CLAP: Isolating Content from Style through Contrastive Learning with Augmented Prompts

Yichao Cai[1], Yuhang Liu[1], Zhen Zhang[2], and Javen Qinfeng Shi[3]

Australian Institute for Machine Learning, University of Adelaide, SA 5000, Australia

{yichao.cai, yuhang.liu01, zhen.zhang02, javen.shi}@adelaide.edu.au

摘要

对比视觉语言模型，例如 CLIP，在各种 dowmsteam 任务中引起了相当大的关注，这主要是由于所学习的特征具有卓越的泛化能力。然而，他们学到的特征经常混合内容和风格信息，这在一定程度上限制了他们在分布变化下的泛化能力。为了解决这一限制，我们采用多模态数据的因果生成视角，并提出通过数据增强进行对比学习，以将内容特征与原始表示分离。为了实现这一目标，我们首先探索图像增强技术，并开发一种方法将它们无缝集成到预训练的类似 CLIP 的模型中，以提取纯内容特征。更进一步，认识到文本数据固有的语义丰富性和逻辑结构，我们探索使用文本增强将潜在内容与风格特征分离。这使得类 CLIP 模型的编码器能够专注于潜在内容信息，通过预先训练的类 CLIP 模型来完善学习的表示。我们在不同数据集上进行的广泛实验证明了零样本和少样本分类任务的显着改进，同时增强了对各种扰动的鲁棒性。这些结果强调了我们提出的方法在完善视觉语言表示和推进多模态学习的最新技术方面的有效性。

关键词： 数据增强；潜在变量；解纠缠

1 引言

以 CLIP [28] 为代表的视觉语言模型因其卓越的泛化能力而受到广泛关注，这种泛化能力是通过利用跨模态对比损失获得的学习特征来实现的 [15,20,28]。然而，尽管在广泛的数据集上进行了预先训练，类似 CLIP 的模型在解开潜在内容信息和潜在风格信息方面仍存在不足。因此，它们不能免受虚假相关性的影响，即，与风格相关的信息被错误地用于预测与任务相关的标签。当存在分布变化或对抗性攻击时，这些限制变得明显，其中虚假相关性经常在不同环境中发生变化。例如：(1) 零样本能力明显依赖于特定的输入文本提示 [16,37,38]；(2)

在少样本学习场景中观察到少样本场景的性能下降 [9, 28]；(3) 对对抗性攻击的敏感性已被探索 [25, 33, 35]。

从因果角度来看，这项工作从一种简单而有效的方法开始，即图像增强，以在类 CLIP 模型的学习表示中理清内容和风格信息。这种方法的灵感来自于因果表示学习理论发展的最新进展 [32]，这表明增强图像可以被解释为对潜在风格变量进行软干预的结果，如图 1a 所示。这种增强会产生自然的数据对，其中内容信息保持不变，而风格信息发生变化。因此，使用对比学习，将不变的内容信息与变化的风格信息隔离开来变得可行。受这一理论进步的推动，我们提出了一种实用方法，将图像增强合并到类 CLIP 模型中，以从原始学习特征中提取内容信息。具体来说，解缠结网络旨在通过使用对比损失和图像增强来微调预训练的 CLIP 模型。

尽管在通过图像增强从类 CLIP 模型学习的原始特征中分离内容和风格信息方面取得了进展，但我们认识到一个固有的局限性：设计足够的图像增强以确保图像中的所有风格因素发生变化通常具有挑战性。从理论上讲，解开内容和风格信息需要改变所有风格因素 [32]。然而，由于图像数据中风格信息的高维性和复杂性，通过图像增强引起潜在风格的充分变化提出了挑战。通过人工设计的图像增强技术实现显着的风格变化（例如将狗的照片转换为草图，同时保留内容但显着改变风格）是非常困难的。

