具有一致表示的对比学习

Zihu Wang *zihu_wang@ucsb.edu*
加州大学圣塔芭芭拉分校电气与计算机工程系

Yu Wang *yu95@ucsb.edu*
电气与计算机工程系 加州大学圣塔芭芭拉分校

Zhuotong Chen *ztchen@ucsb.edu*
电气与计算机工程系 加州大学圣塔芭芭拉分校

Hanbin Hu *hanbinhu@ucsb.edu*
电气与计算机工程系 加州大学圣塔芭芭拉分校

Peng Li *lip@ucsb.edu*
电气与计算机工程系 加州大学圣塔芭芭拉分校

arXiv:2302.01541v2 [cs.CV] 4 Sep 2024

摘要

对比学习在表征学习方面展现出巨大潜力。数据增强通过提供数据的有效视图，无需显式标签，在对比学习中发挥着关键作用。尽管如此，当前方法的效果在很大程度上取决于所采用的数据增强（DA）函数的质量，而这些函数往往是从有限选项中手动选择的。尽管利用多样化的数据增强颇具吸引力，但数据增强和表征学习本身的复杂性可能导致性能下降。为应对这一挑战并促进多样化数据增强的系统性整合，本文提出了具有一致表征的对比学习（CoCor）。 CoCor 的核心是一种新颖的一致性度量，称为数据增强一致性。该度量控制增强输入数据向表征空间的映射，确保这些样本根据所施加的数据增强强度，以一致的方式处于最优位置。此外，我们提出将最优映射位置学习为数据增强的函数，同时保持其相对于数据增强强度所期望的单调性。实验结果表明，与基线方法相比， CoCor 显著提升了所学习表征的泛化能力和迁移能力。 CoCor 的实现见 https://github.com/zihuwang97/CoCor。

1 引言

数据增强（DA）广泛应用于计算机视觉中的监督学习 Ho 等（2019）； Lim 等（2019）； Cubuk 等（2019; 2020）； Li 和 Li（2023），并在常用数据集 Ciregan 等（2012）； Sato 等（2015）； Wan 等（2013）； Krizhevsky 等人（2017）上取得了出色的成果。数据增强也是近期对比学习技术的重要组成部分 Chen 等（2020a）； Tian 等（2020b）； He 等（2020）； Chen 和 He（2021）； Xiao 等人（2020）； Lee 和 Shin（2023）。通过对比损失训练编码器，使其能够学习输入数据的良好视觉表征。对比损失的特点遵循以下原则：在特征空间中，给定数据样本经不同的数据增强函数变换后得到的两个视图具有相关性（相似度），而不同输入样本经变换后得到的视图则表现出不相似性。编码器的有效性

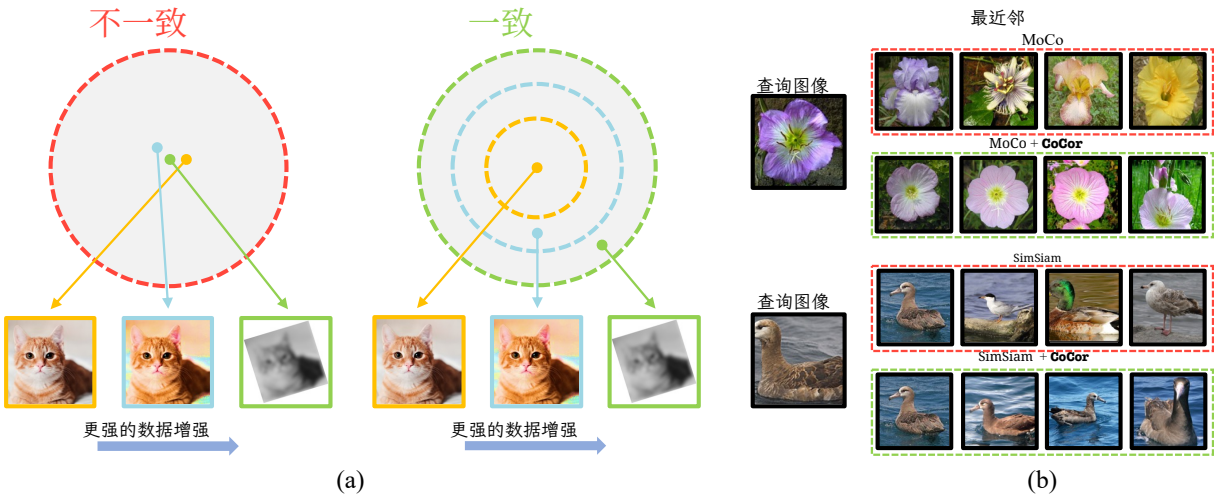


图 1：（a）左：使用标准对比损失训练的编码器可能表现出不一致性，因为同一实例的不同视图被鼓励在特征空间中以相似的方式表示，而不考虑它们之间的实际差异。右：一致的编码器会将增强程度更高的数据向量放置得离原始数据向量更远。这里的圆环表示与中心表征向量具有不同相似度的点。（b）使用预训练编码器，在 CUB-200 Wah 等（2011）和 Flowers102 Nilsback 和 Zisserman（2008）上进行特征空间中的最近邻检索。现有的对比方法 He 等（2020）；Chen 和 He（2021）强制模型对所有数据增强保持不变，因此可能会无意中将彼此不相似的数据聚集在特征空间中。 然而，通过应用一致性，CoCor 可确保只有共享相似潜在语义的数据才会在潜在空间中彼此靠近地分布。

取决于其在无标签数据上的训练效果，这对对比学习的整体性能至关重要，并且有赖于所采用的数据增强的选择。

为了学习有效且具有迁移能力的表征，许多研究致力于通过选择合适的数据增强来改进对比学习。Chen 等（2020a）；Tian 等（2020b）；He 等（2020）；Li 等（2023）；Wang 等（2023）；Van der Sluijs 等（2024）表明，以适当的强度组合不同的增强方式能够提升性能。然而，这些工作将增强方式限制为特定类型在有限强度范围内的随机组合。与此相反，研究表明，采用多样化的数据增强能够有效提升模型捕捉训练数据不变性的能力，从而提高模型的性能（Cubuk 等（2020））、迁移能力（Lee 等人（2021））和鲁棒性（Lopes 等（2019））。因此，近年来对多样化数据增强的探索日益受到关注。PIRL（Misra 和 Maaten（2020））采用额外的增强方式；SwAV（Caron 等（2020））引入多个随机调整大小裁剪，为编码器提供更广泛的数据视图。此外，近期一些工作在对比学习中采用了更强的数据增强组合（Tian 等（2020b）；Wang 和 Qi（2021）；Lee 和 Shin（2023））。

尽管利用多样化数据增强的想法颇具吸引力，但迄今为止，尚无系统性方法能够将大量增强整合到对比学习中。必须注意，数据增强和表征学习的复杂性可能导致性能下降，尤其是在未能审慎选择增强函数时（Tian 等（2020b）；Chen 等（2020a）；Wang 和 Qi（2021））。

为应对上述挑战，本文提出了一种名为“具有一致表征的对比学习”的方法（CoCor）。首先，我们定义一组复合增强，记为 $\mathbf{\Omega}_c$，通过组合多个基本增强构成。每种复合增强都用一个组合向量表示，该向量编码了所采用的基本增强的类型和频次。这组复合增强$\mathbf{\Omega}_c$ 有助于纳入多种数据增强，涵盖广泛的变换类型和整体强度。

关键在于，我们提出了数据增强一致性的概念，称为 数据增强一致性，它为各种复合数据增强生成的视图表征规定了期望属性。具体而言，对于给定的输入样本，更强的数据增强应使其特征空间中的视图远离原始样本的表征。换言之，变换视图与原始输入在特征空间中的相似度会随着数据增强强度的增加而降低，如图1(a)所示。从数据增强的角度来看，未能满足这一属性的编码器被认为是不一致的。

我们提出引入一种新颖的 一致性损失 ，以在编码器的表示空间内强制实现数据增强一致性 。不同于对比损失——后者未考虑所使用数据增强的类型或强度 Chen 等（2020a）； Tian 等（2020b）； He 等（2020）——我们的 一致性损失 会有针对性地引导编码器训练，确保所得特征空间与所应用数据增强的强度相对齐。如图1(b)所示，在最近邻检索任务中，现有的对比学习方法（如 He 等（2020）； Chen 等人（2020b）； Chen 和 He（2021））可能会将不相似的视图聚集在一起，这是由于它们未考虑数据增强所引入的数据方差。然而， CoCor 所强制实现的一致性仅确保具有相似潜在特征的数据在特征空间中彼此接近，同时将数据的不同变体明确分离。此外，我们的实验研究证实，保持数据增强一致性 对于缓解不同数据增强之间的差异至关重要，从而避免特征空间中的潜在性能下降，甚至维度坍塌 Jing 等（2021）； Li 等（2022a）。

所提出的一致性损失取决于如何确定输入样本及其变换视图的表征之间的最优相似度，而该值取决于复合数据增强的强度。然而，这一最优相似度并非先验已知。为应对这一挑战，我们采用数据驱动方法，训练一个神经网络，将每种数据增强的组合向量映射到所需的相似度。鉴于相似度与增强强度之间存在单调关系，我们在神经网络中引入并施加单调性约束。该约束确保更强的复合数据增强对应严格更低的相似度。

所提出的对比　　学习与一致性表征（CoCor）有以下贡献　　：

- 系统地探索大量多样的复合数据增强，以提升编码器在各种下游任务上的性能。
- 提出一个新概念，即数据增强一致性，用于量化输入样本与其变换视图之间的潜在空间相似度对增强强度的单调依赖关系。
- 提出一种新的一致性损失，旨在训练编码器满足数据增强一致性，并利用多样化的增强，同时避免因不一致而导致的性能损失。
- 训练一个具有单调约束的神经网络，以学习最优的潜在空间相似度。

借助一致性表征， CoCor 在各种下游任务上取得了最先进水平的结果。此外，它还可以轻松集成到现有的对比学习框架中，从而有效地对编码器施加数据增强一致性。

2 相关工作

对比学习 近年来在计算机视觉任务及其他领域展现出巨大潜力 Chen 等（2020a）； He 等（2020）； Wang 等（2023）； Caron 等（2020）； Wang 等（2024）。一种常见做法是利用数据增强构造数据的正样本对和负样本对，以定义对比损失 Chen 等（2020a）； Oord 等人（2018）。一些方法强调负样本对在从数据中学习可区分特征方面的重要性 Robinson 等（2020）； Awasthi 等（2022）。MoCo 引入了一个移动平均编码器，并将其与大型负样本记忆队列配对 He 等（2020）； Misra 和 Maaten（2020）。与此相反，Chen 和 He（2021）； Grill 等人（2020）仅从正样本对中学习特征，在不引入负

样本对的情况下取得了最先进水平的性能。此外，监督式对比方法引入与标注相关的信息，以学习更优的表征 Khosla 等人（2020）；Li 等（2022b）。近期研究致力于从表征分布的角度理解对比学习的有效性与局限性 Wang 与 Isola（2020），并深入探讨维度坍塌等问题 Li 等（2022a）；Jing 等（2021）。

