arXiv:2302.02502v2 [cs.LG] 6 Jun 2023

从对抗视角重新思考鲁棒对比学习

Fatemeh Ghofrani*、**Mehdi Yaghouti***、**Pooyan Jamshidi**[1]

摘要

为了加深对鲁棒深度学习的理解，我们深入探究对抗训练对自监督和监督式对比学习以及监督学习的影响。我们的分析揭示，在各种学习算法中，标准训练网络的对抗表征与干净表征之间存在显著差异。值得注意的是，对抗训练能够减轻这些差异，并促进表征收敛至一个通用集合，而不受所用学习方案的影响。此外，增大对抗表征与干净表征之间的相似性，尤其是在网络末端附近，能够增强网络鲁棒性。这些发现为设计和训练有效且鲁棒的深度学习网络提供了宝贵见解。我们的代码已发布于
h
t
t
p
s://
github.com/softsys4ai/CL-Robustness。

1. 引言

近年来，自监督学习取得了显著进展，并在各种应用中实现了最先进的性能。虽然这为从海量无标签数据中学习有效表征铺平了道路，但易受对抗攻击影响仍然是致命威胁。事实证明，对抗训练是监督学习中的一种有效防御方法。这种方法可以解释为一个极小极大优化问题（Madry 等人，2017），其中，模型参数通过最小化针对对抗扰动的训练损失来迭代更新，而这些扰动则由最大化对手损失函数生成。通常的做法是使用相同的损失函数来进行训练和生成对抗攻击，但也有一些研究尝试使用不同的损失函数来探究鲁棒训练（Pal 等人，2021）。Hendrycks 等人（Hendrycks 等人，2019）在训练损失中引入了一个自监督项，以提高监督模型的鲁棒性。Chen 等人（Chen 等人，2020b）率先将对抗训练应用于自监督模型，从而获得鲁棒的预训练编码器，这些编码器可通过微调用于下游任务。

近年来，对比学习方案的鲁棒性及其对抗训练研究受到了广泛关注。对比学习（CL）的核心思想是通过比较语义相似与不相似的样本，学习恰当的表示。近期一些研究将生成的对抗样本用作与锚点数据相似的匹配项，以提升模型的鲁棒性。Kim 等人（Kim 等人，2020）率先利用对比损失生成对抗样本，在不使用任何标签的情况下增强 SimCLR（Chen 等人，2020a）框架的鲁棒性。Moshavash 等人（Moshavash 等人，2021）、Wahed 等人（Wahed 等人，2022）和 Gowal 等人（Gowal 等人，2021）分别将相同技术应用于动量对比（MOCO）（He 等人，2020）、视图间交换分配（SwAV）（Caron 等人，2020）和自举潜在表示（BYOL）（Grill 等人，2020）。Fan 等人（Fan 等人，2021）在对比损失中引入了额外的正则化项，以增强跨任务鲁棒性迁移能力。他们采用伪标签生成技术，避免在下游任务的对抗训练中使用标签。类似地，Jiang 等人（Jiang 等人，2020）考虑了两个对比损失函数的线性组合，以研究在不同配对选择场景下的鲁棒性[1]。

任何对比学习方案的核心步骤之一，都是选择正负样本对。在没有标签信息的情况下，正样本对通常通过数据增强获得，而负样本则从一个小批量中随机选取。然而，这种随机选取策略可能会导致选出假负样本对，即两个样本来自同一类别的情况。在（Gupta 等人，

[1] F. Ghofrani、M. Yaghouti 和 P. Jamshidi 任职于美国南卡罗来纳州哥伦比亚市南卡罗来纳大学工程与计算学院。电子邮箱：ghofrani@email.sc.edu、yaghouti@mailbox.sc.edu、pjamshid@cse.sc.edu。

*这些作者对本文贡献相同。

[1]本文中，“鲁棒”一词均为“对抗鲁棒”的简称。

2022）中，Gupta 等人通过实验证明，对比学习的对抗脆弱性与训练期间使用假负样本对有关。解决这一困境的一种方法是利用标签信息，将自监督对比损失扩展为由（Khosla 等人，2020）提出的监督对比（SupCon）损失。SupCon 损失将同一类别的嵌入表示作为正样本进行对比，并将不同类别的嵌入表示作为负样本进行对比。研究表明，在准确率和超参数稳定性方面，SupCon 优于交叉熵损失（Khosla 等人，2020）。Islam 等人（Islam 等人，2021）开展了一项广泛研究，比较了通过交叉熵、SupCon 和标准对比损失学习到的表征在多个下游任务中的可迁移性。Zhong 等人（Zhong 等人，2022）设计了一系列鲁棒性测试，包括数据损坏，涵盖从像素级伽马失真、图块级打乱到数据集级分布偏移，以量化对比学习与监督学习框架之间的差异。

近期的一些研究开始解释并理解鲁棒网络。Jones 等人（Jones 等人，2022）表明，无论架构或随机初始化如何，对抗鲁棒性都是对网络学习到的函数的一项重要约束。（Cianfarani 等人，2022）开展的研究揭示了对抗训练对表征的影响。研究强调了鲁棒表征缺乏专门化，以及过拟合对鲁棒训练过程中更深层产生的显著影响。然而，需要注意的是，该研究主要关注监督学习算法。为弥补这一研究空白，我们从统一的视角，探究鲁棒训练对采用不同学习方案（对比式、监督式对比学习和监督学习）训练的模型所产生的影响。

贡献在这项工作中，我们开展了多项全面实验，在不同训练场景下比较对比学习、监督式对比学习与标准监督学习的鲁棒性。我们采用 CKA（中心化核对齐）（Kornblith 等人，2019；Nguyen 等人，2020；Subramanian，2021）和线性探测（Alain 和 Bengio，2016）等分析工具，研究并比较通过各种训练方法学习到的逐层表征。实验的设计与实施由以下研究问题驱动：

Q1：从对抗鲁棒性的角度来看，对比学习学到的表征有什么特别之处吗？

问题2：利用标签信息在多大程度上有助于或损害对比学习表征的鲁棒性？

问题3：对抗训练会如何影响监督学习和对比学习所学到的表征（包括相似之处和差异）？

我们的主要发现可总结如下：s:

R1：我们的结果表明，**在标准训练中，不含标签信息的对比学习比其他学习方案的鲁棒性更差**。然而，将标准对比损失与监督交叉熵或监督对比损失相结合，可以利用标签信息提升学习到的表征的鲁棒性。（第3.1节）

R2：从我们的结果中可以观察到，**完整对抗微调对对比学习所学表征的鲁棒性有显著的积极影响**。然而，在监督对比学习或标准监督学习方案中，完整对抗微调并无效果。（第3.2.1节）

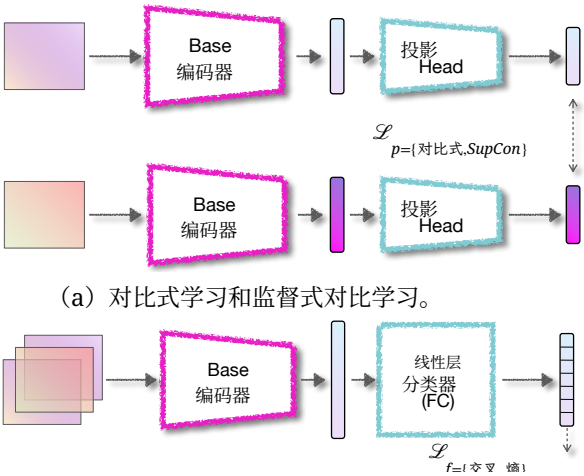
R3：我们的研究揭示了对抗训练对不同学习方案中表征的影响所带来的重要启示。我们观察到，**在各种学习方案中，标准训练网络的对抗表征与干净表征之间存在显著差异**。然而，**经过对抗训练后，我们观察到对抗表征与干净表征之间呈现出显著的相似性**。这表明，**无论采用何种学习方案，对抗训练都会促使表征向一个通用集合收敛，该集合的特征是**[2] **这些特征会在不同模型和任务中稳定出现**(Olah 等人，2020)。此外，我们发现，**提高对抗表征与干净表征之间的相似性，尤其是在网络末端提高这种相似性，可以提升网络的鲁棒性**。这些发现为设计和训练更高效、更有效的鲁棒网络提供了宝贵见解。（第3.2.2节）