更进一步，我们不依赖图像增强，而是探索使用文本增强来理清潜在内容和风格因素。这种转变是由两个关键观察推动的：1）视觉和语言数据共享相同的潜在空间。因此，文本增强也可以用来诱导潜在风格因素的变化，而不是图像增强。2）文本数据本质上具有高度的语义和逻辑结构，使其比图像数据更适合属性操作。因此，通过文本增强实现足够的样式更改比图像增强更可行，有助于将内容与样式信息隔离，参见图 1c 进行视觉比较。例如，将文本从“狗的照片”转换为“狗的草图”在语言模态中是简单的，而在图像数据中实现类似的转换则具有挑战性。受这些观察的启发，我们认为通过文本增强引入风格变化（如图 1b 所示）为学习视觉语言内容特征提供了一种比依赖图像增强更有效的方法。

总之，这篇论文的贡献包括：(1) 旨在解开潜在内容和风格因素，以细化预训练的 CLIP-like 模型的视觉语言特征，我们提出通过数据增强进行对比学习，以微调预训练的 CLIP 的原始特征。就像从因果角度来看的模型一样。(2) 我们提出了一种为预训练的类 CLIP 模型定制的新方法。该方法利用解纠缠网络，该网络使用图像增强的对比学习进行训练，从类 CLIP 模型的图像编码器提供的学习特征中提取潜在内容特征。(3) 我们提出了增强提示对比学习 (CLAP)，从类 CLIP 模型的表示中提取潜在内容特征。首先使用类 CLIP 模型和文本增强的预训练文本编码器训练解缠结网络。随后，训练好的解缠结网络被转移到类 CLIP 模型的图像编码器中。(4) 在大型真实数据集上进行的实验证明了所提出的图像增强和文本增强在零镜头和少镜头性能方面的有效性，以及对扰动的鲁棒性。

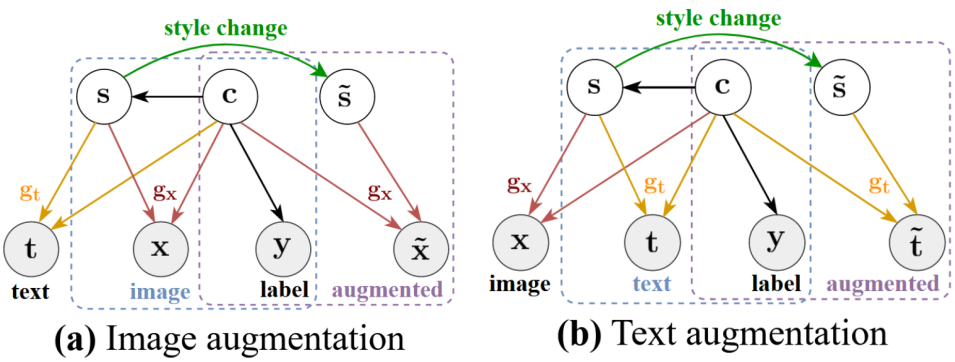


图 1. 因果生成模型

2 相关工作

2.1 对比视觉语言模型

使用跨模态对比损失，CLIP [28] 革命性地引入了可扩展的对比视觉语言模型。通过利用大量来自互联网的图像文本对，展示了前所未有的零样本学习能力和跨数据集的卓越泛化能力，并支持众多下游任务 [15]。ALIGN [7] 扩大了对比视觉-语言建模的规模，训练了多达十亿个图像-文本对，同时集成了视觉转换器的自注意力机制 [7]，进一步提高了性能。尽管取得了成功，但类 CLIP 模型对输入文本提示表现出敏感性 [16, 38]，导致不同提示的性能存在差异。通过提示学习和工程来减轻这种提示敏感性的努力 [6, 10, 16, 37, 38] 侧重于为特定任务定制提示，但并没有从根本上增强 CLIP 的表示。此外，类似 CLIP 的模型容易受到对抗性攻击 [2, 8]，当前的策略 [25, 35] 涉及对抗性自然图像对以提高弹性。我们的工作与特定于任务的方法不同，旨在从解开的角度增强 CLIP 的表示，解决类 CLIP 模型中固有的上述问题。