数据增强 对自然图像进行数据增强，已被证明能够有效提升监督学习模型的泛化能力和鲁棒性（Cubuk 等人，2019年；2020年；Lopes 等（2019）；Krizhevsky 等人（2017年））。然而，在自监督学习中（Chen 等（2020a）；He 等（2020年）），从有限的数据增强集合中谨慎选择数据增强，对于确保最佳性能至关重要。近期的一些研究（Zhang 和 Ma（2022）；Lee 等人（2021年）；Zhang 等人（2022年）；Devillers & Lefort (2023)）不再将正样本对中的两个视图视为同等，而是建议捕捉由不同随机增强所造成的两个视图之间的方差，并将其称为增强感知信息。AugSelf（Lee 等人（2021年））通过训练辅助网络来预测生成正样本对时所用增强之间的差异。LoGo（Zhang 等人（2022年））提出学习每张图像不同尺寸裁剪图像之间的差异。Zhang 和 Ma（2022年）使用两个网络，将增强图像与数据增强参数一同编码，并结合两种嵌入，形成用于对比损失的特征向量。EquiMod（Devillers & Lefort (2023)）训练网络，根据原始数据和所应用的数据增强，预测增强视图的表征。InfoMin（Tian 等（2020b））和 JOAO（You 等人（2021年））等方法提出搜索最优数据增强策略，以构建对比损失中的样本对；但其实证研究表明，更强且更多样的数据增强实际上可能会降低编码器的性能。相反，近期也有一些研究致力于探索强增强的优势（Wang 和 Qi（2021年）；Caron 等（2020年）；Lee 和 Shin（2023））。然而，迄今为止，尚缺乏一种系统性方法，能够在考虑增强强度影响的同时，充分利用多样化的数据增强。

3 预备知识

对比表征学习的目标是学习一个由 $\boldsymbol{\theta_e}$ 参数化的编码器，该编码器将输入数据 $\boldsymbol{x} \in \mathbb{R}^n$ 映射为特征空间 $\mathcal{Z}$ 中维度为 m 的 ℓ_2 归一化特征向量 $\boldsymbol{z}$，即 $f_{\boldsymbol{\theta_e}}(\cdot)$： $\mathbb{R}^n \rightarrow \mathcal{S}^{m-1}$。编码器通过最小化对比损失进行训练。具体而言，它将同一输入样本的两个不同增强版本定义为一个正样本对，并期望它们在特征空间中具有相似的表征。同时，编码器应经过训练，能够在特征空间中区分来自不同输入样本的任意两个增强实例，即一个负样本对。最小化对比损失会拉近正样本对，并推远负样本对 Oord 等人（2018年）；Chen 等（2020a）；Li 等人（2020年）。

4 方法

4.1 多样化复合增强

所提出的具有一致表示的对比学习（CoCor）通过探索由一组基本增强组合而成的多样化数据增强，提升了对比学习的性能。

定义 1 (复合数据增强)。 长度为 l的复合增强（即 $A_i^{<l>}$）定义为：从给定的N_a 个基本增强函数集合$\mathbf{\Omega}_a = \{a_1(\cdot)\ a_2(\cdot), \cdots\ a_{N_a}(\cdot)\}$中随机抽取 l 个增强，并将它们复合起来： $A_i^{<l>} = a_{i(1)} \circ a_{i(2)} \circ \ldots a_{i(l)}$，其中$i(k) \in [1, N_a],\ k = 1, 2 \cdots l$。

我们将所有长度为 l 的复合增强的集合记为$\mathbf{\Omega}_c^{<l>}$，将不同长度的所有复合增强的集合记为$\mathbf{\Omega}_c = \mathbf{\Omega}_c^{<1>} \cup \mathbf{\Omega}_c^{<2>} \cdots$。我们的研究表明，复合增强中基本增强的排列顺序对性能没有显著影响。为简化讨论，我们假设复合增强具有顺序不变性，例如，$a_i \circ\ a_j = a_j \circ a_i$。我们用一个唯一的组合向量表示复合增强，如图2(b)所示。

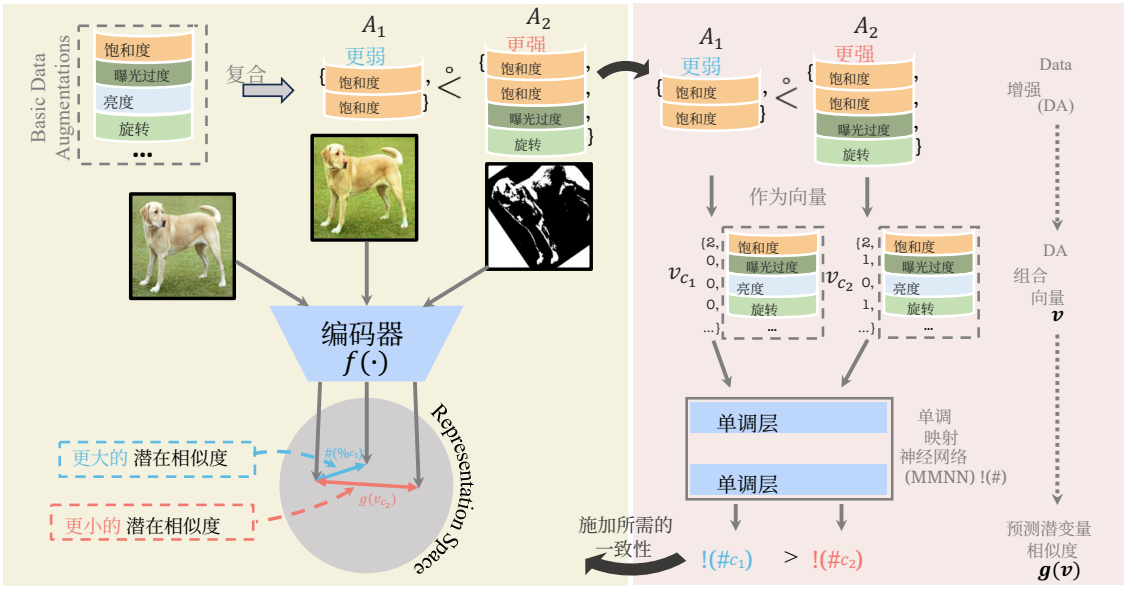


图2：所提方法概述。(a) 所提出的性质——数据增强一致性——确保在表示空间中，经过更强数据增强的数据与原始数据之间的距离，比弱增强视图与原始数据之间的距离更远。(b) 单调映射神经网络（MMNN）以增强组合向量为输入，预测增强视图与原始数据之间的最优潜在相似度。更强的数据增强会使 MMNN 预测的潜在相似度更小。

定义2 (组合向量). 长度为l 的复合增强 $A_i^{<l>}$的组合向量定义为一个 N_a维向量 $\boldsymbol{v}_i = \boldsymbol{v}(A_i^{<l>}) \in \mathbb{N}_0^{N_a}$（$\mathbb{N}_0$ 是所有自然数的集合），其中第j_{th} 个分量 $\boldsymbol{v}_i[j]$ 表示基本增强 a_j 在 $A_i^{<l>}$中应用的次数，且$\sum_{j=1}^{N_a} \boldsymbol{v}_i[j] = l$。

4.2 数据增强一致性

数据增强在对比学习中发挥着关键作用，因此需要精心设计，以揭示原始数据的潜在结构。然而，我们的关键观察是，标准对比学习框架 Chen 等（2020a）；He 等（2020）；Chen 和 He（2021）并未内在地蕴含数据增强（*DA*）一致性这一由本研究提出的基本性质。如果缺乏这种一致性，编码器可能只会在特征空间$\mathcal{Z} \subseteq \mathcal{S}^{m-1}$中将正样本对中的视图拉近，而不考虑它们之间的实际差异。然而，输入样本的弱增强视图和强增强视图未必具有相同的潜在结构；如图1(a)所示，将它们编码到特征空间中的不同区域可能更有利。在这种情况下，试图将正样本对中的视图拉近的编码器会学到错误的表征分布。

为此，我们提出了数据增强一致性这一概念。在数据增强一致性 的约束下，编码器的训练目标不仅是将正样本对聚集在一起，还要根据所应用的数据增强的强度和类型，将它们映射到潜在特征空间$\mathcal{Z}$ 中的各自最优位置。因此，编码器能够保留数据增强引入的关键信息，例如与亮度、锐度和旋转相关的方差，如图2(a)所示。预训练编码器的这一特性提升了其在各种下游任务中的表现，而这些任务中的识别往往离不开此类信息。

我们通过相对于原始输入数据的潜在相似度 来量化映射后的特征空间位置。

定义 3 (潜在相似度)。 给定编码器 f和复合增强 A，输入 $\boldsymbol{x}$的潜在相似度 $l_d(\boldsymbol{x}; f, A)$定义为特征空间 $\mathcal{Z} \subseteq \mathcal{S}^{m-1}$中$\boldsymbol{x}$ 的归一化表示与 $A(\boldsymbol{x})$的归一化表示之间的余弦相似度： $l_d(\boldsymbol{x};\ f, A)\ = f(\boldsymbol{x})^T \cdot\ f(A(\boldsymbol{x}))$。

若要实现数据增强一致性，同一数据经不同增强处理后得到的增强视图，其潜在空间相似度应随增强强度单调递减。我们学习一个映射，即参数化神经网络 $g_{\boldsymbol{\theta_d}}(\cdot)$，它以复合增强 A 的组合向量 $\boldsymbol{v}(A)$ 为输入，并将其映射为最优潜在相似度 $l_d^*(\boldsymbol{v}(A))$，详见第 4.3 节。如果某个最优编码器对于任意给定的复合增强，其潜在相似度都被视为最优，则该编码器被认为完全符合数据增强一致性。对于未完全符合数据增强一致性的编码器，我们进一步定义数据增强一致性水平，即所学习的表征在数据增强方面达到一致的程度。

定义 4 (一致性水平)。 给定一个编码器 f、一组复合增强 $\boldsymbol{\Omega}_c$ 和一个原始输入 $\boldsymbol{x}$，数据增强一致性水平（DACL）定义为编码器与最优潜在相似度之间的偏差：

$$\mathrm{DACL}(\boldsymbol{x}; f, \boldsymbol{\Omega}_c) = \mathbb{E}_{A\sim\boldsymbol{\Omega}_c}[|l_d(\boldsymbol{x}; f, A) - l_d^*(\boldsymbol{v}(A))|] \tag{1}$$

DACL 通过比较给定编码器 f 对 $\boldsymbol{x}$ 增强视图的潜在相似度与相应的最优潜在相似度 $l_d^*(\boldsymbol{v}(A))$，衡量该编码器在所有可能的复合增强下的一致性水平。

通过 DACL，可以通过最小化下文定义的一致性损失 ，训练由 $\boldsymbol{\theta_e}$ 参数化且尚未完全符合 DA 一致性的编码器 $f_{\boldsymbol{\theta_e}}(\cdot)$，使其更加符合 DA 一致性：

$$\begin{aligned}\mathcal{L}_{\text{consistent}}(\boldsymbol{\theta_e}) &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{X}, A\sim\boldsymbol{\Omega}_c}[|\mathrm{DACL}(\boldsymbol{x}; f_{\boldsymbol{\theta_e}}, \boldsymbol{\Omega}_c)|] \\ &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{X}, A\sim\boldsymbol{\Omega}_c}[|l_d(\boldsymbol{x}; f_{\boldsymbol{\theta_e}}, A) - l_d^*(\boldsymbol{v}(A))|] \\ &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{X}, A\sim\boldsymbol{\Omega}_c}[|l_d(\boldsymbol{x}; f_{\boldsymbol{\theta_e}}, A) - g_{\boldsymbol{\theta_d}}(\boldsymbol{v}(A))|]\end{aligned} \tag{2}$$