2. 方法

本节将介绍我们对以下三种学习方案的鲁棒性进行比较研究的方法：

- 对比学习（*CL*）：在 SimCLR 的标准框架中，对比学习通过最小化投影到潜在空间中的表示之间的对比损失来训练基础编码器（图1a）。提取的特征将用于训练下游任务中的线性分类器。

[2]一种表征包含某一层中的所有特征。



（a）对比式学习和监督式对比学习。

（b）监督学习。

图*1*。所研究学习方案的训练过程。

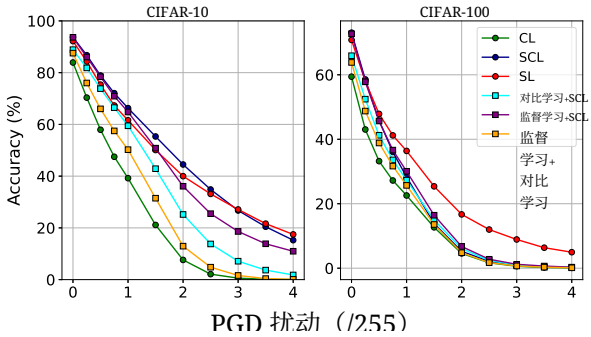


图*2*。**将标签信息融入对比学习，可以增强所得表征的鲁棒性**。我们在标准训练设置下，比较了不同学习方案在 CIFAR-10 和 CIFAR100 数据集上抵御对抗样本时的测试准确率。与其他学习方案相比，不使用标签的对比学习鲁棒性较低。我们还观察到，半监督学习方案 SL+CL 或 SCL+CL 的鲁棒性能优于 CL 方案。

- 监督式对比学习（*SCL*）：一种在（Khosla 等人，2020）中提出的对比学习监督式扩展，通过利用标签信息来避免选择错误正样本对。

- 监督学习（*SL*）：网络由基础编码器和一个作为线性分类器的全连接层组成（见图 1b）。在这种情况下，使用真实标签与预测标签之间的交叉熵来训练网络参数。

对比学习和监督式对比学习的训练过程包括以下两个阶段：

预训练阶段：该阶段的目标是通过在给定数据集 $\mathscr{D}_p$ 上最小化自监督损失 $\mathscr{L}_p(\boldsymbol{\theta}_b, \boldsymbol{\theta}_{ph})$，训练基础编码器参数 $\boldsymbol{\theta}_b$。其中，$\boldsymbol{\theta}_{ph}$ 是投影头的参数向量，该投影头用于将基础编码器的输出映射到低维潜在空间，并在该空间中应用 $\mathscr{L}_p$。

监督微调阶段：该阶段的目标是通过在带标签数据集 $\mathscr{D}_f$ 上最小化监督损失 $\mathscr{L}_f(\boldsymbol{\theta}_c)$，训练线性分类器参数 $\boldsymbol{\theta}_c$。线性分类器学习将预训练阶段提取的表征映射到标签空间，其中$\mathscr{L}_f$ 是交叉熵损失（有关方法的完整细节，请参见附录 A）。

我们考察上述训练阶段的标准训练和鲁棒训练变体，以比较不同学习方案之间的对抗鲁棒性。表A.1总结了附录 A 中针对对比学习和监督式对比学习方案在不同训练阶段可能训练场景下所研究的全部训练组合。

3. 实验

我们的目标是从对抗视角理解，与监督学习相比，对比学习从数据中学习表征的方式是否存在差异。为此，我们在表 A.1所示的不同训练场景下，对 CIFAR-10 和 CIFAR-100 图像分类基准上的对比学习（CL）、监督式对比学习（SCL）和监督学习（SL）的鲁棒性进行了广泛实验评估。在后续所有实验中，我们在同一数据集上训练基础编码器和线性分类器。实验设置见附录 B。我们使用两种威胁模型评估不同场景下的鲁棒性：威胁模型 I 涉及通过交叉熵损失生成的端到端攻击，攻击者完全了解基础编码器和线性分类器的架构及网络参数。在威胁模型-II 中，攻击专门针对基础编码器（详见附录 B.1）。

3.1. 对比学习：通过标准训练分析鲁棒性与标签的影响

众所周知，经过标准训练的神经网络容易受到对抗样本的攻击，但采用不同学习方法训练的模型，其易受攻击的程度可能有所不同。因此，我们旨在研究通过标准训练得到的各种学习算法的脆弱性，以评估其性能。威胁模型 I 下，结果随扰动大小变化的情况如图 2所示。我们使用 20 步l_∞ 投影梯度下降（PGD）（Madry 等人，2017）攻击，并采用不同的扰动生成对抗

攻击来开展本实验。除前文提到的学习方案外，我们还研究了它们的组合，包括监督学习与监督式对比学习的组合（记为 SL+SCL），以及另外两种半监督版本：对比学习与监督学习的组合（记为 SL+CL），或与监督式对比学习的组合（记为 CL+SCL）。这些组合旨在回答以下问题：利用标签信息是否有助于提升对比学习表征的鲁棒性？如我们所见，不使用标签信息的对比学习比其他学习方案的鲁棒性更差。我们还观察到，半监督学习方案 SL+CL 或 CL+SCL 的鲁棒性能优于 CL 方案。附录 C 使用 t-SNE 对 CIFAR-10 数据集上所有这些学习方案所学习到的表征进行了可视化。

为避免标签信息与对抗训练的影响之间可能产生的混淆，我们在后续实验中排除了半监督版本。附录 D提供了威胁模型-II 下的更多结果，其中攻击仅针对基础编码器生成。

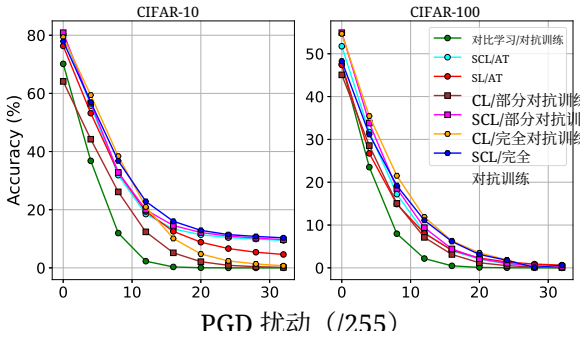


图*3*。 **完全对抗训练成功地将标签信息融入通过对比学习训练得到的鲁棒网络架构中，从而显著提升了整体鲁棒性。** 我们通过不同的对抗训练场景，比较了不同学习方案在 CIFAR-10 和 CIFAR100 数据集上应对不同 PGD攻击时的测试准确率。结果表明，完全对抗训练能够有效地将标签信息融入通过 CL 学习方案获得的对抗训练网络中，从而提升鲁棒性。此外，在不同对抗训练场景下，SCL 的性能略有差异。

3.2. 对抗训练：比较表征

在此，我们首先比较不同对抗训练场景下的性能。随后，采用一组解释工具（例如 CKA 和线性探测）来考察对抗训练如何影响隐藏层中学习到的表征。

3.2.1. 直接比较

本实验旨在评估以下场景在威胁模型 I 下的模型鲁棒性：i）使用对抗训练训练基础编码器，然后冻结基础编码器，单独训练线性分类器（对抗训练）；ii）通过对抗训练训练基础编码器，然后冻结基础编码器，并使用对抗训练单独训练线性分类器（部分对抗训练）；iii）通过对抗训练训练基础编码器，然后增强端到端模型的鲁棒性（完全对抗训练）。后一种情况是指，首先使用 $\mathscr{L}_{CL}$ 或 $\mathscr{L}_{SCL}$ 对基础编码器参数进行对抗训练，然后在对线性分类器进行对抗训练时微调这些参数，其中对抗样本使用

