集中的一幅图像及其 16 种数据增强共同定义的 200 个流形上进行评估。阴影区域表示均值周围 95% 的置信区间。

I 较小数据集上的分类性能

在下面的表 4 中，我们报告了在较小数据集上使用 ResNet-50 骨干网络训练时，我们的方法以及 Barlow Twins 和 SimCLR 的性能。

J 批量大小依赖性

对比式自监督学习方法最常被指出的缺点之一是，要在下游任务中取得良好表现，就需要使用较大的批量大小进行训练；而对嵌入的互相关矩阵/协方差矩阵施加约束的非对比式方法（例如 VICReg 或 Barlow Twins（77；7））则更适合使用较小批次进行训练。还值得注意的是，可以通过多种方式缓解对比方法对大批量大小的需求，例如维护一个记忆库（74），或使用缓慢更新的动量编码器（41）。鉴于我们的方法既不完全属于对比式，也不完全属于非对比式（因为它直接作用于嵌入矩阵的谱），我们想知道其表现会如何随训练批量大小而变化。我们使用批量大小为 256、512、1024、2048、4096 的设置在 ImageNet-1k 上进行预训练，并评估线性分类

表 4：使用各种数据集和目标函数训练的表征的线性分类器 Top-1 分类准确率。注：对于 ImageNet-100 上的 Barlow Twins，我们报告 da Costa 等人 (23) 的结果，该结果使用 ResNet-18 骨干网络，因为我们未能获得更好的性能。对于 ImageNet-100 上的 MMCR，我们测试了 2 个视图（与基线一致）和 4 个视图，结果按（2 个视图）/（4 个视图）的格式呈现

方法	CIFAR-10	CIFAR-100	STL-10	ImageNet-100
Barlow Twins（我们的复现）	90.91	67.91	89.96	80.38*
SimCLR（我们的复现）	92.22	70.04	**91.11**	79.64
MMCR（$\lambda = 0.0$）	**93.53**	69.87	90.62	81.52/**82.88**
MMCR（$\lambda = 0.01$）	93.39	**70.94**	90.77	81.28/82.56

准确率。令人鼓舞的是，我们观察到，即使是所测试的最小批量大小，性能也仅略有下降。图 8 展示了这次扫描的结果，并将其与 Barlow Twins 和 SimCLR 进行了比较。请注意，这些实验使用了两个视图，并采用了附录 D 中所述的线性学习率缩放方法。未来的工作应致力于更好地理解各种超参数对学习到的表征质量的影响。

一个重要细节是，在这项实验中，我们训练 MMCR 时没有使用动量编码器。He 等人（41）指出，动量编码器会增大有效小批量大小，因为缓慢变化的权重会编码来自先前批次的信息。然而，我们在此明确关注对批次大小的依赖性，因此消除了这一架构因素的干扰（Zbontar 等人（77）和 Chen 等人（14）均未使用动量编码器）。

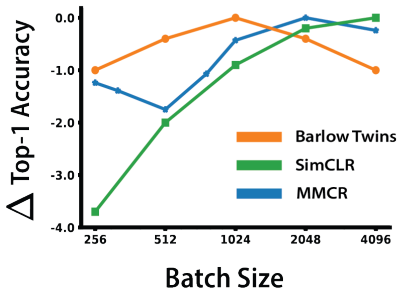

图 8：三种方法的 Top-1 性能相对于各自最佳设置的下降幅度。Barlow Twins 和 SimCLR 的数据均摘自 Zbontar 等人（77）。

K Brain-Score 的其他详细信息

Brain-Score 根据模型预测神经元对图像的实测反应的能力来评估模型。这里简要介绍该指标（完整说明请参见 (64)）。设 $\boldsymbol{y} \in \mathbb{R}^N$ 表示单个（生物）神经元对一组 N 训练图像的平均反应，$\boldsymbol{X} \in \mathbb{R}^{N \times K}$ 表示 K 模型神经元对相同图像的反应。Brain Score 首先求解线性回归问题 $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w}$，以得到权重 $\boldsymbol{w}$。随后，模型预测其对一组留出图像的反应 $\boldsymbol{y}'$。接着计算 $\boldsymbol{y}'$ 与 y 之间的皮尔逊相关系数，特定数据集的得分则为该数据集中各个神经元预测能力的中位数。然后，对每个数据集的不同训练-测试划分所得得分取平均值。整个脑区的得分（如表2所示）是对训练-测试划分及不同数据集的得分取平均值。对每个数据集（通常介于 1e-3 和 1e-2 之间）的均值标准误，在同一脑区内按平方和开方后求和，再除以数据集数量，以得出表2中报告的误差。在表 5 中，我们将脑区得分拆分为各个数据集的得分。

模型	V1.0	V1.1	V2.0	V4.0	V4.1	V4.2
MMCR	0.270	0.718	0.311	0.577	0.627	0.492
SimCLR	0.224	0.776	0.288	0.576	0.626	0.48
BYOL	0.274	0.727	0.291	0.585	0.626	0.48
MoCo	0.273	.726	0.293	0.57	0.629	0.492
Barlow	0.276	.721	0.293	0.568	0.626	0.493
SwAV	0.252	.723	0.296	0.568	0.614	0.469
模型	**V4.3**	**IT.0**	**IT.1**	**IT.2**	**IT.3**	
MMCR	0.226	0.554	0.558	0.545	0.424	
SimCLR	0.224	0.552	0.545	0.518	0.456	
BYOL	0.216	0.55	0.545	0.516	0.41	
MoCo	0.215	0.54	0.560	0.550	0.437	
Barlow	0.221	0.545	0.547	0.518	0.412	
SwAV	0.202	0.533	0.537	0.518	0.405	