我们的实证研究表明，另一种一致性损失能够带来更好的性能；这种替代方案的详细信息见附录。

4.3 学习最优潜在相似度

4.3.1 用于最优潜在相似度的可学习模型

要使用公式 2 中提出的一致性损失来训练编码器，我们需要针对给定的复合增强 A，提供最优潜在相似度 $l_d^*(\boldsymbol{v}(A))$，而这取决于 A 的增强强度。然而，定义增强强度颇具挑战，因为两种复合增强可能包含不同的基本增强类型，因此难以定义和比较它们的强度。

为了解决这个问题，我们学习一个映射，即参数化神经网络 $g_{\boldsymbol{\theta_d}}(\cdot)$，它以复合向量为输入，并将其映射为最优潜在相似度 $l_d^*(\boldsymbol{v}(A))$。这种方法免去了直接对不同增强的强度进行建模的需要。

4.3.2 对可学习模型施加单调性

更进一步，我们不仅使用神经网络近似最优潜在相似度 l_d^*，还施加一项重要的单调性，以更好地学习所需的最优潜在相似度。

随着复合增强的整体强度增加，例如加入更多基础增强，增强数据与原始数据之间的相似度预计会降低。例如，尽管直接比较两种不同增强 `GaussianBlur` 和 `Sharpness` （分别通过高斯核模糊图像和调整边缘对比度）的效果可能颇具挑战，但在复合数据增强中应用`Sharpness` 两次，必然会比应用一次对原始输入造成更大的失真。

[**数据增强强度比较算子**]因此，如图3所示，我们正式定义一个算子 $\mathring{>}$ ，用于比较集合$\boldsymbol{\Omega}_c$中复合数据增强的强度： $A_i \mathring{>} A_j$ 当且仅当$\boldsymbol{v}(A_i) \mathring{>} \boldsymbol{v}(A_j)$，其中 $\mathring{>}$ 逐元素作用于组合向量： $\boldsymbol{v}(A_i) \mathring{>} \boldsymbol{v}(A_j)$意味着$\boldsymbol{v}(A_i)[k] \geq \boldsymbol{v}(A_j)[k]$， $\forall k \in [1, N_a]$，并且 $\exists k \in [1, N_a]$使得 $\boldsymbol{v}(A_i)[k] > \boldsymbol{v}(A_j)[k]$。

换言之，当且仅当在 A_i 中应用任意一种基本增强的次数不少于在 A_j中应用该增强的次数，且至少有一种基本增强在 A_i 中的应用次数多于在 A_j中时，才认为复合增强 A_i 强于 A_j （即 $A_i \mathring{>} A_j$）。

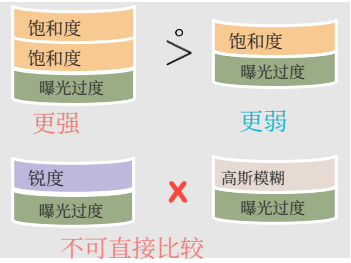


图3：当且仅当一种复合增强包含另一种复合增强的所有组成部分，并额外包含基本增强时，才认为前者强于后者。

我们的可学习神经网络模型 $g_{\boldsymbol{\theta}_d}(\cdot)$以
增强 A_i 的组合向量作为输入，并将其映射为特征空间中$l_d^*(\boldsymbol{v}(A_i))$的最优潜在相似度： $l_d^*(\boldsymbol{v}(A_i)) = g_{\boldsymbol{\theta}_d}(\boldsymbol{v}(A_i))$。我们在$g_{\boldsymbol{\theta}_d}(\cdot)$上强制要求其关于输入 $\boldsymbol{v}(A_i)$满足单调性：$g_{\boldsymbol{\theta}_d}(\boldsymbol{v}(A_i)) > g_{\boldsymbol{\theta}_d}(\boldsymbol{v}(A_j))$，若 $v(A_i) \mathring{>} v(A_j)$，则 $\forall A_i$, $\forall A_j \in \boldsymbol{\Omega}_c$。换言之，**更强的数据增强会产生更小的潜在相似度**。如图2(b)所示，我们通过在 $g_{\boldsymbol{\theta}_d}(\cdot)$中引入 You 等人（2017）的单调线性嵌入层来强制满足单调性。我们将 $g_{\boldsymbol{\theta}_d}(\cdot)$称为单调映射神经网络（MMNN）。

4.4 一致性编码器训练

所提出的 CoCor 方法将公式 2 中的一致性损失纳入整体损失函数$\mathcal{L}_u$ ，以便在给定无标签数据和 MMNN 参数 $\boldsymbol{\theta_d}$ 的情况下优化编码器参数 $\boldsymbol{\theta_e}$ ：

$$\mathcal{L}_u(\boldsymbol{\theta_e}|\boldsymbol{\theta_d}) = \mathcal{L}_{\text{contrast}}(\boldsymbol{\theta_e}) + \mathcal{L}_{\text{consistent}}(\boldsymbol{\theta_e}|\boldsymbol{\theta_d}) \tag{3}$$

在公式 3 中，编码器参数 $\boldsymbol{\theta_e}$的优化取决于 MMNN 参数 $\boldsymbol{\theta_d}$。因此，我们将 $\boldsymbol{\theta_e}$ 对 $\boldsymbol{\theta_d}$ 的依赖关系表示为 $\boldsymbol{\theta_e}(\boldsymbol{\theta_d})$，并通过求解双层优化问题来优化 $\boldsymbol{\theta_d}$:

$$\min_{\boldsymbol{\theta_d}} \quad \text{CE}(\boldsymbol{\theta_e^*}(\boldsymbol{\theta_d})) \tag{4}$$

$$\text{s.t.} \quad \boldsymbol{\theta_e^*}(\boldsymbol{\theta_d}) = \arg\min_{\boldsymbol{\theta_e}} \mathcal{L}_u(\boldsymbol{\theta_e}|\boldsymbol{\theta_d}) \tag{5}$$

公式 4 中的顶层问题基于少量带标签数据上的交叉熵损失 CE，对 MMNN 和线性分类器进行优化（公式 4 中未明确显示线性分类器）。每次迭代中，都会求解公式 5 中的底层问题，以更新无标签数据上的编码器参数 $\boldsymbol{\theta_e}$。关于 θ_e对 θ_d 的依赖关系及 θ_d 的更新规则的推导，我们留待附录说明。

我们在算法1中总结了所提出方法的算法流程。

CoCor 的设计使其兼容各种最先进的自监督学习框架。CoCor 可以通过将方程3中的$\mathcal{L}_{\text{contrast}}$ 替换为该方法特定的损失函数，集成到现有方法中。新增的一致性损失可施加数据增强一致性，并提升这些框架中预训练编码器的泛化能力和迁移能力。

算法1： CoCor的算法流程**输入：** 初始编码器、MMNN 和分类器参数 $\boldsymbol{\theta}_e^{(0)}$、$\boldsymbol{\theta}_d^{(0)}$ 和 $\boldsymbol{\theta}_c^{(0)}$，训练轮数 N，无标签数据加载器 $\mathcal{D}_u$，带标签数据加载器 $\mathcal{D}_l$。**对于**i=1 **到** N **执行**1. 从 $\mathcal{D}_u$ 和 $\mathcal{D}_l$分别抽取无标签数据 $\boldsymbol{x}_u$ 和带标签数据（$\boldsymbol{x}_l, \boldsymbol{y}_l$）。2. 调用方程5，在固定 MMNN 参数 $\boldsymbol{\theta}_d^{(i-1)}$ 的情况下，使用 $\boldsymbol{x}_u$ 更新编码器参数 $\boldsymbol{\theta}_e^{(i)}$。3. 调用方程4，在固定编码器参数$\boldsymbol{\theta}_e^{(i)}$ 的情况下，使用（$\boldsymbol{x}_l, \boldsymbol{y}_l$）更新 MMNN $\boldsymbol{\theta}_d^{(i)}$ 和分类器 $\boldsymbol{\theta}_c^{(i)}$。**输出：** 参数为 $\boldsymbol{\theta}_e^{N}$的训练后编码器

5 实验研究

我们将 CoCor 整合到多种最先进的对比学习方法中，以展示其性能和通用性：MoCo（Chen 等人（2020b）；He 等（2020））是一种采用大型负样本记忆队列的双编码器方法；SimSiam（Chen 和 He（2021））仅利用正样本对；SupCon（Khosla 等人（2020））是一种利用带标签数据增强表征学习的监督式对比学习方法。我们将这三种方法分别称为基线方法。我们在多种数据集上进行线性评估和目标检测，将这些基线方法预训练得到的编码器与采用 CoCor 和基线方法组合进行预训练得到的编码器进行性能比较。我们还将 CoCor 与近期考虑增强感知信息的研究（Zhang 和 Ma（2022）；Lee 等人（2021）；Devillers & Lefort (2023)），以及采用比传统对比学习更强的数据增强的研究（Lee 和 Shin（2023）；Wang 和 Qi（2021））进行了比较。为进一步揭示CoCor的工作机制，我们对潜在相似度和可能出现的维度坍塌进行了事后分析，并开展了若干消融研究。

除非 另有说明，本文所呈现的所有实验结果均由 u 复现 s.

5.1 编码器预训练的实验设置

基线方法的编码器在大型 ImageNet-1K 数据集 Russakovsky 等人（2015）及其子集 ImageNet-100 Tian 等人（2020a）上进行预训练，并采用两种不同的骨干编码器架构：ResNet-50 和 ResNet-34 He 等人（2016）。基于 MoCo 和 SimSiam 的方法，其编码器预训练 200个训练轮次；而 SupCon 的编码器则经过 100 个训练轮次的预训练。所有预训练实验的批量大小均设为 256。其他基线方法设置遵循 Khosla 等人（2020）；Chen 等人（2020b）；Chen 和 He（2021）。

为确保公平比较，加入 CoCor 的预训练采用与相应基线相同的实验设置。 CoCor 中的 MMNN 是一个三层多层感知机，包含单调线性嵌入层 You 等人（2017）和 ReLU 单元。虽然大多数结果是在仅使用预训练数据集 1% 的标签来调整 MMNN 时展示的，但我们也在消融研究中展示了 CoCor 的良好表现：表5.4 中带标签数据的使用量进一步大幅减少。 CoCor 采用基础数据增强集 Ω_a，其中包含 14 种常用增强函数，详情见附录。我们对若干特定长度的复合数据增强集 $\Omega_c^{<l>}$施加数据增强一致性约束。对于大多数预训练实验，我们取 $l = 1$，2，3，即对每个样本，采样三个长度分别为 1、2 和 3 的复合数据增强，并将其应用于样本，以获得三个数据视图来计算一致性损失。关于数据增强和预训练设置的更多详情见附录。

5.2 主要结果

线性评估 每个预训练编码器的参数均被冻结，并与一个线性分类器配对；该分类器按照 Krizhevsky 等人（2017）的线性评估协议进行微调，详情见附录。线性评估在以下数据集上进行：Cifar-10/100（Krizhevsky 等人，2009）、CUB-

表 1：线性评估的 Top-1 准确率（%）。所有 ResNet-50 骨干编码器均在 ImageNet-100 上进行预训练。

方法	训练轮数	IN-100	CIFAR-10	CIFAR-100	CUB200	Caltech-101	SUN397	Food101	Flowers102	Pets
MoCo Chen 等人（2020b）	200	67.04	82.15	59.17	21.52	79.23	39.43	52.97	71.77	56.94
MoCo + CoCor（**我们的方法**）	200	**71.66**	**83.87**	**59.64**	**22.57**	**81.51**	**41.95**	**54.81**	**75.65**	**60.23**
SimSiam Chen 和 He（2021）	200	73.92	84.66	61.82	26.13	85.11	45.96	58.73	82.40	65.60
SimSiam + CoCor（**我们的方法**）	200	**83.70**	**86.89**	**66.36**	**34.50**	**87.85**	**49.92**	**62.48**	**87.95**	**76.04**
SupCon Khosla 等人（2020）	100	80.16	84.30	61.33	26.46	85.04	41.95	52.45	76.99	73.75
SupCon +CoCor（**我们的方法**）	100	**82.14**	**85.49**	**62.18**	**27.40**	**87.12**	**42.87**	**53.93**	**79.04**	**75.95**