$\mathscr{L}_{CE}$。针对 CIFAR-10 和 CIFAR100 数据集，在 20 步不同 PGD攻击下的结果如图 3所示。附录 E还展示了在 CIFAR100 数据集上，针对一种名为 Auto-attack 的最先进对抗攻击（Croce与Hein，2020）时，不同场景下的鲁棒性。此外，附录 F 还展示了威胁模型-II下的更多结果，其中攻击仅针对基础编码器生成。结果表明，主要有以下两点发现：

（i）在各种评估场景中，完全对抗训练下的对比学习始终优于其他学习方案，在标准准确率和对抗攻击下的鲁棒性方面均有显著提升。上一节的结果表明，在标准训练期间利用标签信息能够提升对比学习方案的鲁棒性。值得注意的是，完全对抗训练还能有效地将标签信息融入鲁棒网络架构，从而显著提升整体鲁棒性。这一发现与先前的研究（Zhai等人，2019；Fan等人，2021）一致。这些研究探讨了如何利用无标签数据的伪标签，并证明伪标签有助于增强神经网络的对抗鲁棒性。通过融入标签信息，对比学习方案得以利用额外的宝贵监督信息，强化抵御对抗攻击的能力，并提升整体鲁棒性。（ii）在对比学习的对抗训练和完全对抗训练场景下，SCL 的性能略有差异。这表明，SCL 在对抗训练场景下学到的表征已足以实现可接受的鲁棒性。

3.2.2. 特征比较

先前的结果从表征学习的角度提出了一个重要问题：神经网络经过对抗训练后，各层会发生什么变化？我们考察了不同鲁棒学习方案所学到的内部层表征，以深入探讨这一问题。

对抗表征与干净样本的表征在标准训练网络中存在显著差异。我们首先通过以下方式展开研究——

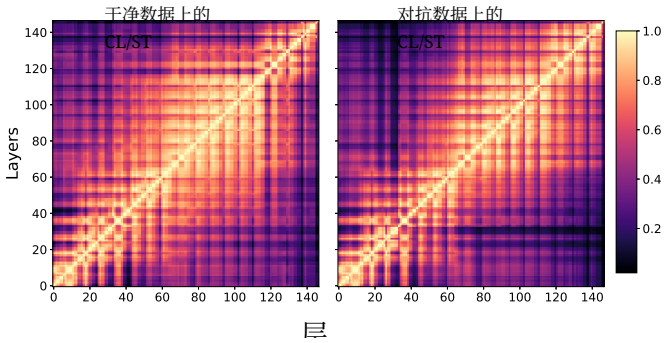


图*4.* **采用不同学习算法的标准训练网络所产生的表征，在对抗样本与干净样本之间呈现出显著差异。** 我们计算了采用 CL 方案训练的标准训练网络中所有层组合的表征相似度，同时考察了干净数据和对抗数据。

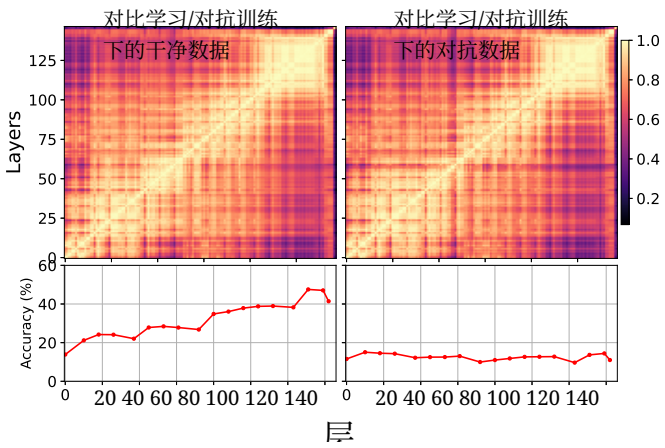


图*5*。 **无论采用何种学习方案，对抗训练网络中的对抗表征与干净表征都表现出显著的相似性。** 我们计算了采用对比学习方案训练的对抗训练网络中所有层组合之间表征的相似性，同时考虑干净数据和对抗数据。

使用 CKA 研究各模型的内部表征结构。CKA 是一种用于衡量两组特征之间相似性的指标。为了解不同学习方案如何提取表征，我们针对采用不同学习方案训练的模型，取其中每一对层 X 和 Y，计算它们在干净样本和对抗样本上的 CKA 相似性。图 4 和 G.1以热图形式展示了标准场景下，不同学习算法在干净样本和对抗样本上的结果。值得注意的是，我们观察到三种学习方案的内部表征结构存在明显差异：(i) 与 CL 相比，SCL 和 SL 方案中网络末端附近的层与其他层的相似性较低；(ii) 标准训练网络中干净样本与对抗样本的表征之间存在差异，凸显了这些网络易受对抗样本攻击的脆弱性。先前的研究（Mitrovic等人，2020）假设数据由内容和风格组成，且只有内容与未知的下游任务相关。此外，该假设还认为内容与风格彼此独立，也就是说，风格变化不会影响底层内容。相比之下，对抗扰动被视为自然数据分布中的一种特定分布变化（Zhang等人，2020），它们会在保留内容的同时改变风格，因此肉眼无法察觉。基于这些假设，标准训练网络中对抗样本与干净样本之间显著的差异表明，这类网络无法提取在两种样本上保持一致的内容相关表征。

无论采用何种学习方案，对抗训练网络中的对抗表征与干净表征都表现出很高的相似性。我们还对鲁棒模型进行了先前的比较，考察对抗训练模型内部每一对层 X 和 Y，鲁棒性.....

由不同学习方案训练，并计算其在干净样本和对抗样本上的 CKA 相似度。此外，我们将线性探测用作概念工具，以更好地理解神经网络内部的动态，以及各个中间层所发挥的具体作用。图5 和 G.2 展示了对抗训练实验结果中的几项观察：（i）无论采用何种学习方案，与标准训练相比，跨层相似性都有所增强。图中的亮度更高，清楚地体现了这一点。（ii）通过对抗训练得到的网络，其对抗表征与干净表征高度相似。先前的研究（Jones 等人，2022）表明，鲁棒训练能够有效减轻对抗扰动的影响，使鲁棒网络对干净样本和对抗样本产生相似的表征。然而，他们的研究仅关注监督学习方案。我们的结果支持并拓展了这些发现，表明无论采用何种学习方案，干净表征与对抗表征之间的相似性都成立。（iii）比较 AT 与其对应的完全对抗训练（见图 G.2）所获得的表征时，我们发现 CL 框架中的长程相似度显著提高。这种相似性的增强带来了标准准确率和对抗准确率的显著提升。值得注意的是，完全对抗训练通过将标签信息纳入网络，显著提升了整体鲁棒性。相比之下，在 AT 和完全对抗训练两种情形下，SCL 和 SL 学到的表征差异很小。这些细微的表征差异只带来了微乎其微的性能差异，表明 AT 中使用的标签信息已经提供了足够的

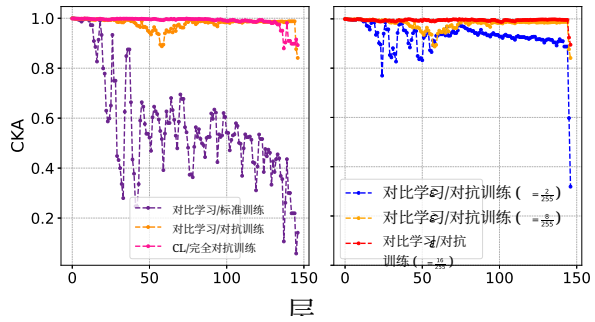


图6。**提高对抗表征与干净表征之间的相似性，尤其是在网络末端附近，可以提升鲁棒性**。 左）比较对比学习中的干净表征与对抗表征发现，标准训练网络中的二者存在显著差异。对抗训练显著缩小了这种差异，但在中间层观察到相似性下降，从而影响其性能（对比学习/对抗训练）。完全对抗训练有效缓解了相似性下降，增强了整体模型鲁棒性。右）在对抗训练过程中，我们提高扰动预算，以改变对抗攻击的强度。这使对抗表征与干净表征之间的相似性更高，尤其是在网络末端附近。