2.2 解纠缠表征学习

旨在将数据中的内在潜在因素分离成不同的、可控的表示，解缠结的表示学习有利于各种应用 [19,30,34]。具体来说，在分类任务中，研究表明，可以通过更有效地解开不变内容变量来增强模型的性能和针对数据分布扰动的鲁棒性，而无需完全识别所有内在潜在变量 [17, 21–23]。在单一模式中，[39] 等研究表明对比学习 [4, 12, 13] 可以潜在地逆转数据生成过程，有助于分离表示。此外，[32] 表明图像增强可以通过显着的风格变化帮助将内容变量从潜在空间中分离出来。[14] 采用混合技术进行数据增强，实现更丰富的跨模式匹配。与这些方法不同，我们的方法侧重于采用文本增强来解开潜在内容变量，引入一种独特的方法来学习精致的视觉语言表示。

3 本文方法

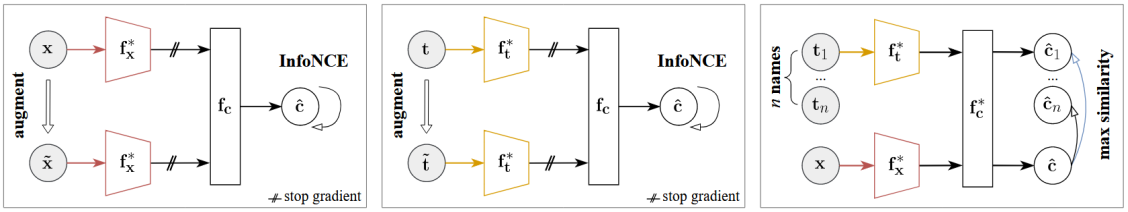
3.1 本文方法概述

如图 1 的因果生成模型所示，在所提出的模型当中，图片所提取出来的特征一般被区分为两个特征：一个对应于潜在内容特征 $\mathbf{c}$，另一个对应于潜在风格变量 $\mathbf{s}$，而在类 CLIP 模型

中，潜在变量空间被划分为两个子空间，一个是图片对应的特征 $\mathbf{x}$，另一个为文本对应的特征 $\mathbf{t}$。我们的因果生成模型可以由以下公式表示出来：

$$s := g_s(c), x := g_x(c, s), t := g_t(c, s), y := g_y(c). \tag{1}$$

该公式表明，风格特征受到内容影响，而图片同时受到内容和风格影响，文本页同时受到内容和风格影响，而标签仅仅与内容相关。[32] 表明通过要求所有潜在风格变量发生变化潜在内容变量可以从图片中识别出来。这种变化可以通过图像增强来实现，即 $\widetilde{x}$ 是由 $\widetilde{s}$ 以及 $\mathbf{c}$ 组成的，它是通过对原始潜在风格变量 s 进行软干预而产生的。但是图像增强无法保证完全对图片所有的潜在风格变量进行改变，因此该方法存在着不小的局限性，因此该论文又提出了另一种方法来对风格变量进行干预，通过文本增强对所有潜在风格变量进行足够的改变比图像增强更可行，由于文本数据中的高语义和逻辑结构，该论文深度研究了使用文本增强将内容信息与风格信息分离。



(a) Image augmentation **(b)** Text augmentation **(c)** Zero-shot inference

图 2. 通过数据增强完善 CLIP。(a) 训练涉及一个解缠结的网络 f_c，利用原始图像对 x 和 $\tilde{x}$ 上的对比损失，CLIP 的图像编码器 f_x^* 保持冻结梯度。(b) 使用 CLIP 的固定文本编码器 f_t^*，通过增强文本提示 t 和 t 的对比学习，实现更有效的内容特征学习。(c) 推理阶段：经过训练的解缠网络 f_c^* 与 CLIP 的文本和图像编码器 f_t^* 和 f_x^* 集成，以实现对输入图像 x 和类别的零样本推理。