表 2：线性评估的 Top-1 准确率（%）。所有 ResNet-50 骨干编码器均在 ImageNet-1K 上预训练。

方法	训练轮数	IN-1K	CIFAR-10	CIFAR-100	CUB200	Caltech-101	SUN397	Food101	Flowers102	Pets
MoCo Chen 等人（2020b）	200	67.56	92.24	74.33	41.54	92.00	58.28	69.84	88.86	81.74
MoCo + CoCor（**我们的方法**）	200	**72.83**	**93.28**	**76.19**	**44.11**	**93.10**	**60.15**	**71.10**	**90.39**	**83.08**
SimSiam Chen 和 He（2021）	200	70.93	93.08	76.24	51.05	93.13	60.81	71.18	92.11	85.01
SimSiam + CoCor（**我们的方法**）	200	**73.25**	**94.04**	**78.12**	**54.50**	**94.51**	**63.09**	**73.31**	**92.84**	**87.24**
SupCon Khosla 等人（2020）	100	74.18	93.17	76.55	52.30	93.28	61.03	71.54	91.34	85.35
SupCon + CoCor（**我们的方法**）	100	**76.24**	**94.61**	**77.91**	**55.73**	**94.66**	**63.80**	**74.09**	**92.17**	**87.21**

200（Wah 等（2011））、加州理工学院-101（Fei-Fei 等人，2004）、SUN397（Xiao 等人，2010）、Food101（Bossard 等人，2014）、Flowers102（Nilsback 和 Zisserman，2008）、牛津-IIIT宠物（Pets）（Parkhi 等人，2012）、Aircraft（Maji 等人，2013）以及 StanfordCars（Krause 等人，2013）。

ImageNet-100 和 ImageNet-1K 预训练编码器的分类准确率见表 1 和表 2。值得注意的是， CoCor 显著提升了预训练编码器在各类分类任务上的表现。这一改进表明， CoCor 通过引入更广泛的数据增强，并补充特征空间中的语义，增强了编码器的泛化能力。

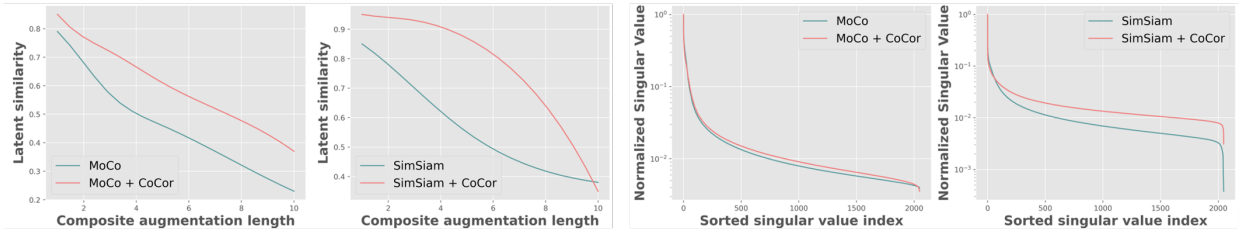
目标检测我们在 VOC2007+2012 Everingham 等人（2009）和 COCO2017 Lin 等人（2014）数据集上对预训练编码器进行微调，以用于目标检测下游任务。我们使用 Detectron2 Wu 等人（2019），将预训练编码器转换为采用 ResNet50-C4 骨干网络的广义 R-CNN Girshick 等人（2014）检测器。在 VOC2007 和 COCO2007 上训练的模型随后在相应数据集的测试集上进行评估。

与基线方法相比，表 3 显示， CoCor 在 AP、 AP_{50}和 AP_{75}等标准指标下显著提高了准确率，这表明 CoCor 有助于提升预训练编码器的迁移能力。值得注意的是， CoCor 在检测具有挑战性的小目标和中目标时取得了显著改进， AP_s 和 AP_m 分数的提高也证明了这一点。

事后研究 我们通过比较在预训练过程中使用和未使用一致性损失的编码器的潜在相似度，来可视化式 2 所提出的一致性损失的影响。表4 (a)展示了 ImageNet- 100 数据集上，各种特定长度的复合增强的平均潜在相似度。显然，使用 CoCor 会使不同数据增强长度范围内的潜在相似度值更大，这表明具有相同身份的图像视图聚类得更紧密。这与 Wang 和 Isola（2020）提出的分析一致：即使不考虑对增强强度的依赖，由较弱增强生成的视图聚类得更紧密，也与编码器性能提升相关。此外，图1(b)中的图像检索任务结果表明，与基线方法相比， CoCor 能够学习到语义一致性更强的特征空间。

表 3：VOC 和 COCO 目标检测任务上的迁移学习结果。

方法	预训练 训练轮数	VOC07+12 AP	 AP_{50}	 AP_{75}	COCO AP	 AP_{50}	 AP_{75}	 AP_s	 AP_m	 AP_l
MoCo He 等（2020）	200	50.23	76.52	54.42	34.23	55.08	36.60	16.67	37.27	50.97
MoCo + CoCor（**我们的方法**）	200	**51.11**	**77.41**	**54.88**	**39.13**	**58.30**	**42.19**	**23.19**	**43.45**	**52.13**
SimSiam Chen 和 He（2021）	200	52.04	77.86	57.11	34.15	55.21	36.50	15.30	37.37	50.49
SimSiam + CoCor（**我们的方法**）	200	**54.31**	**80.60**	**59.58**	**34.80**	**56.29**	**37.05**	**17.53**	**39.88**	**51.74**
SupCon Khosla 等人（2020）	100	47.41	75.83	53.17	34.26	55.11	36.35	14.34	37.40	50.56
SupCon +CoCor（**我们的方法**）	100	**49.03**	**77.22**	**54.19**	**35.67**	**56.10**	**37.98**	**16.63**	**38.71**	**51.29**



(a) (b)

图4：（a）：在 ImageNet-100 上评估的、采用和未采用一致性损失预训练的编码器在不同复合增强长度下的潜在相似度；（b）：学习到的潜在表示的奇异值。

这种引入的一致性确保具有相似潜在语义的数据彼此靠近，而不相似的实例则在特征空间中保持距离。

维度坍塌评估 我们通过对特征向量进行主成分分析（PCA），展示在 ImageNet-100 上预训练的编码器所得到的潜在表示的奇异值。维度坍塌通常与观测到较小的奇异值有关。然而，如图4 (b) 所示，与两个基线方法相比，这种情况在由 CoCor 预训练的编码器中不太可能出现。尽管有研究认为，更强的数据增强 Ji ng 等（2021）和更大的数据集 Li 等（2022a）可能导致维度坍塌，但在 CoCor编码器中并未观察到这一问题。在传统的对比学习方法中，编码器被迫对增强引入的特征方差保持不变，这可能导致特征空间的某些维度取值为常数。 CoCor通过施加一致性来捕获这种方差，因此可能缓解了维度坍塌问题。

表4：SimSiam、现有增强感知方法和 CoCor 在线性评估中的 Top-1 准确率（%）对比。编码器（ResNet-34）在 ImageNet-100 上训练 200个训练轮次。

方法	ImageNet100	CUB200	Flowers102	StanfordCars
SimSiam Chen 和 He（2021）	63.90	29.10	76.10	19.60
AugSelf Lee 等（2021年）	75.96	29.96	82.96	29.41
Hierarchical Zhang 和 Ma（2022）	67.10	33.90	83.60	20.70
EquiMod Devillers 和 Lefort (2023)	65.45	30.09	77.91	19.06
SimSiam+CoCor（**我们的**）	**77.76**	**34.07**	**83.88**	**29.93**

表5：采用更强数据增强的方法在线性评估中的 Top-1 准确率（%）对比。编码器（ResNet-50）在 ImageNet-1K 上训练 200个训练轮次。

方法	ImageNet1K	CUB200	Flowers102	Aircraft
CLSA Wang 和 Qi（2021）	69.4	40.20	89.97	61.24
R´enyiCL Lee 和 Shin（2023）	72.6	**45.05**	90.17	61.80
MoCo+CoCor（**我们的方法**）	**72.8**	44.11	**90.39**	**63.18**

5.3 与最先进水平的比较

与增强感知方法的比较 近期，若干研究 Devillers & Lefort (2023); Lee 等（2021）; Zhang 和 Ma（2022）在对比学习中考虑了增强感知信息。AugSelf Lee 等（2021）通过训练辅助网络来预测生成正样本对时所应用的数据增强之间的差异。Zhang 和 Ma（2022）使用两个网络，将增强图像与数据增强参数一同编码，并结合两个嵌入以形成用于对比损失的特征向量。EquiMod Devillers & Lefort (2023) 训练网络，根据原始数据和所应用的数据增强来预测增强视图的表征。表4展示了这些方法与 CoCor 在线性评估中的性能比较。所有编码器均采用 ResNet-34 架构，并在 ImageNet-100 上预训练 200个训练轮次。与 Lee 等（2021）; Zhang 和 Ma（2022）; Devillers & Lefort (2023) 相比， CoCor 引入了强得多的数据增强以及更多样的数据增强类型。 CoCor的表现始终优于这些方法，这可能归因于它有效利用了范围显著更广的数据增强，同时保持了增强强度中定义明确的单调性。

与采用更强数据增强的方法进行比较 近期一些对比学习方法采用了更强的数据增强。CLSA Wang 和 Qi（2021）利用弱增强视图之间的分布散度，作为弱增强视图与强增强视图之间分布散度的监督信号。R´enyiCL Lee 和 Shin（2023）提出了一种新的对比目标，使用 R´enyi 散度来处理强增强数据。表5展示了这些方法与我们方法的比较。这里，所有编码器均为 ResNet-50，在 ImageNet-1K 上训练 200个训练轮次。与 R´enyiCL 和 CLSA 不同，后两者未明确区分不同类型的数据增强（DAs），而 CoCor 通过使用 MMNN 区分不同的数据增强类型来解决这一问题。此外， CoCor 引入了不仅更强而且更多样的数据增强，并在不同强度的数据增强之间实现一致性；相比之下，R´enyiCL 和 CLSA 在每个预训练过程中只采用单一强度的数据增强。

5.4 消融研究

基础数据增强 定义了编码器需要学习的增强感知信息。为了更好地理解这些数据增强在预训练期间的作用，我们将它们分为两组：Ω_{color} 和 Ω_{affine}。Ω_{color} 包括改变每个像素颜色但保持其位置不变的增强，例如 Brightness 和 Contrast。Ω_{affine} 由仿射变换组成，例如 Rotate 和 Shear。附录中给出了 Ω_{color} 和 Ω_{affine} 的组合方式。使用 CoCor，我们分别在 ImageNet-100 上使用 Ω_{color}、Ω_{affine}和 $\Omega_{\text{color}} \bigcup \Omega_{\text{affine}}$对三个编码器进行了 500个训练轮次的预训练。表 6 展示了在 ImageNet-100 以及两个细粒度数据集 CUB-200 和 Flowers102 上的线性评估结果。结果表明， CoCor 的整体表现优于 LooC、Xiao 等人（2020）和 AugSelf、Lee 等人（2021）；后两者提升了模型在各种下游识别任务中的迁移能力。值得注意的是，尽管 $\Omega_{\text{color}} \bigcup \Omega_{\text{affine}}$ 在所测试的数据集上取得了最佳总体结果，但改进基础数据增强可以提升 CoCor 在特定下游任务上的表现。例如，颜色感知信息对于区分 Flower102 数据集中的花卉图像至关重要（Nilsback 和 Zisserman（2008）），而仿射变换引入的位置相关信息则不那么重要。因此，仅应用颜色相关增强，同时保持对仿射变换的不变性，可以提升在 Flower102 上的表现。