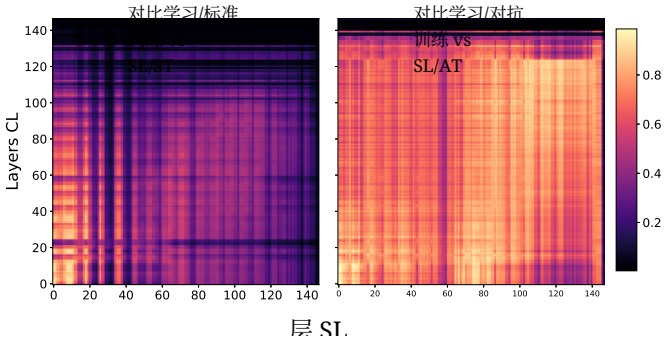


图7。**对抗训练促进了向通用表征集的收敛**。标准训练网络在不同对比学习方案之间表现出显著不同的对抗表征，尤其是在较高层中。然而，在应用对抗训练后，层之间的相似性增加，表明网络逐渐转向提取通用表征集。

鲁棒性，足以满足 SCL 和 SL 的需求。因此，完全对抗训练不会引入额外信息来进一步提升 SCL 和 SL 的鲁棒性。

提高对抗表征与干净表征之间的相似性，尤其是在网络末端附近，可以提升鲁棒性。 为深入理解对抗表征与干净表征之间的差异，我们比较模型在干净数据上各层 X 的输出与同一模型在对抗样本上对应层 Y 的输出。图6 和图G.3 左侧的结果表明，采用标准训练训练的网络中，对抗表征与其干净表征存在显著差异，尤其是在网络末端附近，而且无论采用何种学习方案，这一现象都存在。相比之下，对抗训练显著减轻了对抗扰动的影响，使鲁棒网络对干净样本和对抗样本产生相似的表征。值得注意的是，我们观察到对比学习/对抗训练的中间层存在相似性下降，这或许可以解释其性能为何低于其他鲁棒学习方案。为验证这一解释，我们在完全对抗训练后的对比学习模型上（其鲁棒性得到显著提升）进行了类似实验，并比较了所得表征。如图6左侧所示，完全对抗训练减少了相似性下降，增强了模型的鲁棒性。为进一步深入了解这一现象，我们通过在训练期间改变对抗攻击的强度开展了消融研究，具体做法是调整

增大的扰动预算（ϵ）。如图6 和图G.4右侧所示，我们观察到

更强的对抗扰动增强了对抗表征与其对应的干净表征之间的相似性，尤其是在网络的后续层中。这一重要发现证实了（Cianfarani 等人，2022）报告的先前结果，并凸显了我们将这一观察扩展到对比学习方案中的创新贡献。此外，从图G.5的热图中，我们可以观察到，增强对抗扰动的强度会提高不同层之间的长程相似性。

与标准训练不同，无论采用何种学习方案，对抗训练都会收敛至一个通用表征集。 在这里，我们进行跨模型比较，以衡量一个模型中采用特定学习方案训练得到的所有层 X 与另一个采用不同学习方案训练的模型中的所有层 Y 之间的相似性。图7和 G.6左侧展示了采用不同标准学习方案训练的模型的结果。我们观察到，除低层外，其他层提取的表征彼此高度不同。然而，应用对抗训练后（结果见右侧），采用不同学习方案训练的网络中各层之间的相似性显著提高，表明它们倾向于提取通用表征集。这一发现与先前的一项研究（Jones 等人，2022）一致。该研究指出，无论架构如何，鲁棒网络都会收敛至通用表征集。我们的结果将这一发现拓展至对比学习方案，为更普遍的论断提供了实证证据。具体而言，我们的研究表明，无论采用何种学习方案，对抗训练都会促进模型收敛至通用表征集。

4. 结论

本研究比较了对比式对比学习、监督式对比学习与标准监督学习的鲁棒性。结果表明，在对比学习中融入标签信息有助于提升鲁棒性。对抗训练缩小了对抗表征与干净表征之间的差异，使其趋向于收敛至通用表征集。对抗表征与干净表征之间相似性的提升增强了鲁棒性，尤其是在更深层。这些发现为优化鲁棒学习方案提供了有价值的见解。

致谢

本研究部分受美国国家科学基金会（资助编号 2233873、2007202 和 2107463）、美国国家航空航天局（资助编号 80NSSC20K1720）以及 Chameleon Cloud 的支持。

参考文献

Alain, G.andBengio, Y. 使用线性分类器探针理解中间层。 *arXiv* 预印本*arXiv:1610.01644*，2016 年。

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P. 和 Joulin, A. 通过对比簇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。

Chen, T., Kornblith, S., Norouzi, M. 和 Hinton, G. 用于视觉表示对比学习的简单框架。在国际机器学习会议，第1597–1607页。PMLR，2020a。

Chen, T., Liu, S., Chang, S., Cheng, Y., Amini, L. 和 Wang, Z. 对抗鲁棒性：从自监督预训练到微调。在*IEEE/CVF*计算机视觉与模式识别会议论文集，第699–708页，2020b。

Cianfarani, C., Bhagoji, A. N., Sehwag, V., Zhao, B., Zheng, H. 和 Mittal, P. 通过表示相似性的视角理解鲁棒学习。神经信息处理系统进展，35:34912–34925，2022。

Croce, F. 和 Hein, M. 使用多样化无参数攻击集成可靠评估对抗鲁棒性。载于国际机器学习会议，第2206–2216 页。PMLR，2020。

Fan, L., Liu, S., Chen, P.-Y., Zhang, G. 和 Gan, C. 对比学习何时能在从预训练到微调的过程中保留对抗鲁棒性？神经信息处理系统进展，34:21480–21492，2021。

Gowal, S., H$_{\text{uang}}$, P.-S^{van} den Oord, A., Mann, T^{an} d Kohli, P. 面向低标签、高数据场景的自监督对抗鲁棒性。载于国际学习表征会议，2021。Grill, J.-B., Strub, F., Altch´e, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M. 等。自举自身潜变量——一种自监督学习新方法。神经信息处理系统进展，33：21271–21284，2020。

Gupta, R.、Akhtar, N.、Mian, A. 和 Shah, M. 关于对比自监督学习更高的对抗易感性。*arXiv* 预印本 *arXiv:2207.10862*，2022。

He, K., Fan, H., Wu, Y., Xie, S. 和 Girshick, R. 用于无监督视觉表征学习的动量对比。在*IEEE/CVF*计算机视觉与模式识别会议论文集，第 9729–9738 页，2020。

Hendrycks, D., Mazeika, M., Kadavath, S. 和 Song, D. 使用自监督学习可以提升模型的鲁棒性和不确定性。神经信息处理系统进展，32，2019。

Islam, A., Chen, C.-F. R., Panda, R., Karlinsky, L., Radke, R. 和 Feris, R. 关于通过对比学习获得的视觉表征可迁移性的广泛研究。在*IEEE/CVF*国际计算机视觉会议论文集，第 8845–8855 页，2021。

Jiang, Z., Chen, T., Chen, T. 和 Wang, Z. 通过对抗对比学习进行鲁棒预训练。神经信息处理系统进展，33：16199–16210，2020。

Jones, H. T., Springer, J. M., Kenyon, G. T., 和 Moore, J. S. 训练一个就等于训练了全部：跨架构相似性随鲁棒性提升而增加。载于人工智能中的不确定性，第 928–937 页。PMLR，2022。

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., 和 Krishnan, D. 监督式对比学习。神经信息处理系统进展，33：18661–18673，2020。

Kim, M., Tack, J., 和 Hwang, S. J. 对抗式自监督式对比学习。神经信息处理系统进展，33：2983–2994，2020。

Kornblith, S., Norouzi, M., Lee, H. 和 Hinton, G. 重新审视神经网络表征的相似性。发表于国际机器学习会议，第 3519–3529 页。PMLR，2019。