3.2 解纠缠网络架构

解缠结网络采用多层感知器（MLP）架构。为了充分受益于预训练的 CLIP 文本编码器，我们构建了一个具有零初始化投影的残差 MLP，充当解缠结网络，如图 3 所示。这种设计可以直接从预训练的表示空间中学习，避免了随机启动点，受到 ControlNet 的零转换操作 [36] 的启发，我们在残差 MLP 中适应零线性操作。在此架构中，主分支包括一个零初始化、无偏差线性层，位于 SiLU 激活和正常初始化线性层组合之后。传统上，初始线性层之前、位于第一和第二线性层之间以及第二线性层之后的特征的维度分别被命名为输入 dim、潜在 d_{mid} 和输出 d_{out} 维度。为了纠正输入和输出维度之间的任何不匹配，网络在快捷路径内采用最近邻下采样，从而确保输入特征的对齐和清晰度的保留。在推理阶段，引入权重参数 > 0，以在与输入特征集成之前调制从主分支发出的特征部分，而该参数在整个训练阶段保持恒定为 1。

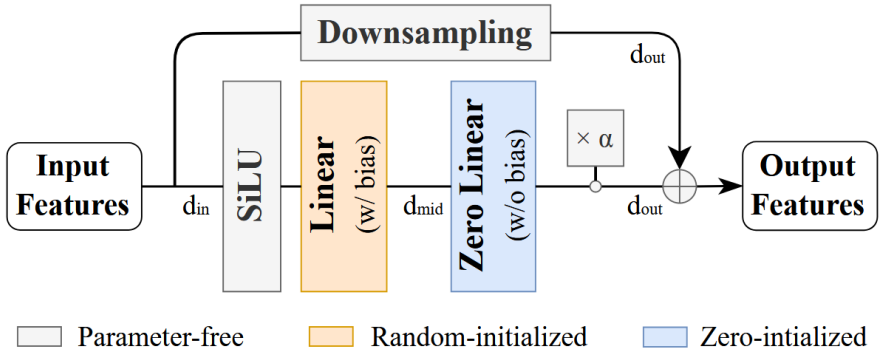


图 3. 解缠结网络的结构。该架构包含一个残差块，该残差块具有零初始化、无偏差线性层，可从输入特征空间开始优化。当输入和输出维度不同时，利用下采样操作来实现对齐。在推理过程中，标量参数 在组合之前平衡主分支和输入特征。

3.3 通过图片增强将内容与风格分离

虽然最近的研究 [32] 通过对比学习和数据增强来进行内容风格解耦提供了保证，但是对于如何应用到视觉语言模型中还没有较好的思路，理论研究建议使用 InfoNCE 损失函数 [26] 来提取内容信息，如下所述：

$$\mathcal{L}\left(\mathbf{f};\{\mathbf{x}_i,\tilde{\mathbf{x}}_i\}_{i=1}^{b},\tau\right)=-\frac{1}{b}\sum_{i=1}^{b}\log\frac{\exp\left[\langle\mathbf{f}\left(\mathbf{x}_i\right),\mathbf{f}\left(\tilde{\mathbf{x}}_i\right)\rangle/\tau\right]}{\sum_{j=1}^{b}\exp\left[\langle\mathbf{f}\left(\mathbf{x}_i\right),\mathbf{f}\left(\tilde{\mathbf{x}}_j\right)\rangle/\tau\right]}, \tag{2}$$

其中，b 为 batch 大小，x_i 为第 i 张图片，$\tilde{x_i}$ 为图片增强后的第 i 张图片，$f(x_i)$ 为第 i 张图片的特征。

我们将其扩展以利用增强图像（以下简称 “Im.Aug”）的对比学习来完善预先训练的视觉语言模型。如图 2a 所示，我们在 CLIP 的预训练图像编码器之上训练一个解缠结网络。为了提高训练效率和所提出方法的可用性，我们冻结了预训练的图像编码器。基于 InfoNCE 损失，Im.Aug 的学习目标制定如下：

$$\mathbf{f}_{\mathbf{c}}^{*}=\underset{\mathbf{f}_{\mathbf{c}}}{\operatorname{argmin}}\underset{\{\mathbf{x}_i\}_{i=1}^{b}\in\mathcal{D}_{\mathbf{x}}}{\mathbb{E}}\mathcal{L}\left(\mathbf{f}_{\mathbf{c}}\circ\mathbf{f}_{\mathbf{x}}^{*};\{\mathbf{x}_i,\tilde{\mathbf{x}}_i\}_{i=1}^{b},\tau\right), \tag{3}$$