表 6：LooC、AugSelf 和 CoCor 的比较。 CoCor 使用基础数据增强池进行预训练，其中包括颜色相关增强、仿射变换及其组合。报告 Top-1 线性评估准确率。编码器（ResNet-50）在 ImageNet-100 上训练 500个训练轮次。

方法	ImageNet100	CUB200	Flowers（5样本）	Flowers（10样本）
MoCo* Chen 等人（2020b）	81.0	36.7	67.9	77.3
LooC（颜色）* Xiao 等人（2020）	81.1	40.1	68.2	77.6
LooC（旋转）* Xiao 等人（2020）	80.2	38.8	70.1	79.3
LooC（颜色、旋转）* Xiao 等人（2020）	79.2	39.6	70.9	80.8
MoCo + AugSelf* Lee 等人（2021）	82.4	37.0	81.7	84.5
SimSiam + AugSelf* Lee 等人（2021）	82.6	45.3	86.4	88.3
SimSiam +CoCor（颜色）	83.5	46.5	**89.1**	**90.2**
SimSiam + CoCor（仿射变换）	81.8	45.2	83.5	86.1
SimSiam + CoCor（颜色+仿射变换）	**83.7**	**47.9**	88.9	89.6

*由于 LooC 的官方实现尚未发布，我们采用 Lee 等人（2021）和 Xiao 等人（2020）的结果。为确保公平比较，我们对本表中的所有方法均采用相同的实验设置。

表 7：使用 MMNN 训练和未使用 MMNN 训练的 CoCor 模型的线性评估。编码器（ResNet-50）在 ImageNet-100 上训练 200个训练轮次。

	基线	不使用 MMNN	使用 MMNN
MoCo + CoCor	67.04	69.51	**71.66**
SimSiam +CoCor	73.92	82.18	**83.70**

MMNN 的有效性 CoCor 利用 MMNN 模型学习各种复合增强的最佳潜在相似度l_d^* 。为了检验这一可学习 MMNN 模型的优势，我们用手动优化的模型替代它：通过大量反复试验，针对每种复合增强长度手动搜索最佳的 l_d^* 值，以获得最佳线性评估表现。附录中包含了关于 l_d^* 选择的更详细实验设置和结果。

我们使用 MMNN 和基于复合增强长度 $l = 1$、2、3 的手动模型，在 ImageNet-100 上对基于 SimSiam 和 MoCo 构建的 ResNet-50 编码器进行预训练，共训练 200 个训练轮次。结果如表 7 所示，其中还包括 MoCo 和 SimSiam 基线方法的性能。与未使用 MMNN（即使用手动模型）训练的编码器相比，使用 MMNN 训练的编码器显然具有优势。

MMNN 训练中带标签数据量的影响 我们进一步研究用于训练 MMNN 的监督信号量对编码器性能的影响。我们在 ImageNet-100 上对基于 SimSiam 构建的 ResNet-50 CoCor 编码器进行 100 轮预训练，同时使用数据集中 100%、10%、1% 和 0.1% 的带标签数据训练 MMNN。随后通过在 ImageNet-100 上进行线性评估来评估预训练骨干网络。表 8 显示，将 MMNN 训练数据量从 100% 大幅减少到 0.1%，性能仅下降 0.94%；此外，所有四种 CoCor 模型的表现都比基线 SimSiam 高出 15% 以上。

预训练中的复合增强长度 我们提出的复合增强能够生成多样且信息丰富的视图。我们通过实验检验复合增强强度的影响：每次预训练实验中仅采用一种复合增强长度。我们在 SimSiam 和 MoCo 上运行 CoCor ，分别采用长度为 1、2、3、4 的单一增强，以及长度为 1、2 和 3 的组合。所有编码器均为在 ImageNet-100 上预训练 50 轮的 ResNet-50，随后在 ImageNet-100 上进行评估。

如表5所示， CoCor 能够利用不同长度的增强，并且在所有这些长度上组合数据增强还能进一步提升性能。与 Chen 等（2020a）呈现的结果不同；

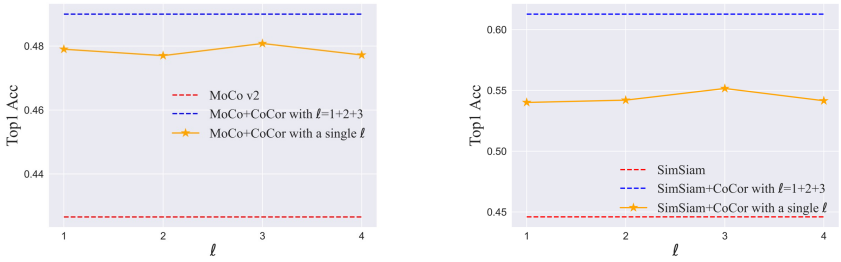


图 5：在（a）MoCo 和（b）SimSiam 上，使用单一长度l 的复合增强以及长度为 1、2 和 3 的组合训练得到的 CoCor 编码器的线性评估。

表 8：使用不同数量带标签数据训练的 CoCor 模型的性能。编码器（ResNet-50）在 ImageNet-100 上训练 50 轮。

标签比例	100%	10%	1%	0.1%	SimSiam
Top-1准确率	75.97	75.82	75.79	75.46	59.20

Tian 等（2020b）的研究表明，强增强可能会降低性能。我们进行了额外实验，采用强度大得多的复合增强，以展示 CoCor利用此类增强的能力。这些结果见附录，表明与近期在方法中引入强增强的研究 Lee 和 Shin（2023）；Wang 和 Qi（2021）相比， CoCor能够从更强的数据增强中获益。

6 讨论

本文介绍了 CoCor，这是一种系统化方法，旨在探索对比学习中的多样化数据增强。我们的贡献包括引入数据增强一致性，以量化目标潜在相似度对所应用数据增强的依赖关系。我们提出了一种一致性损失，以引导编码器训练，从而提升数据增强一致性。为了实现数据增强一致性，我们采用数据驱动方法来学习最优依赖关系，并将其应用于编码器训练。 CoCor 的有效性通过在各种任务和数据集上的大量实验结果得到了证实。有关 MMNN 更新规则、其他实验设置以及扩展结果的更多细节，请参阅附录。

作为一个潜在的未来研究方向，我们建议探索针对数据示例学习最优潜在相似度。具体而言，将 MMNN 扩展为以数据实例作为输入，或可考虑由数据方差引起的潜在相似度方差。这种方法有望学习从数据增强到潜在相似度的更准确映射。

参考文献

Pranjal Awasthi、Nishanth Dikkala 和 Pritish Kamath。《在对比学习中，更多负样本是否必然有害？》，2022年。

Lukas Bossard、Matthieu Guillaumin 和 Luc Van Gool。Food-101——使用随机森林挖掘判别性成分。载于计算机视觉—*ECCV 2014*：第*13*届欧洲会议，瑞士苏黎世，*2014*年*9*月*6*日至*12*日，论文集，第*VI*部分 *13*，第446–461页。施普林格，2014年。

Mathilde Caron、Ishan Misra、Julien Mairal、Priya Goyal、Piotr Bojanowski 和 Armand Joulin。通过对比簇分配进行无监督视觉特征学习。神经信息处理系统进展，33:9912–9924，2020年。

Ting Chen、Simon Kornblith、Mohammad Norouzi 和 Geoffrey Hinton。用于视觉表征对比学习的简单框架。*arXiv* 预印本 *arXiv:2002.05709*，2020a。Xinlei Chen 和 何恺明。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。Xinlei Chen、Haoqi Fan、Ross Girshick 和 何恺明。基于动量对比学习的改进基线方法。*arXiv* 预印本 *arXiv:2003.04297*，2020b。Dan Ciregan、Ueli Meier 和 Jürgen Schmidhuber。用于图像分类的多列深度神经网络。载于*2012*年 *IEEE*计算机视觉与模式识别会议，第 3642–3649 页。IEEE，2012年。Ekin D Cubuk、Barret Zoph、Dandelion Mane、Vijay Vasudevan 和 Quoc V Le。Autoaugment：从数据中学习增强策略。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 113–123 页，2019年。Ekin D Cubuk、Barret Zoph、Jonathon Shlens 和 Quoc V Le。Randaugment：通过缩减搜索空间实现实用的自动化数据增强。载于*IEEE/CVF*计算机视觉与模式识别会议研讨会论文集，第 702–703 页，2020年。Alexandre Devillers 和 Mathieu Lefort。Equimod：用于改进视觉实例判别的等变模块。载于国际学习表征会议，2023年。Mark Everingham、Luc Van Gool、Christopher KI Williams、John Winn 和 Andrew Zisserman。Pascal视觉物体类别（VOC）挑战赛。《国际计算机视觉期刊》，88:303–308，2009年。李飞飞、Rob Fergus 和 Pietro Perona。从少量训练样本中学习生成式视觉模型：在101个物体类别上测试的增量式贝叶斯方法。载于*2004*年计算机视觉与模式识别会议研讨会，第 178–178 页。IEEE，2004年。Ross Girshick、Jeff Donahue、Trevor Darrell 和 Jitendra Malik。用于精准目标检测和语义分割的丰富特征层次结构。载于*IEEE*计算机视觉与模式识别会议论文集，第 580–587 页，2014年。Jean-Bastien Grill、Florian Strub、Florent Altché、Corentin Tallec、Pierre Richemond、Elena Buchatskaya、Carl Doersch、Bernardo Avila Pires、Zhaohan Guo、Mohammad Gheshlaghi Azar 等。自举潜在表示：一种自监督学习新方法。神经信息处理系统进展，33:21271–21284，2020年。何恺明、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。何恺明、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。Daniel Ho、Eric Liang、Xi Chen、Ion Stoica 和 Pieter Abbeel。基于种群的数据增强：高效学习增强策略调度。载于国际机器学习会议，第 2731–2741 页。PMLR，2019年。Li Jing、Pascal Vincent、Yann LeCun 和 Yuandong Tian。理解对比式自监督学习中的维度坍塌。*arXiv* 预印本 *arXiv:2110.09348*，2021年。Prannay Khosla、Piotr Teterwak、Chen Wang、Aaron Sarna、Yonglong Tian、Phillip Isola、Aaron Maschinot、Ce Liu 和 Dilip Krishnan。监督式对比学习。神经信息处理系统进展，33:18661–18673，2020年。