迈向能够抵御对抗攻击的深度学习模型。*arXiv* 预印本*arXiv:1706.06083*，2017 年。

Mitrovic, J., McWilliams, B., Walker, J., Buesing, L. 和 Blundell, C. 通过不变因果机制进行表征学习。*arXiv* 预印本*arXiv:2010.07922*，2020。Moshavash, M., Eftekhari, M. 和 Bahraman, K. 基于动量对比自监督的对抗鲁棒性训练。《计算与安全杂志》，8(1):33–43，2021。Nguyen, T., Raghu, M. 和 Kornblith, S. 宽深网络学到的内容是否相同？揭示神经网络表征如何随宽度与深度而变化。*arXiv*预印本*arXiv:2010.15327*，2020。Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M. 和 Carter, S. 深入探索：电路简介。*Distill*，2020。doi: 10.23915/distill.00024.001. https://distill.pub/2020/circuits/zoom-in。Pal, M., Jati, A., Peri, R., Hsu, C.-C., AbdAlmageed, W. 和 Narayanan, S. 使用混合对抗训练对深度说话人识别进行对抗防御。载于*ICASSP 2021-2021 IEEE*国际声学、语音与信号处理会议（*ICASSP*），第6164–6168页。IEEE，2021。Subramanian, A. torch cka。https://github.com/AntixK/PyTorch-Model-Compare，2021。Wahed, M., Tabassum, A. 和 Lourentzou, I. 通过置换簇分配进行对抗对比学习。*arXiv*预印本*arXiv:2204.10314*，2022。Zhai, R., Cai, T., He, D., Dan, C., He, K., Hopcroft, J. 和 Wang, L. 对抗鲁棒泛化只需要更多无标签数据。*arXiv*预印本*arXiv:1906.00555*，2019。Zhang, C., Zhang, K. 和 Li, Y. 从因果视角看神经网络的鲁棒性。神经信息处理系统进展，33:289–301，2020。Zhong, Y., Tang, H., Chen, J., Peng, J. 和 Wang, Y.-X. 自监督学习是否比监督学习更鲁棒？*arXiv* 预印本*arXiv:2206.05259*，2022。

附录

下面，我们提供各章节的补充信息和其他结果。

A. 方法细节

对比学习和监督式对比学习的训练过程包含两个主要阶段：

- **预训练阶段：** 在此阶段，主要目标是通过使用数据集 $\mathscr{D}p$ 最小化自监督损失函数$\mathscr{L}p(\boldsymbol{\theta}b,\ \boldsymbol{\theta}ph)$，来训练基础编码器参数$\boldsymbol{\theta}b$ 。参数$\boldsymbol{\theta}ph$ 对应于投影头，它将基础编码器的输出映射到一个较低维度的潜在空间，在该空间中应用自监督损失 $\mathscr{L}_p$。

- **监督微调阶段：** 在此阶段，主要目标是通过使用带标签数据集 $\mathscr{D}_f$ 最小化监督损失函数 $\mathscr{L}f(\boldsymbol{\theta}c)$，来训练线性分类器参数$\boldsymbol{\theta}_c$。线性分类器的作用是利用交叉熵损失 $\mathscr{L}_f$，将预训练阶段获得的表征映射到带标签空间。

在预训练阶段的标准训练中，对于对比式和监督对比式方案，$\mathscr{L}_p$ 分别由 (Chen 等人，2020a) 和 (Khosla 等人，2020) 中引入的 $\mathscr{L}_{CL}(\boldsymbol{\theta}_b,\boldsymbol{\theta}_{ph};\ \boldsymbol{x}',\boldsymbol{x}'')$ 和 $\mathscr{L}_{SCL}(\boldsymbol{\theta}_b,\boldsymbol{\theta}_{ph};\boldsymbol{x}',\ \boldsymbol{x}'',\boldsymbol{y})$ 给出。这里，$\boldsymbol{x}'$ 和 $\boldsymbol{x}''$ 是同一小批量的两个变换视图，$\boldsymbol{y}$ 标签向量在 $\mathscr{L}_{SCL}$ 中用于避免假阳性配对选择，具体说明见 (Khosla 等人，2020)。

在表征的鲁棒预训练或对抗训练场景中，我们采用一个受对抗对比学习启发的损失函数 $\mathscr{L}_p$（Jiang 等人，2020）。该损失函数表示为两个项的线性组合，定义如下：

$$\begin{aligned}\mathscr{L}_p(\boldsymbol{\theta}_b,\boldsymbol{\theta}_{ph}) &= \alpha\,\mathscr{L}_{CL}(\boldsymbol{\theta}_b,\boldsymbol{\theta}_{ph};\boldsymbol{x}',\boldsymbol{x}'')\\ &+ \beta\,\mathscr{L}_{CL}(\boldsymbol{\theta}_b,\boldsymbol{\theta}_{ph};\boldsymbol{x},\boldsymbol{x}_{adv})\end{aligned}\quad(1)$$

and

$$\begin{aligned}\mathscr{L}_p(\boldsymbol{\theta}_b,\boldsymbol{\theta}_{ph}) &= \alpha\,\mathscr{L}_{SCL}(\boldsymbol{\theta}_b,\boldsymbol{\theta}_{ph};\boldsymbol{x}',\boldsymbol{x}'',\boldsymbol{y})\\ &+ \beta\,\mathscr{L}_{SCL}(\boldsymbol{\theta}_b,\boldsymbol{\theta}_{ph};\boldsymbol{x},\boldsymbol{x}_{adv},\boldsymbol{y})\end{aligned}\quad(2)$$

分别对应对比式和监督式对比方案。在这些公式中，$\boldsymbol{x}'$ 和 $\boldsymbol{x}''$ 是 $\boldsymbol{x}$经过变换后的视图，而 $\boldsymbol{x}_{adv}$ 则是通过在每个给定的小批量 $\boldsymbol{x}$ 上迭代最大化相应损失函数而生成的PGD攻击，如下所示：

$$\boldsymbol{x}^{t+1}=\Pi_{\boldsymbol{x}+\boldsymbol{S}}(\boldsymbol{x}^t+\alpha sgn(\nabla_{\boldsymbol{x}}\mathscr{L}_{CL}(\boldsymbol{\theta}_b,\boldsymbol{\theta}_{ph};\boldsymbol{x},\boldsymbol{x}_{adv}))\quad(3)$$

and

$$\boldsymbol{x}^{t+1}=\Pi_{\boldsymbol{x}+\boldsymbol{S}}(\boldsymbol{x}^t+\alpha sgn(\nabla_{\boldsymbol{x}}\mathscr{L}_{SCL}(\boldsymbol{\theta}_b,\boldsymbol{\theta}_{ph};\boldsymbol{x},\boldsymbol{x}_{adv},\boldsymbol{y}))\quad(4)$$

分别对应对比式和监督式对比方案。这里，$\Pi_{\boldsymbol{x}+\boldsymbol{S}}$ 表示将扰动投影到允许扰动集合 $\boldsymbol{S}$ 中，而 α 是步长。在这种设置下，最小化 $\mathscr{L}_p$ 相当于同时最小化由系数 α 和 β加权的各项。本研究中，我们取 $\alpha=\beta$，以避免在变换项和对抗项之间进行不合理的优先排序。

我们考虑微调阶段鲁棒训练的两种方案。在部分对抗训练中，我们仅通过最小化损失函数来更新线性分类器参数– θ_c ，

$$\mathscr{L}_f(\boldsymbol{\theta}_c)=\alpha\,\mathscr{L}_{CE}(\boldsymbol{\theta}_c;\boldsymbol{x},\boldsymbol{y})+\beta\,\mathscr{L}_{CE}(\boldsymbol{\theta}_c;\boldsymbol{x}_{adv},\boldsymbol{y})\quad(5)$$