表 5：六种自监督模型在 11 个不同的、采自猕猴的电生理数据集上的 Brain-score 比较 (64)。数据集 V1.0 和 V2.0 来自 Freeman 等人 (30)，V1.1 来自 Marques 等人 (53)，V4.0、V4.1、IT.0 和 IT.1 来自 Majaj 等人 (51)。(V4/IT).(2/3) 是 brainscore 所标注的 SanghaviJozwik2020 和 SanghaviMurty2020 数据集。

L 关于谱性质的更多细节

参与率 我们首先提取了各模型对 ImageNet 验证集图像的响应所对应的 2048 维特征向量。按照测试分类器时所采用的设置，将图像调整为 256×256，然后进行中心裁剪至 224×224。对于每个模型，我们对得到的特征矩阵进行 10 次独立的有放回重采样。对于每个重采样得到的特征数据集，我们计算经验协方差矩阵、相应的特征值谱以及相应的参与率（ L_1 与特征向量 L_2 范数之比的平方）。

衰减系数 使用上述步骤得到的谱在尾部（最不显著的特征值附近）均迅速衰减。为避免在估计衰减系数时受到这些尾部的偏差影响，我们仅考虑排名前 2000 的特征值。为了估计衰减系数，我们对特征值的对数与其索引的对数进行线性回归，拟合出一条回归线。我们针对每个模型的自助法谱重复这一拟合过程，以获得均值的标准误。对于所有模型，线性回归都与数据高度吻合，观测到的最小 R^2 值为 0.95。

M 目标检测

为确保通过 MMCR 获得的表征不会过度专用于分类任务，我们还在目标检测任务上评估了性能最高的模型。我们遵循文献（41；77）的方法，使用 Faster R-CNN 检测头和 C-4 主干网络，在 VOC07+12数据集上对表征网络进行微调（使用 train+val 划分进行训练，并在 VOC07 测试集上评估）。除初始学习率（设为 0.12 ）外，所有设置均与 He 等人（ 41 ）的设置相同；此外，我们也使用 detectron2 库进行评估。使用 8 个视图并进行 100 轮预训练的 MMCR，其 AP50 达到 81.9（这是三次独立微调的平均值，标准差为 0.2），这表明我们的方法生成的表征并不局限于物体识别；完整结果见表 **??**。作为参考，使用 MoCo v2 训练 200 轮得到的表征，其 AP50（该数据集最常用的评估指标）为 82.4（为保持一致性，下表报告的是我们对 MoCo 进行 100 轮预训练所获得的结果）。

N 局限性

受时间和计算资源限制，我们无法对所有设计选择进行全面优化。例如，我们没有探究投影网络宽度和深度变化的影响。此外，我们将模型限定为使用在 ImageNet-1k 上预训练的 ResNet-50 编码网络

方法	平均精度均值 AP50		AP75
Barlow	53.1	80.9	57.7
SwAV	54.4	81.6	61.0
MMCR	54.6	81.9	60.6
SimCLR	54.7	81.7	60.2
MoCo v2	55.6	**82.3**	61.7
BYOL	**56.0**	**82.3**	**62.0**

表 6：我们遵循（41；77），使用 Faster R-CNN 检测头和 C-4 主干网络，在 VOC07+12 数据集上对表征网络进行微调。

数据集训练 100 个训练轮数。这一选择使我们能够在适度的计算预算下，与近期开发的几种自监督学习方法进行公平比较，但无法回答该方法如何扩展到更大规模问题这一问题。

我们还注意到，尽管与现有方法相比，我们的目标函数具有较低的计算复杂度，但其评估过程并不容易在多台机器上进行分布式计算。这是因为我们需要对质心矩阵进行奇异值分解，这要求将分布式前向传播的输出汇集到一台机器上。未来，针对奇异值分解的高效分布式计算方法或许有助于缓解这一问题。

动量编码器的影响

在其他设置均保持不变的情况下，我们发现，使用动量编码器在分类性能方面略有优势。下文报告了 MMCR 使用 2、4 和 8 个视图进行训练时，有无动量编码器对应的冻结线性评估准确率。

	2 个视图	4 个视图	8 个视图
使用动量编码器	69.5	71.4	72.1
不使用动量编码器	68.4	70.2	71.5

表 7：展示了使用上述设置、通过动量编码器或不使用动量编码器训练得到的表征在线性评估协议下的 Top-1准确率。可以看出，随着视图数量增加，动量编码器带来的正样本表征多样性增益会逐渐减弱。

P MMCR受益于更长时间的预训练

为验证MMCR能从更长时间的预训练中受益，我们调整了设置，使用两个视图进行预训练，将基础学习率降至0.4，并训练1000轮。此外，在评估该模型时，我们还调整了预训练线性分类器的超参数（分类器以0.3的学习率和2048的批量大小训练100轮）。我们没有针对这一设置进行任何超参数调优，因此这些结果代表了双视图MMCR的性能下限。

方法	Top-1 准确率
SimCLR	69.3
MoCov2	71.1
SimSiam	71.3
SwAV（不使用多裁剪）	71.3
MMCR（2 个视图）	73.0
Barlow Twins	73.2
BYOL	74.3
SwAV（使用多裁剪）	75.3
NNCLR	75.6
ReLICv2（69）	**77.1**

表8：展示了使用各种自监督学习框架训练 1000 训练轮数后，表征在线性评估协议下的 Top-1准确率。