Simon Kornblith、Jonathon Shlens 和 Quoc V Le。更好的 ImageNet 模型是否具有更强的迁移能力？载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第2661–2671页，2019年。Jonathan Krause、Jia Deng、Michael Stark 和李飞飞。收集一个大规模细粒度汽车数据集。2013年。Alex Krizhevsky、Geoffrey Hinton 等人。从微小图像中学习多层特征。2009年。Alex Krizhevsky、Ilya Sutskever 和 Geoffrey E Hinton。使用深度卷积神经网络进行 ImageNet 分类。《ACM通讯》，60(6):84–90，2017年。Hankook Lee、Kibok Lee、Kimin Lee、Honglak Lee 和 Jinwoo Shin。通过增强感知自监督提升表征的迁移能力。载于神经信息处理系统进展，34: 17710–17722，2021年。Kyungmin Lee 和 Jinwoo Shin。Renyicl：使用偏斜 Rényi 散度进行对比表征学习，2023年。Alexander C Li、Alexei A Efros 和 Deepak Pathak。理解非对比孪生表征学习中的坍塌。载于欧洲计算机视觉会议，第490–505页。施普林格，2022a。Junnan Li、Pan Zhou、Caiming Xiong 和 Steven CH Hoi。无监督表征的原型对比学习。《style id='9'>arXiv预印本 arXiv:2005.04966，2020年。Lujun Li 和 Anggeng Li。A2-aug：自适应自动数据增强。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第2266–2273页，2023年。Tianhong Li、Peng Cao、Yuan Yuan、Lijie Fan、Yuzhe Yang、Rogerio S Feris、Piotr Indyk 和 Dina Katabi。面向长尾识别的定向监督对比学习。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第6918–6928页，2022b。Wei Li、Jiahao Xie 和 Chen Change Loy。用于自监督视觉预训练的相关性图像建模。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第15105–15115页，2023年。Sungbin Lim、Ildoo Kim、Taesup Kim、Chiheon Kim 和 Sungwoong Kim。快速自动增强。《style id='17'>神经信息处理系统进展，32，2019年。Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Dollár 和 C Lawrence Zitnick。Microsoft COCO：情境中的常见物体。载于计算机视觉—*ECCV 2014*：第*13*届欧洲计算机视觉会议，苏黎世，瑞士，*2014*年*9*月*6*日至*12*日，论文集，第*V*部分 *13*，第740–755页。施普林格，2014年。Raphael Gontijo Lopes、Dong Yin、Ben Poole、Justin Gilmer 和 Ekin D Cubuk。通过图像块高斯增强，在不牺牲准确率的情况下提升鲁棒性。《style id='21'>arXiv预印本 arXiv:1906.02611，2019年。Subhransu Maji、Esa Rahtu、Juho Kannala、Matthew Blaschko 和 Andrea Vedaldi。Aircraft 的细粒度视觉分类。《style id='23'>arXiv预印本 arXiv:1306.5151，2013年。Ishan Misra 和 Laurens van der Maaten。前置任务不变表征的自监督学习。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第6707–6717页，2020年。Maria-Elena Nilsback 和 Andrew Zisserman。大规模类别自动花卉分类。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。Aaron van den Oord、Yazhe Li 和 Oriol Vinyals。基于对比预测编码的表征学习。《style id='29'>arXiv预印本 arXiv:1807.03748，2018年。Omkar M Parkhi、Andrea Vedaldi、Andrew Zisserman 和 CV Jawahar。猫和狗。载于*2012*年*IEEE*计算机视觉与模式识别会议，第3498–3505页。IEEE，2012年。

Hieu Pham、Zihang Dai、Qizhe Xie 和 Quoc V Le。元伪标签。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第11557–11568页，2021年。Joshua Robinson、Ching-Yao Chuang、Suvrit Sra 和 Stefanie Jegelka。使用困难负样本进行对比学习。《style id='3'>arXiv预印本 arXiv:2010.04592，2020年。Olga Russakovsky、Jia Deng、Hao Su、Jonathan Krause、Sanjeev Satheesh、Sean Ma、Zhiheng Huang、Andrej Karpathy、Aditya Khosla、Michael Bernstein 等人。ImageNet 大规模视觉识别挑战。《style id='5'>《国际计算机视觉期刊》，115:211–252，2015年。Ikuro Sato、Hiroki Nishimura 和 Kensuke Yokoi。APAC：使用神经网络进行增强模式分类。《style id='7'>arXiv预印本 arXiv:1505.03229，2015年。Yonglong Tian、Dilip Krishnan 和 Phillip Isola。对比多视图编码。载于计算机视觉—*ECCV 2020*：第*16*届欧洲计算机视觉会议，英国格拉斯哥，*2020*年*8*月*23*日至*28*日，论文集，第*XI*部分 *16*，第776–794页。施普林格，2020a。Yonglong Tian、Chen Sun、Ben Poole、Dilip Krishnan、Cordelia Schmid 和 Phillip Isola。什么样的视图才适合对比学习？《style id='11'>神经信息处理系统进展，33:6827–6839，2020b。Rogier Van der Sluijs、Nandita Bhaskhar、Daniel Rubin、Curtis Langlotz 和 Akshay S Chaudhari。探索适用于胸部X光片孪生表征学习的图像增强。载于深度学习医学影像，第444–467页。PMLR，2024年。Catherine Wah、Steve Branson、Peter Welinder、Pietro Perona 和 Serge Belongie。Caltech-UCSD Birds-200-2011 数据集。2011年。Li Wan、Matthew Zeiler、Sixin Zhang、Yann Le Cun 和 Rob Fergus。使用 DropConnect 对神经网络进行正则化。载于国际机器学习会议，第1058–1066页。PMLR，2013年。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。Xiao Wang 和 Guo-Jun Qi。使用更强的数据增强进行对比学习。《style id='19'>arXiv预印本 arXiv:2104.07713，2021年。Zihu Wang、Hanbin Hu、Chen He 和 Peng Li。使用半监督对比学习，结合优化的潜在表征学习与数据增强，识别晶圆图模式。载于*2023*年 *IEEE*国际测试会议（*ITC*），第141–150页。IEEE，2023年。Zihu Wang、Lingqiao Liu、Scott Ricardo Figueroa Weston、Samuel Tian 和 Peng Li。利用合成数据学习用于自监督细粒度视觉识别的判别性特征。《style id='23'>arXiv预印本 arXiv:2407.14676，2024年。Xi Weng、Yunhao Ni、Tengwei Song、Jie Luo、Rao Muhammad Anwer、Salman Khan、Fahad Shahbaz Khan 和 Lei Huang。在自监督学习中调制频谱。《style id='25'>arXiv预印本 arXiv:2305.16789，2023年。Yuxin Wu、Alexander Kirillov、Francisco Massa、Wan-Yen Lo 和 Ross Girshick。Detectron2。 `https://github.com/facebookresearch/detectron2`，2019年。Jianxiong Xiao、James Hays、Krista A Ehinger、Aude Oliva 和 Antonio Torralba。SUN数据库：从修道院到动物园的大规模场景识别。载于*2010*年*IEEE*计算机学会计算机视觉与模式识别会议，第3485–3492页。IEEE，2010年。Tete Xiao、Xiaolong Wang、Alexei A Efros 和 Trevor Darrell。对比学习中哪些内容不应进行对比。《style id='31'>arXiv预印本 arXiv:2008.05659，2020年。

Yuxuan Yin、Yu Wang 和 Peng Li。通过采用优化的未标记数据采样的半监督学习进行高维贝叶斯优化。*arXiv*预印本 *arXiv:2305.02614*，2023年。

Seungil You、David Ding、Kevin Canini、Jan Pfeifer 和 Maya Gupta。深度格网络与部分单调函数。神经信息处理系统进展，30，2017年。

Yuning You、Tianlong Chen、Yang Shen 和 Zhangyang Wang。自动化图对比学习。发表于国际机器学习会议，第12121–12132页。PMLR，2021年。

Junbo Zhang 和 Kaisheng Ma。重新思考对比学习中的增强模块：利用扩展视图学习层次化增强不变性。发表于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第16650–16659页，2022年。

Tong Zhang、Congpei Qiu、Wei Ke、Sabine Süsstrunk 和 Mathieu Salzmann。《利用局部与全局表征：一种新的自监督学习策略》。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 16580–16589 页，2022年。

A 使用替代的一致性损失改进 CoCor

原始一致性损失定义如下：

$$\mathcal{L}_{\text{consistent}}(\boldsymbol{\theta_e}|\boldsymbol{\theta_d}) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{X},a_c\sim\boldsymbol{\Omega_c}}\left[|l_d(\boldsymbol{x};f_{\boldsymbol{\theta_e}},a_c) - g_{\boldsymbol{\theta_d}}(\boldsymbol{v_c}(a_c))|\right] \tag{6}$$

然而，在我们的初步实验中，我们发现，与公式 6 相比，以下一致性损失能带来更好的下游任务性能。

$$\mathcal{L}_{\text{consistent}}(\boldsymbol{\theta_e}|\boldsymbol{\theta_d}) = \mathbb{E}_{\boldsymbol{\Omega_c^{<l>}}\sim\boldsymbol{\Omega_c}}[\text{softplus}[\mathbb{E}_{\boldsymbol{x}\sim\mathcal{X},a_{c,i}^{<l>}\sim\boldsymbol{\Omega_c^{<l>}}}[l_d(\boldsymbol{x};f,a_{c,i}^{<l>}) - g_{\boldsymbol{\theta_d}}(\boldsymbol{v_c}(a_{c,i}^{<l>}))]]] \tag{7}$$



图6：来自同一数据集类别的图像。目标物体的大小和颜色可能不同，在图像中的位置也可能不同。因此，相同的增强对不同图像可能产生不同的效果。此处以 ImageNet 为例进行说明。

其动机是减少视觉数据集中图像之间的高方差。因此，完全相同的复合增强可能会在特征空间中产生不同的潜在相似度。例如，如图6所示，不同图像中的目标物体可能具有不同的大小和形状，且位于不同位置。因此，强制所有增强视图共享相同的最优相似度，未必能带来最佳性能。为此，公式 7 计算每种增强组合长度下一个小批量中的平均潜在相似度。因此，公式 7 对批次中的平均潜在相似度施加数据增强一致性约束，而非针对每张图像施加约束，从而有助于更好地处理方差。

此外，能够大幅扭曲一张图像的数据增强，未必能同样大幅扭曲另一张图像的恒等变换。然而，绝对值损失会使这些扭曲较小的视图远离原始数据的表征。因此，我们使用 Softplus 替代绝对值计算，因为对于负数，Softplus 产生的梯度较小。

B MMNN 更新规则的推导

在我们的实现中，编码器的参数（$\boldsymbol{\theta_e}$）和 MMNN 的参数（$\boldsymbol{\theta_d}$）按照公式 8 交替更新。

$$\begin{aligned}\boldsymbol{\theta}'_{\boldsymbol{e}} &= \boldsymbol{\theta_e} - \eta_e \cdot \nabla_{\boldsymbol{\theta_e}}\mathcal{L}_u(\boldsymbol{\theta_e}|\boldsymbol{\theta_d}) \\ \boldsymbol{\theta}'_{\boldsymbol{d}} &= \boldsymbol{\theta_d} - \eta_d \cdot \nabla_{\boldsymbol{\theta_d}}\text{CE}(\boldsymbol{\theta}'_{\boldsymbol{e}}(\boldsymbol{\theta_d}))\end{aligned} \tag{8}$$

利用整体损失函数优化编码器时，$\boldsymbol{\theta_e}$ 以学习率 η_e 进行的更新可写为：

$$\boldsymbol{\theta}'_{\boldsymbol{e}} = \boldsymbol{\theta_e} - \eta_e \cdot (\nabla_{\boldsymbol{\theta_e}}\mathcal{L}_{\text{contrast}} + \nabla_{\boldsymbol{\theta_e}}\mathcal{L}_{\text{consistent}}) \tag{9}$$