作为第二种替代方案，完整对抗微调采用以下损失函数，

$$\begin{aligned}\mathscr{L}_f(\boldsymbol{\theta}_b,\boldsymbol{\theta}_c) &= \alpha\,\mathscr{L}_{CE}(\boldsymbol{\theta}_b,\boldsymbol{\theta}_c;\boldsymbol{x},\boldsymbol{y})\\ &+ \beta\,\mathscr{L}_{CE}(\boldsymbol{\theta}_b,\boldsymbol{\theta}_c;\boldsymbol{x}_{adv},\boldsymbol{y})\end{aligned}\quad(6)$$

用于重新调整基础编码器参数 $\boldsymbol{\theta}_b$ 并训练线性分类器。在这些公式中，$\boldsymbol{x}_{adv}$ 是通过在给定的每个小批量 $\boldsymbol{x}$ 上迭代最大化交叉熵损失而生成的 PGD攻击。这里我们将 $\alpha=\beta$ 与预训练阶段并行进行。

我们研究上述各训练阶段的标准训练和鲁棒训练方法，以比较不同学习方案的对抗鲁棒性。表 A.1 总结了所研究的所有训练组合，包括对比学习和监督式对比学习方案中训练阶段的各种场景。

表 *A.1*。 训练场景汇总。

场景	预训练阶段	微调阶段
ST	标准训练	标准训练（固定 $\boldsymbol{\theta}_b$）
AT	对抗训练	标准训练（固定 $\boldsymbol{\theta}_b$）
部分对抗训练	对抗训练	部分对抗训练（固定 $\boldsymbol{\theta}_b$）
完整对抗训练	对抗训练	完整对抗训练

B. 实验设置

我们的实验设置与对比学习领域的代表性工作相似：SimCLR 使用了（Chen 等人，2020a），SupCon 使用了（Khosla 等人，2020）。所有场景均使用 ResNet-50 作为基础编码器，并使用双层多层感知机网络作为投影头。损失函数使用 Adam优化器进行优化，学习率为 0.0003。标准场景下，我们以 128 的小批量大小训练每个模型 200 个 epoch；对抗场景下则使用 256。在所有对抗训练场景中，除非指定了 l_∞ 范数和最大扰动上限 $\epsilon = 8/255$ 的特定值，否则使用 5 步投影梯度下降（PGD）攻击生成对抗扰动。模型在测试时使用 PGD攻击和最先进的 Auto-Attack（Croce与Hein，2020）进行评估。我们报告所有场景的测试集Top-1准确率，以评估上述场景。

B.1. 威胁模型

为评估每种场景的鲁棒性，我们考虑了两种不同的威胁模型：

• 威胁模型 I（由 $\mathscr{L}_{CE}$生成的端到端攻击）：在该威胁模型中，攻击者完全了解基础编码器和线性层分类器的架构及网络参数，但不了解预训练阶段所使用的投影头。攻击通过使用交叉熵损失以端到端方式生成。

• 威胁模型-II（针对由$\mathscr{L}_p$生成的基础编码器的攻击）：在威胁模型-II中，攻击者完全了解预训练阶段的所有组成部分，包括架构、网络参数、损失函数和训练数据集。

C. 标准训练场景下学习方案的 t-SNE 可视化

图 C.1 展示了在 CIFAR-10 数据集上使用 t-SNE 可视化的 CL、SCL、SL、SL+SCL、SL+CL 和 CL+SCL 学习方案所学到的表征。这里，我们使用标签为对应每个数据点的标记着色。ST 场景下所示的结果清楚地表明，与 CL 方案相比，SCL 和 SL 的类别边界更加清晰。此外，我们还观察到，在这两种半监督学习方案中使用一些标签信息，即 SL+CL 和 CL+SCL，可以使类别分离得更加清晰。这表明，半监督学习版本会使对手难以成功
扰动图像，从而获得更稳健的预测结果。

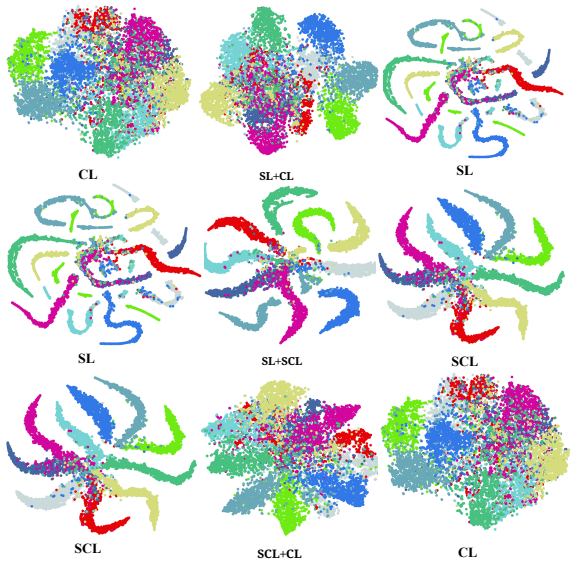


图 *C.1.***半监督学习方案（SL+CL 和 SCL+CL）比对比学习（CL）方案更能清晰地区分类别。**我们使用 t-SNE 在 CIFAR-10 数据集上可视化了不同学习方案（CL、SCL、SL、SL+SCL、SL+CL 和 SCL+CL）所学到的表征。结果表明，在 ST 场景下，SCL 和 SL 的类别边界都比 CL 更清晰。此外，在半监督学习方案（SL+CL 和 SCL+CL）中融入标签信息能增强类别分离效果，表明其抵御对抗扰动的鲁棒性有所提升。

D. 威胁模型-II 下通过标准训练获得的鲁棒性

威胁模型-II 不适用于监督式学习方案，因为基础编码器和线性分类器是以端到端方式联合训练的。表 D.1 报告了在 CIFAR-10 和 CIFAR-100 数据集上针对不同 40 步 PGD攻击的结果。从结果中可以观察到，使用 $\mathscr{L}_{SCL}$ 训练的模型比 $\mathscr{L}_{CL}$ 更稳健；在后者中，攻击是针对基础编码器生成的。这表明，自监督对比学习中的假负对选择会降低模型的稳健性，这与（Gupta 等人，2022）报告的结果一致。

E. 不同对抗训练场景在多种 AutoAttack 下的性能

在此，我们评估了不同场景在 CIFAR100 数据集上抵御不同 Auto-Attack 的鲁棒性。结果如图 E.1所示。对比分析表明，完全对抗训练能够有效提升基于对比学习的网络的鲁棒性。这种提升

表 *D.1*。 **在针对基础编码器生成的对抗攻击下，SCL 比 CL 方案更鲁棒**。 在 ST 场景中，根据威胁模型-II（针对基础编码器生成的攻击）评估对比学习方案（包括 CL 和 SCL）的性能。最佳性能以粗体标出。威胁模型-II 不适用于 SL 方案，因为基础编码器和线性分类器是以端到端方式联合训练的。

模型	数据集	标准训练	投影梯度下降 (4/255)	投影梯度下降 (4/255)	投影梯度下降(16/255)
SCL	CIFAR-10	93.56	**30.3**	**12.9**	**10.06**
CL		84.27	19.58	9.61	7.19
SCL	CIFAR-100	73.38	**7.87**	**3.28**	**2.16**
CL		60.28	5.52	0.94	0.23

是通过整合标签信息实现的，从而增强了抵御对抗攻击的鲁棒性。此外，在不同的对抗训练场景下，SCL 的性能略有差异。

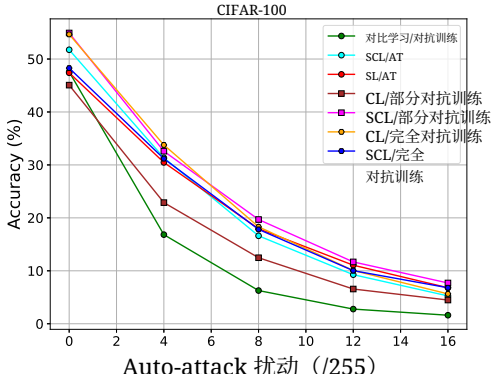


图 *E.1*。 **完全对抗训练有效地将标签信息融入通过对比学习训练的鲁棒网络，显著提升了整体鲁棒性**。 对 CIFAR-100 数据集上不同学习方案针对各种 Auto-攻击的测试准确率进行比较分析，结果表明，通过整合标签信息，完全对抗训练能够有效提升基于对比学习的网络的鲁棒性，从而增强其抵御对抗攻击的能力。此外，在不同对抗训练场景下，SCL 的性能略有差异。