其中，D_x 为图像数据集，f_c 为我们要训练的解纠缠网络，f_x^* 为冻结参数的图像编码器，变量 x_i 指的是从 D_x 采样的图像，$\tilde{x_i}$ 是其增强视图。

对于图像增强过程，使用了对比学习实践中常用的技术 [4,5,32]，特别是随机裁剪和颜色失真。同时我们设计了对图像增强的提示词，然后使用稳定的扩散模型 [31]，生成包含训练数据 D_x 的合成图像，稳定扩散的模板提示的创建基于对象大小、颜色、图像类型和艺术风格等属性。属性包括对象的 10 种颜色和 3 种尺寸，图像的 8 种类型和 2 种艺术风格。通过将这些属性组装成 “[对象大小][对象颜色][类] 的 [艺术风格][图像类型]” 等提示，我们为每个类生成 480 个唯一文本，每个提示从这些文本中合成一个图像。

Object Color	Object Size	Image Type	Art Style
yellow, green, black, blue, multicolored, orange, red, white, brown, purple	large, small, normal sized	painting, cartoon, infograph, sketch, photograph, clipart, mosaic art, sculpture	realistic, impressionistic

图 4. 基于模板的提示.

3.4 通过文本增强将内容与风格分离

尽管通过图像增强在解开内容和风格方面取得了进展，但由于图像中风格信息的高维性和复杂性，充分改变图像中的所有风格因素仍然具有挑战性。现有的图像增强技术很难通过增强来实现实质性的风格改变，这对于完全解开 [32] 至关重要。相反，文本数据本质上具有高度的语义和逻辑结构，使其比图像数据更适合属性操作。为了进一步探索内容的解开，提出了增强提示对比学习（CLAP）。如图 2b 所示，CLAP 采用 InfoNCE 损失在 CLIP 的预训练文本编码器上训练解纠缠网络，保持编码器的梯度固定，类似于 Im.Aug。利用更简单的文本结构，以前用于合成图像的基于模板的提示现在用作训练文本数据集，用 D_t 表示。利用与 Im.Aug 中相同的解缠结网络，CLAP 的学习目标概述如下：

$$f_c^* = \arg\min_{f_c} \mathbb{E}_{\{t_i\}_{i=1}^b \in \mathcal{D}_t} \left[\mathcal{L}(f_c \circ f_t^*; \{t_i, \tilde{t}_i\}_{i=1}^b, \tau) + \lambda \mathcal{L}(f_c \circ f_t^*; \{t_i^c, \tilde{t}_i\}_{i=1}^b, 1) \right], \tag{4}$$

其中，D_t 为文本数据集，f_c 为我们要训练的解纠缠网络，f_t^* 为冻结 t 参数的文本编码器，t_i 为从文本数据集中提取出来的第 i 个文本，$\tilde{t}_i$ 为进行文本增强后的第 i 个文本。t_i^c 为该文本句子的对应类别名称。

为了确保文本提示在不影响其内容的情况下进行风格变化，论文开发了用于合成文本提示的特定增强技术。我们采用了 EDA 中的随机删除 (RD) 和随机交换 (RS) 技术，并对其进行自定义以适合我们的提示结构。为了避免通过引入新的对象名称或更改文本提示的核心思想而无意中更改内容，我们的增强方法不包括随机单词插入或替换。我们的主要增强技术是对象大小删除（OSD）、对象颜色删除（OCD）、图像类型删除（ITD）、艺术风格删除（ASD）和交换提示顺序（SPO），每种技术都有一定的概率应用。为了丰富训练样本群体，我们使用了额外的增强，称为 IGN（插入高斯噪声）。遵循提示学习方法 [47, 48] 的初始化协议，我们将标准差为 0.02、噪声长度等于 4 的零均值高斯噪声插入到标记化提示中。

Original	OSD	OCD	ITD	ASD	SPO
a realistic painting of a large red car	a realistic painting of a red car	a realistic painting of a large car	a realistic of a large red car	a painting of a large red car	a large red car in a realistic painting

图 5. 即时增强技术。使用特定增强技术从原始文本提示生成各种增强视图：OSD（对象大小删除）、OCD（对象颜色删除）、ITD（图像类型删除）、ASD（艺术风格删除）和 SPO（交换提示顺序）.