应用链式法则后， $\boldsymbol{\theta_d}$ 以学习率 η_d 进行的更新可改写为:

$$\begin{aligned} \boldsymbol{\theta_d'} &= \boldsymbol{\theta_d} - \eta_d \cdot (\nabla_{\boldsymbol{\theta_d}} \mathrm{CE}(\boldsymbol{\theta_e'}(\boldsymbol{\theta_d})), \\ \text{where}: \quad \nabla_{\boldsymbol{\theta_d}} \mathrm{CE}(\boldsymbol{\theta_e'}(\boldsymbol{\theta_d})) &= \frac{\partial}{\partial \boldsymbol{\theta_e'}} \mathrm{CE}(\boldsymbol{\theta_e'}) \cdot \frac{\partial \boldsymbol{\theta_e'}}{\partial \boldsymbol{\theta_d}} \end{aligned} \tag{10}$$

在公式 10 中， $\frac{\partial}{\partial \boldsymbol{\theta_e'}} \mathrm{CE}(\boldsymbol{\theta_e'})$ 可通过反向传播计算。因此，我们重点计算 $\frac{\partial \boldsymbol{\theta_e'}}{\partial \boldsymbol{\theta_d}}$。

—

首先，我们按照 Pham 等人（2021年）和 Yin 等人（2023年）采用的方法，将公式 10 中的 $\boldsymbol{\theta_e'}$ 替换为公式 9 中经过一步更新的 $\boldsymbol{\theta_e'}$。此外，我们也像 Pham 等人（2021年）那样，忽略 $\boldsymbol{\theta_e}$ 对 $\boldsymbol{\theta_d}$ 的高阶依赖。因此， $\frac{\partial \boldsymbol{\theta_e'}}{\partial \boldsymbol{\theta_d}}$ 可展开如下: —

$$\begin{aligned} \frac{\partial \boldsymbol{\theta_e'}}{\partial \boldsymbol{\theta_d}} &= \frac{\partial}{\partial \boldsymbol{\theta_d}} (\boldsymbol{\theta_e} - \eta_e \cdot (\nabla_{\boldsymbol{\theta_e}} \mathcal{L}_{\text{contrast}} + \nabla_{\boldsymbol{\theta_e}} \mathcal{L}_{\text{consistent}})) \\ &= -\eta_e \cdot \frac{\partial}{\partial \boldsymbol{\theta_d}} \frac{\partial}{\partial \boldsymbol{\theta_e}} \mathcal{L}_{\text{consistent}} \\ &= -\eta_e \cdot \frac{\partial}{\partial \boldsymbol{\theta_d}} \frac{\partial}{\partial \boldsymbol{\theta_e}} \mathbb{E}_{l \in \ell} [\text{softplus}[\mathbb{E}_{i \in \mathcal{B}} \ [(f_{\boldsymbol{\theta_e}}(\boldsymbol{x_i^l})^T \cdot f_{\boldsymbol{\theta_e}}(\boldsymbol{x_i})) - g_{\boldsymbol{\theta_d}}(\boldsymbol{v_c}(a_{c,i}^{<l>}))]] \end{aligned} \tag{11}$$

其中， $\boldsymbol{x_i^l} = a_{c\,i}^{<l>}$, $(\boldsymbol{x_i})$ 表示增强数据， ℓ 表示所用复合增强的长度集合，而$\mathcal{B}$ 表示小批量中的数据索引集合。

为便于后续部分的推导，我们定义:

$$\begin{aligned} k^l &= \mathbb{E}_{i \in \mathcal{B}} [(f_{\theta_e}(\boldsymbol{x_i^l})^T \cdot f_{\boldsymbol{\theta_e}}(\boldsymbol{x_i})) - g_{\boldsymbol{\theta_d}}(\boldsymbol{v_c}(a_{c,i}^{<l>})] \\ \mathrm{sim}^l &= \mathbb{E}_{i \in \mathcal{B}} [(f_{\boldsymbol{\theta_e}}(\boldsymbol{x_i^l})^T \cdot f_{\boldsymbol{\theta_e}}(\boldsymbol{x_i}))] \end{aligned} \tag{12}$$

接着，我们将公式 11 与公式 12 合并:

$$\begin{aligned} \frac{\partial \boldsymbol{\theta_e'}}{\partial \boldsymbol{\theta_d}} &= -\eta_e \cdot \mathbb{E}_{l \in \ell} [\frac{\partial}{\partial \boldsymbol{\theta_d}} (\frac{-e^{k^l}}{1 + e^{k^l}} \cdot \frac{\partial}{\partial \boldsymbol{\theta_e}} \mathrm{sim}^l)] \\ &= -\eta_e \cdot \mathbb{E}_{l \in \ell} [\frac{e^{k^l}}{(1 + e^{k^l})^2} \cdot \frac{\partial}{\partial \boldsymbol{\theta_e}} \mathrm{sim}^l \cdot \frac{\partial}{\partial \boldsymbol{\theta_d}} \mathbb{E}_{i \in \mathcal{B}} [g_{\boldsymbol{\theta_d}}(\boldsymbol{v_c}(a_{c,i}^{<l>}))]] \end{aligned} \tag{13}$$

将公式13代入公式10，可得:

$$\nabla_{\boldsymbol{\theta_d}} \mathrm{CE}(\boldsymbol{\theta_e'}(\boldsymbol{\theta_d})) = \frac{\partial}{\partial \boldsymbol{\theta_e'}} \mathrm{CE}(\boldsymbol{\theta_e'}) \cdot \ (-\eta_e \cdot \mathbb{E}_{l \in \ell} [\frac{e^{k^l}}{(1 + e^{k^l})^2} \cdot \frac{\partial}{\partial \boldsymbol{\theta_e}} \mathrm{sim}^l \cdot \frac{\partial}{\partial \boldsymbol{\theta_d}} \mathbb{E}_{i \in \mathcal{B}} [g_{\boldsymbol{\theta_d}}(\boldsymbol{v_c}(a_{c,i}^{<l>}))]]) \tag{14}$$

此外，在公式14中， $\frac{\partial}{\partial \boldsymbol{\theta_e'}} \mathrm{CE}(\boldsymbol{\theta_e'})$可以用一阶泰勒展开近似表示:

$$\mathrm{CE}(\boldsymbol{\theta_e'}) - \mathrm{CE}(\boldsymbol{\theta_e}) \approx (\boldsymbol{\theta_e'} - \boldsymbol{\theta_e}) \cdot \frac{\partial}{\partial \boldsymbol{\theta_e'}} \mathrm{CE}(\boldsymbol{\theta_e'}) \tag{15}$$

因此，我们有:

$$\frac{\partial}{\partial \boldsymbol{\theta_e'}} \mathrm{CE}(\boldsymbol{\theta_e'}) \approx \frac{\mathrm{CE}(\boldsymbol{\theta_e'}) - \mathrm{CE}(\boldsymbol{\theta_e})}{\boldsymbol{\theta_e'} - \boldsymbol{\theta_e}} = \frac{\mathrm{CE}(\boldsymbol{\theta_e'}) - \mathrm{CE}(\boldsymbol{\theta_e})}{-\eta_e \cdot (\nabla_{\boldsymbol{\theta_e}} \mathcal{L}_{\text{contrast}} + \nabla_{\boldsymbol{\theta_e}} \mathcal{L}_{\text{consistent}})} \tag{16}$$

现在，我们可以将公式14改写为：

$$\nabla_{\boldsymbol{\theta_d}}\mathrm{CE}(\boldsymbol{\theta}_{\boldsymbol{e}}^{'}(\boldsymbol{\theta_d})) \approx \frac{\mathrm{CE}(\boldsymbol{\theta}_{\boldsymbol{e}}^{\prime}) - \mathrm{CE}(\boldsymbol{\theta_e})}{(\nabla_{\boldsymbol{\theta_e}}\mathcal{L}_{\mathrm{contrast}} + \nabla_{\boldsymbol{\theta_e}}\mathcal{L}_{\mathrm{consistent}})} \cdot \mathbb{E}_{l\in\ell}[\frac{e^{k^l}}{(1+e^{k^l})^2} \cdot \frac{\partial}{\partial \boldsymbol{\theta_e}}\mathrm{sim}^l \cdot \frac{\partial}{\partial \boldsymbol{\theta_d}}\mathbb{E}_{i\in\mathcal{B}}[g_{\boldsymbol{\theta_d}}(\boldsymbol{v_c}(a_{c,i}^{<l>}))]] \tag{17}$$

此外，公式17还可以进一步简化。注意，$\nabla_{\boldsymbol{\theta_e}}\mathcal{L}_{\mathrm{contrast}} + \nabla_{\boldsymbol{\theta_e}}\mathcal{L}_{\mathrm{consistent}}$ 可以近似为：

$$\nabla_{\boldsymbol{\theta_e}}\mathcal{L}_{\mathrm{contrast}} + \nabla_{\boldsymbol{\theta_e}}\mathcal{L}_{\mathrm{consistent}} := \nabla_{\boldsymbol{\theta_e}}\mathcal{L}_u(\boldsymbol{\theta_e}) \approx \frac{\mathcal{L}_u(\boldsymbol{\theta}_{\boldsymbol{e}}^{\prime}) - \mathcal{L}_u(\boldsymbol{\theta_e})}{\boldsymbol{\theta}_{\boldsymbol{e}}^{\prime} - \boldsymbol{\theta_e}} \tag{18}$$

同样，$\frac{\partial}{\partial \boldsymbol{\theta_e}}\mathrm{sim}^l$ 的近似形式可以写为：

$$\begin{gathered}\frac{\partial}{\partial \boldsymbol{\theta_e}}\mathrm{sim}^l \approx \frac{\mathrm{sim}^l(\boldsymbol{\theta}_{\boldsymbol{e}}^{\prime}) - \mathrm{sim}^l(\boldsymbol{\theta_e})}{\boldsymbol{\theta}_{\boldsymbol{e}}^{\prime} - \boldsymbol{\theta_e}} \\ \text{where:}\quad \mathrm{sim}^l(\boldsymbol{\theta}_{\boldsymbol{e}}^{\prime}) = \mathbb{E}_{i\in\mathcal{B}}[(f_{\boldsymbol{\theta}_{\boldsymbol{e}}^{\prime}}(\boldsymbol{x_i^l})^T \cdot f_{\boldsymbol{\theta}_{\boldsymbol{e}}^{\prime}}(\boldsymbol{x_i}))]\end{gathered} \tag{19}$$

最后，将公式18和公式19与公式17结合，$\nabla_{\boldsymbol{\theta_d}}\mathrm{CE}(\boldsymbol{\theta}_{\boldsymbol{e}}^{'}(\boldsymbol{\theta_d}))$可以改写如下：

$$\nabla_{\boldsymbol{\theta_d}}\mathrm{CE}(\boldsymbol{\theta}_{\boldsymbol{e}}^{'}(\boldsymbol{\theta_d})) \approx (\mathrm{CE}(\boldsymbol{\theta}_{\boldsymbol{e}}^{\prime}) - \mathrm{CE}(\boldsymbol{\theta_e})) \cdot \mathbb{E}_{l\in\ell}[\frac{e^{k^l}}{(1+e^{k^l})^2} \cdot \frac{\mathrm{sim}^l(\boldsymbol{\theta}_{\boldsymbol{e}}^{\prime}) - \mathrm{sim}^l(\boldsymbol{\theta_e})}{\mathcal{L}_u(\boldsymbol{\theta}_{\boldsymbol{e}}^{\prime}) - \mathcal{L}_u(\boldsymbol{\theta_e})} \cdot \frac{\partial}{\partial \boldsymbol{\theta_d}}\mathbb{E}_{i\in\mathcal{B}}[g_{\theta_d}(\boldsymbol{V}(A))]] \tag{20}$$