F. 威胁模型-II 下通过对抗训练实现鲁棒性（跨任务鲁棒性不可迁移）

在对比学习中，任何利用基础编码器学习到的表征的目标都称为下游任务。在本实验中，我们比较了在对比式和监督式对比学习方案中仅对基础编码器进行对抗训练时模型的鲁棒性。表F.1 展示了模型在 CIFAR-10 和 CIFAR-100 数据集上抵御不同 40 步 PGD攻击的性能。与威胁模型 I 相比，我们的研究结果在威胁模型-II 表明，仅对基础编码器进行对抗训练所得到的鲁棒模型无法迁移到下游任务中。这种鲁棒性无法在不同任务间迁移的现象称为跨任务

鲁棒性和可迁移性方面的挑战也已在（Fan 等人，2021）中报告。

G. 不同学习方案的表征结构

本节提供正文第3.2.2 节相关的所有图表。

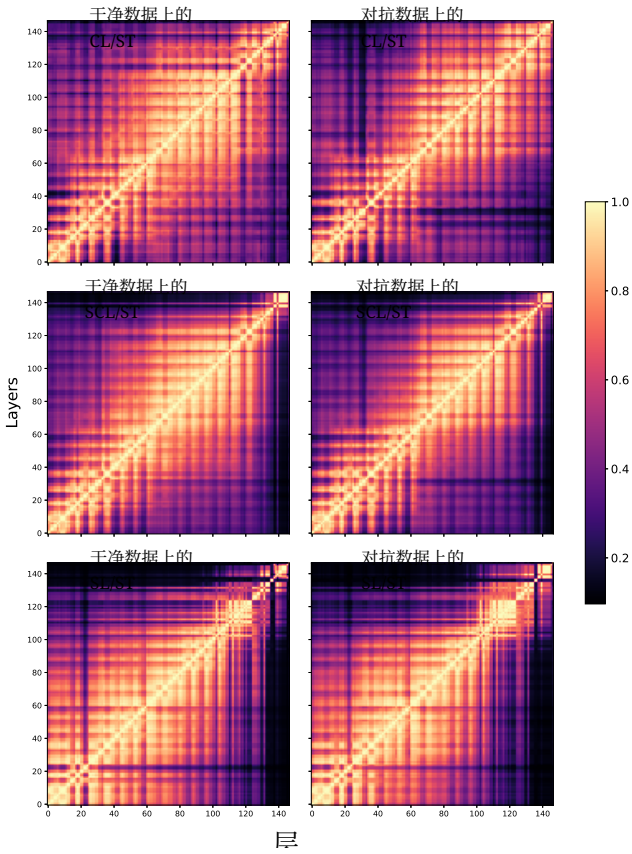


图 *G.1*。**无论采用何种学习算法，标准训练网络获取的表征在对抗样本与干净样本之间都存在显著差异**。 我们计算了使用不同学习方案训练的标准训练网络中所有层组合之间的表征相似度，同时考察干净数据和对抗数据。三种学习方案（SCL、SL 和 CL）的内部表征结构（第一列）存在明显差异。与 SCL 和 SL 相比，CL 在整个网络中展现出更一致的表征。此外，标准训练网络中的干净表征与对抗表征存在显著差异（第一列与第二列相比）。

表 *F.1.* **仅对基础编码器进行对抗训练所获得的鲁棒性，无法有效迁移到下游任务。** 本文以 CIFAR-10 和 CIFAR-100 数据集上的 Top-1 准确率为指标，将对抗训练场景下的对比学习方案（包括对比学习和 SCL）与基线对抗训练监督学习方案进行比较。模型在威胁模型 I（通过交叉熵损失生成的端到端攻击）和威胁模型-II（针对基础编码器生成的攻击）下进行评估。显然，对抗训练能够提升威胁模型-II 下对比学习和 SCL 的鲁棒性。然而，结果表明，这些鲁棒模型在遭受端到端攻击或处于威胁模型 I 下时会失去鲁棒性。

模型	数据集	标准训练	对抗训练	由交叉熵损失生成的端到端攻击			针对基础编码器生成的攻击		
		干净样本	干净样本	PGD（4/255）	PGD（8/255）	投影梯度下降 (16/255)	投影梯度下降 (4/255)	投影梯度下降 (8/255)	投影梯度下降 (16/255)
SL	CIFAR-10	91.73	76.33	53.2	32.52	10.5	NA	NA	NA
SCL		93.56	80.22	11.99	55.83	31.45	**59.36**	**62.25**	**59.38**
CL		84.27	70.13	36.86	11.91	0.26	**69.1**	**65.88**	**49.8**
SL	CIFAR-100	69.56	47.4	26.68	14.8	3.8	NA	NA	NA
SCL		73.38	51.8	31.7	17.11	3.65	**37.09**	**30.02**	**24.5**
CL		60.28	47.58	23.58	7.83	0.43	**46.18**	**42.44**	**26.9**

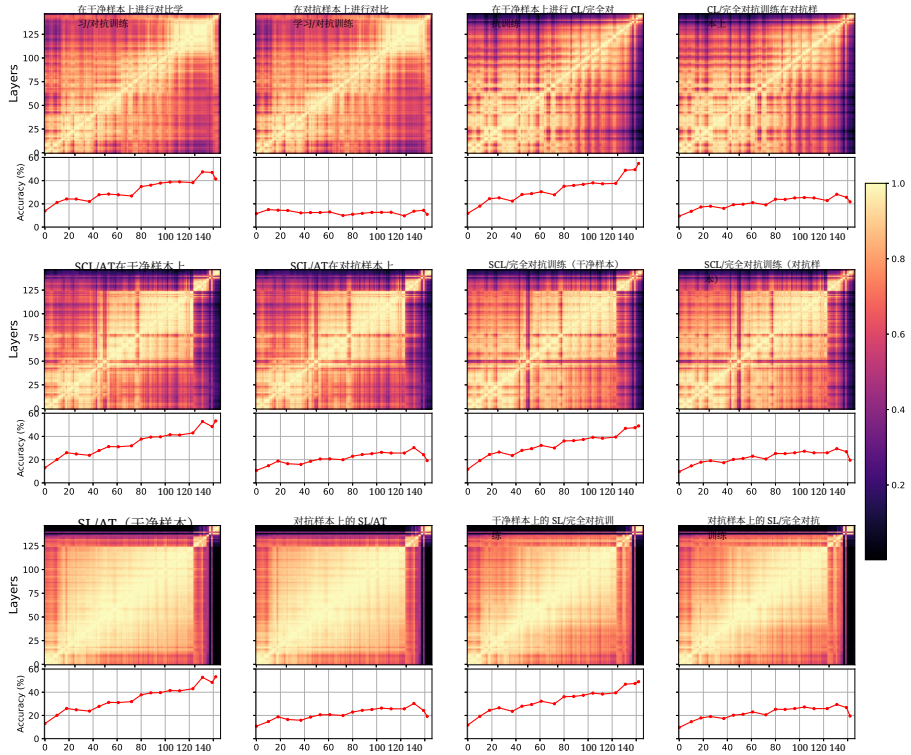
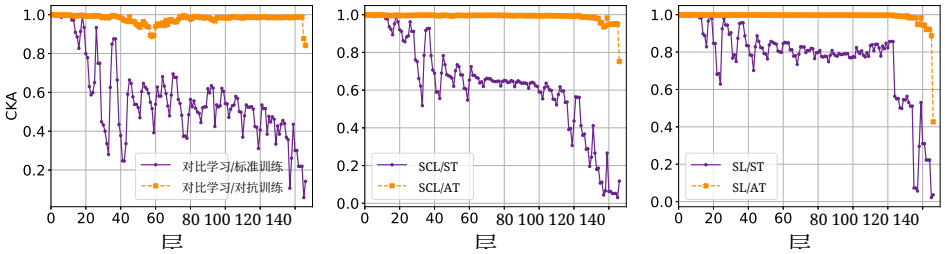
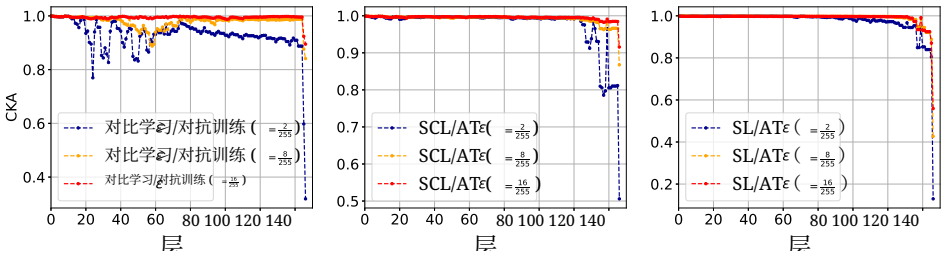