3.5 推理

在训练完成后，解纠缠网络 f_c^* 可以与 CLIP 的图像和文本编码器无缝集成，增强零样本能力, 图像与文本经过 CLIP 对应的图像编码器和文本编码器后，再经过解纠缠网络提取出来的特征进行相似度比对，找对最符合的那个类别。这种集成保留了 CLIP 的零样本功能，同时通过改进内容的分离来实现精细的功能。

4 复现细节

4.1 与已有开源代码对比

这个项目的代码是基于原论文的代码 https://github.com/YichaoCai1/CLAP 上修改的。使用了其生成不同文本的代码，自己复现了其解纠缠网络架构、调用了 CLIP 模型以及其训练过程，同时对不同的数据集进行了不同的配置文件编写，以及调用了不同的对抗攻击方法对图片进行加噪检验其模型效果。其中，主要使用了四个数据集进行训练以及零样本、少样本推理。四个数据集分别为：PACS（7 个类别）[18]、VLCS（5 个类别）[1]、OfficeHome（65 个类）[29]、DomainNet（345 个类）[27]。

4.2 实验环境搭建

本实验的训练以及推理过程均是在 2 张 P100 上进行的，torch 版本为 2.5.1，torchvision 版本为 0.20.1, numpy 版本为 2.1.3, cuda 版本为 12.4, Im.Aug 和 CLAP 使用 ViT-B/16 CLIP 模型实现。为了简洁起见，我们呈现每个数据集跨域的平均结果。

5 实验结果分析

使用训练完成的模型，在各个数据集上进行了零样本，少样本（1-shots、4-shots、8-shots、16-shots、32-shots）学习，同时使用了 FGSM、PGD-20、CW-20 三种对抗攻击方法，在其基

础上检验模型对图片的检测正确率。除此之外，在 PACS 数据集上使用了不同的方法，对样本特征进行了 t-sne 可视化，来更加直观化的展示出模型的效果。

Prompt	Method	PACS	VLCS	Off.Home	Dom.Net	Overall
ZS(C)	CLIP	91.6	71.6	75.8	54.3	73.3
	Im.Aug	92.7	75.9	78.0	53.2	74.9
	CLAP	**93.5**	**79.3**	**80.4**	**56.8**	**77.5**
ZS(CP)	CLIP	91.4	77.2	75.5	54.5	74.6
	Im.Aug	92.6	78.7	78.2	53.0	75.6
	CLAP	**93.3**	**80.1**	**80.6**	**56.9**	**77.7**
ZS(PC)	CLIP	91.8	77.4	76.4	54.7	75.0
	Im.Aug	92.7	78.9	78.8	53.9	76.0
	CLAP	**93.8**	**80.4**	**81.2**	**57.2**	**78.1**
ZS(NC)	CLIP	88.2	65.5	69.4	48.7	67.9
	Im.Aug	92.3	76.2	70.3	46.2	71.2
	CLAP	**93.4**	**79.9**	**74.5**	**50.4**	**74.5**

表 1. 三个不同提示的零样本结果：“C” 代表 “[class]”, “CP” 代表 “a photo of a [class]”, “PC” 代表 “a [class] in a photo ”，以及动态提示 “NC” 表示 “[noise][class]” 表明 CLAP 在所有数据集上始终优于 CLIP , Im.Aug 的零样本性能，