注意，在公式20中，所有导数项都可以通过反向传播直接计算，从而得到一种可扩展算法。

C CoCor 中的候选增强

用于构成复合增强的基本增强集合 Ω_a 由 14 种不同类型的数据增强组成，即 Ω_a = {`AutoContrast`, `Brightness`, `Color`, `Contrast`, `Rotate`, `Equalize`, `Identity`, `Posterize`, `Sharpness`, `ShearX`, `ShearY`, `Solarize`, `TranslateX`, `TranslateY`}。

在第 5.4 节中，我们对基本数据增强的组成进行了消融研究，如表 6 所示。在这些实验中，Ω_a 被分为两个集合：Ω_{color} 和 Ω_{affine}。两个集合的组成如下：Ω_{color} = {`AutoContrast`, `Brightness`, `Color`, `Contrast`, `Equalize`, `Identity`, `Posterize`, `Sharpness`, `Solarize`, } 和 Ω_{affine} = {`Rotate`, `Identity`, `ShearX`, `ShearY`, `TranslateX`, `TranslateY`}。

D 预训练的实验设置

所有预训练实验均使用随机梯度下降（SGD）作为优化器。采用余弦衰减调度器来调整学习率。所有预训练均采用 256 的训练数据批量大小。各方法所使用的详细设置如下。

- **MoCo** He 等（2020）；Chen 等人（2020b）：优化器的初始学习率设为 3×10^{-2}，权重衰减设为 1×10^{-4}，动量设为 0.9。对于 MoCo 的双编码器，特征空间维度为 128，内存队列大小为 65536，更新键编码器的动量为 0.999。对于一致性损失，查询编码器同时以原始数据及其增强视图作为输入。不过，查询编码器在对原始输入进行编码时会停止梯度。对增强视图进行编码时，编码器会接收梯度。

• **SimSiam** Chen 和 He（2021）：SGD 优化器的初始学习率设为 5×10^{-2}，权重衰减设为 1×10^{-4}，动量设为 0.9。预测器、投影器和输出维度均为 2048。对于一致性损失，原始数据的表征通过让数据依次经过停止梯度的骨干网络和投影器来获得。增强视图则经过骨干网络、投影器和预测器，以获得其表征。在计算增强视图的表征时，这三个模型都会接收梯度。

• **SupCon** Khosla 等人（2020）：我们为优化器设置了 3×10^{-2} 的初始学习率、1×10^{-4} 的权重衰减和 0.9 的动量。投影头输出的特征空间维度为 128。在一致性损失中，编码器以原始数据及其增强视图作为输入。对原始输入进行编码时，编码器停止梯度。计算增强视图的表征时，编码器接收梯度。

E 线性评估的实验设置

我们遵循 Chen 等（2020a）、Lee 等人（2021）和 Kornblith 等人（2019）的线性评估协议。在冻结的预训练骨干网络上训练一个单层线性分类器。 `Resize`、 `CentorCrop` 和`Normalization` 用作训练变换。采用 L-BFGS 优化器来最小化经过 ℓ_2 正则化的交叉熵损失。在每个数据集的验证集上选择最优正则化参数。最后，使用最优正则化参数训练模型，并报告所得的测试准确率。

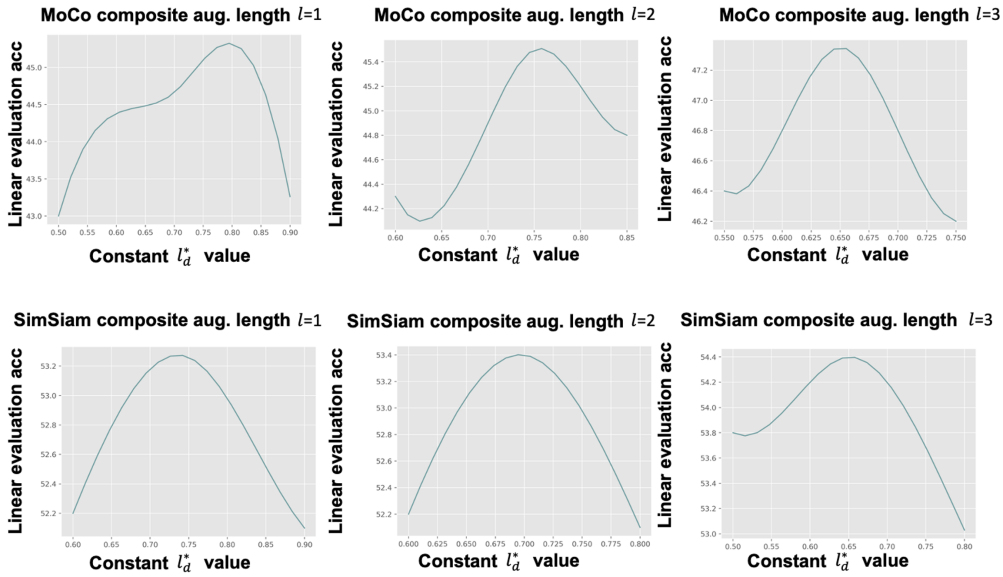


图 7：仅使用一种长度的复合增强训练的 CoCor 模型的线性性能准确率。使用多个不同的常数 l_d^*值运行采用各个长度复合增强的 CoCor。

F 不含 MMNN 的 CoCor 消融研究

为了评估 MMNN 的效果，我们运行 CoCor，并将 l_d^* 替换为固定常数。实际操作中，计算一致性损失时，每种复合增强长度都使用一个常数。我们通过多次实验，为每种复合增强长度选出相应的常数。所有模型均为在 ImageNet-100 上预训练 50 个训练轮数的 ResNet-50。图 7 展示了 CoCor 使用一种复合数据增强长度并搭配不同常数 l_d^*运行时的线性评估结果。对于每种复合增强长度，都存在一个能在线性评估中取得最佳结果的最优常数。这些

最优常数随后用于与 MMNN 进行比较。MoCo 和 SimSiam 的最佳常数列于表 9。

ℓ	1	2	3
MoCo	0.80	0.75	0.65
SimSiam	0.75	0.70	0.60

表 9：用于与 MMNN 比较的常数 l_d^*

G CoCor使用更强的数据增强

CoCor 是一种系统性方法，旨在利用多种增强所提供的信息。为了进一步测试不同强度的复合增强的效果，我们在 MoCo 和 SimSiam 上运行 CoCor，单次增强长度分别为 1、2、3、5 和 10。我们在 ImageNet-100 上训练 ResNet-50，共训练 50 轮。我们在 ImageNet-100 上对预训练骨干网络进行线性分类评估。表 10 展示了评估准确率。

ℓ	1	2	3	5	10	基线
MoCo	47.90	47.70	48.08	45.16	43.07	42.66
SimSiam	54.01	54.20	55.16	53.50	52.70	44.60

表 10：仅在一种长度的复合增强上应用一致性损失所训练编码器的线性评估 l 。

如表 10 所示，在预训练中引入相对较弱（较短）的数据增强，可提升模型在下游任务上的表现。在各种下游任务上评估预训练模型时，用于测试模型性能的测试数据是自然且真实的图像。因此，在预训练过程中向编码器提供失真程度较低的视图，有助于编码器在推理时更好地识别自然图像。

还值得注意的是，CoCor 可以利用长度为 10 的极强增强来提升编码器相较于基线的性能，其强度甚至超过了近期研究中使用的数据增强，例如采用强增强的 CLSA Wang 和 Qi（2021）以及 R´enyiCL Lee 和 Shin（2023）。尽管已有研究表明，过强的数据增强可能会过度扭曲数据的恒等特征，从而给训练过程带来过多噪声。这些噪声可能导致性能下降 Chen 等（2020a）；Tian 等（2020b），部分维度坍塌，甚至完全维度坍塌 Li 等（2022a）；Jing 等（2021）。然而，CoCor 能够从高度失真的视图中捕获关键信息，并利用这些强增强丰富特征空间的语义。

H CoCor 与更多基线对比学习方法的兼容性。

为展示 CoCor 与更多基线方法（例如 BYOL（Grill 等人（2020））和 INTL（Weng 等人（2023））的兼容性，我们在这些额外的基线方法上实现了 `CoCor`。表 11 展示了这些编码器在各种数据集上的线性评估结果。这些结果表明，CoCor 不仅提升了 MoCo、SimSiam 和 SupCon 等对比方法所学习表征的泛化能力，还将这些优势扩展到了其他现有的对比学习方法。

I 训练时间比较

`CoCor`将所提出的一致性损失应用于由各种数据增强生成的视图。为了解决关于 `CoCor` 相较于对应基线方法会额外耗费训练时间的顾虑，我们在表 12 中比较了预训练时间和模型性能。尽管 CoCor 每个训练轮次所需的时间更长，

表 11：线性评估的 Top-1 准确率（%）。所有 ResNet-50 骨干编码器均在 ImageNet-100 上预训练。

方法	IN-100	CIFAR-100	CUB200	Caltech-101	SUN397	Food101
BYOL Grill 等人（2020）	68.05	60.13	20.27	78.90	39.03	53.01
BYOL + CoCor（**我们的方法**）	**72.30**	**62.42**	**22.75**	**80.51**	**42.64**	**55.32**
INTL Weng 等人（2023）	74.58	62.73	29.05	87.45	47.06	60.38
INTL + CoCor（**我们的方法**）	**76.98**	**63.06**	**31.40**	**88.81**	**48.27**	**64.27**

CoCor 仍能以更短的训练时间取得优于对应基线方法的性能。此外，CoCor 的训练耗时与使用强多视图增强的 CLSA Wang 和 Qi（2021年）相近。然而，在训练时间相近的情况下，CoCor 的性能显著优于 CLSA。

表 12：在配备 4 个 A100 GPU 的机器上的训练时间比较。为便于比较，给出了 ImageNet-1K 上的训练时间（小时）/Top-1 线性评估结果（%）。

方法 \ 训练轮数	50	100	200
MoCo He 等（2020）	13.4h/64.78	26.8h/66.81	53.6h/67.22
SimSiam Chen 和 He（2021年）	13.2小时/66.89	26.4小时/68.12	52.8小时/70.93
CLSA Wang 和 Qi（2021年）	18.1小时/67.95	36.2h/69.15	72.4h/71.19
MoCo + CoCor	18.5h/69.26	37.0h/70.76	74.0h/72.83
SimSiam + CoCor	18.1小时/70.90	36.2小时/72.03	72.4小时/73.25

J CoCor 中数据增强顺序的影响。

为了评估 CoCor 中数据增强顺序的影响，我们针对多种固定的数据增强顺序进行了实验。我们使用具有特定数据增强顺序的 MoCo + CoCor 训练编码器。结果如表13所示。例如，在 CoCor-fixed 1/2 配置中，数据增强的顺序是预先设定的。在固定顺序的情形下，例如 CoCor-fixed 1，只要图像同时选择了旋转和亮度增强，就始终先应用旋转。表13中的结果表明，在 CoCor 中，数据增强的顺序不会显著影响编码器在下游任务上的性能。因此，在最终实验中，我们采用随机的数据增强顺序，以探索更广泛的数据增强变化。

表13：采用不同固定数据增强顺序训练的 CoCor 模型在 ImageNet-100 上的线性评估结果。编码器在 ImageNet-100 上训练 200个训练轮次。

	MoCo	CoCor - 固定 1	CoCor - 固定 2
Top-1准确率	67.04	71.32	71.47