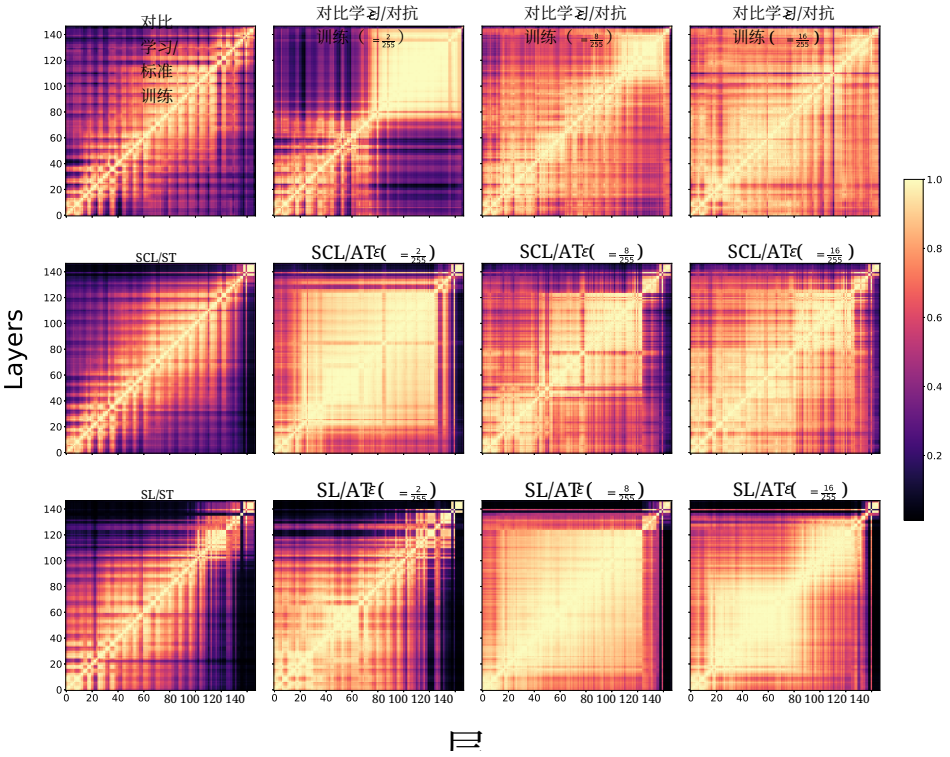
图 *G.2*。 **无论采用何种学习方案，对抗训练网络中对抗表征与干净表征之间都具有较高的相似性。** 我们通过比较不同学习方案中的层对，并计算它们在干净样本和对抗样本上的 CKA 相似度，分析鲁棒模型。我们采用线性探测来深入了解网络动态以及中间层的作用。结果表明，与标准训练相比，跨层相似性得到增强，这一点可由图中更高的亮度体现出来。此外，通过对抗训练得到的网络中，对抗表征与干净表征之间也表现出显著的相似性。进一步比较 AT 与其对应的完全对抗训练（Full AT）所获得的表征后，我们观察到 CL 中的长程相似度显著增强。这种相似性的提升使标准准确率和对抗准确率都得到大幅提高。相比之下，在 AT 和完全对抗训练（Full AT）场景下，SCL 和 SL 学到的表征差异较小，因此其性能仅出现轻微变化。



图*G.3.* **对比对抗表征及其对应的干净表征。** 比较模型不同层中的干净表征和对抗表征后发现，标准训练网络中的两者存在显著差异。对抗训练缩小了这种差异，使鲁棒网络中的干净样本和对抗样本具有相似的表征。然而，CL/AT 在中间层的相似性有所下降，这或许可以解释其性能为何低于其他鲁棒学习方案。



图*G.4.* **提高对抗表征与干净表征之间的相似性能够增强鲁棒性，尤其是在网络末端附近。** 提高对抗表征与干净表征之间的相似性能够增强鲁棒性。在对抗训练期间，我们增大扰动预算，以改变对抗攻击的强度。这使对抗表征与干净表征之间的相似性提高，尤其是在网络末端附近。



图*G.5.* **无论采用何种具体学习方案，对抗训练都会促使层间长程相似性的形成。** 我们在训练期间改变对抗攻击的强度。我们观察到，增大对抗扰动的强度会使层间长程相似性稳定出现，与所采用的学习方案无关。

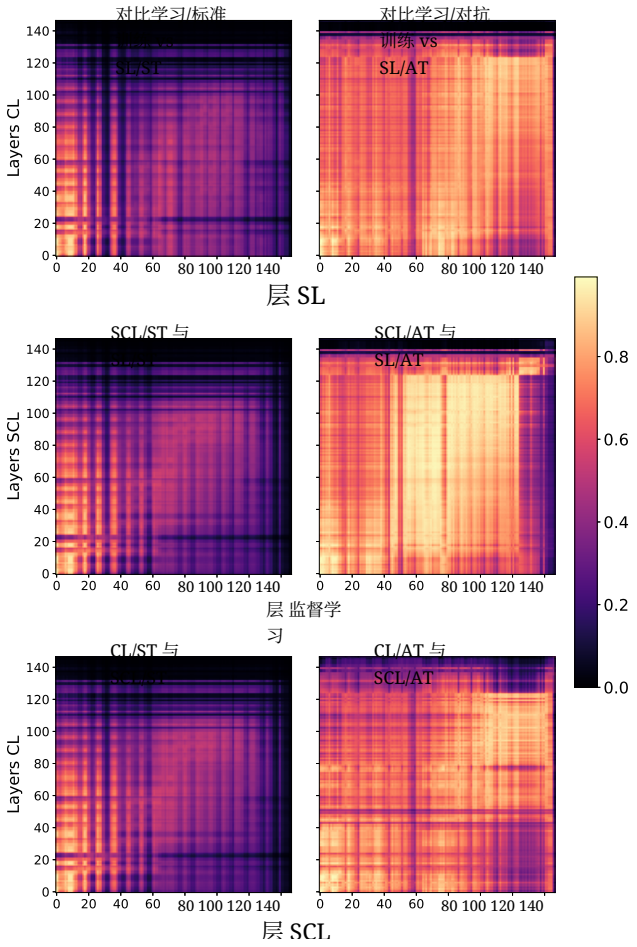


图 *G.6*。**与标准训练网络不同，通过对抗训练得到的网络在不同学习方案下展现出显著的对抗表征相似性**。采用不同学习方案训练的标准训练网络之间的跨模型 CKA 热图表明，这些方案提取的对抗表征各不相同，尤其是在网络中较多的较高层中。跨模型比较显示，应用对抗训练后，不同学习方案所对应层之间的相似性有所提高，这表明模型正转向提取一套通用表征。