零样本性能为了评估零样本能力，CLAP 使用三个特定的固定提示进行评估：ZS(C)，仅使用 “[class]” 中的类名称; ZS(PC)，格式为 “a photo of a [class]”; ZS(CP)，结构为 “a [class] in a photo”。为了彻底检查零样本熟悉程度，还使用了动态提示 ZS(NC)，格式为 “[noise][class]”，其中 “[class]” 表示引入均值为 0 ，标准差为 0.02 的高斯噪声。

如表所示，CLAP 在每个数据集的所有评估提示中均超过了 CLIP 和 Im.Aug。与 CLAP 相对于 CLIP 实现的零样本性能统一增强不同，Im.Aug 显示了不一致的结果。仔细检查发现 CLAP 相对于 CLIP 的优越性对于动态 ZS(NC) 提示尤其重要。这证明了与原始 CLIP 表示相比，CLAP 在显着提高零样本性能方面的有效性。

我们对四个数据集中每个域的 1 次、4 次、8 次和 16 次线性探针进行评估。如图 4 所示，CLAP 在小样本学习场景中显着优于 CLIP 和 Im.Aug。值得注意的是，在 1-shot 设置中，CLAP 在 PACS、VLCS、OfficeHome 和 DomainNet 数据集上比线性探针 CLIP 模型实现了 +1.9%、+7.7%、+4.6% 和 +2.5% 的性能改进，分别。与 Im.Aug 观察到的改进相比，这些改进尤其显着，展示了 CLAP 在少数场景中的功效。

为了评估对抗性鲁棒性，通过生成对抗性样本，针对著名的对抗性攻击方法（例如 FGSM [11]、PGD [24] 和 CW [3]）评估零样本（ZS(C)）和单样本分类器。

对于 FGSM，使用 1 次对抗性迭代，而对于 PGD 和 CW，使用 20 次迭代，所有迭代的 epsilon 均为 0.031。如表所示。如图 5 所示，与基于 CLIP 表示的分类器相比，使用 CLAP 表示的分类器表现出对这些对抗性攻击的卓越弹性。在四个数据集上，CLAP 的零样本分类器相对于 FGSM 超过 CLIP +7.9% ，相对于 PGD-20 超过 +1.3%，相对于 CW-20 超过 +1.2% 。这些数字明显超过了 Im.Aug 所实现的相对于 FGSM 的 +4.8%、相对于 PGD-20 的 +0.8%

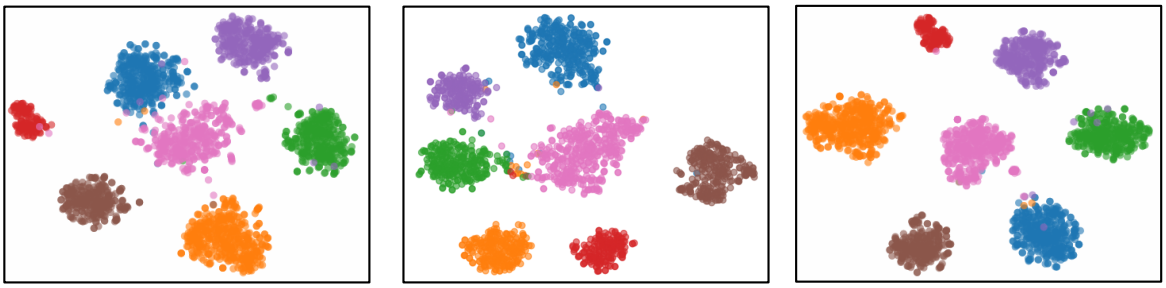
以及相对于 CW-20 的 0.3% 的性能改进。结果表明，CLAP 有效增强了零样本场景中针对对抗性攻击的鲁棒性。

Avg. top-1 acc.(%) under aversarial attacks

Adversarial attack	Method	PACS	VLCS	Off.Home	Dom.Net	Overall
FGSM	CLIP	81.3	67.5	54.3	20.1	55.8
	Im.Aug	82.4	69.2	55.2	35.6	60.6
	CLAP	**84.2**	**72.1**	**56.9**	**41.8**	**63.7**
PGD-20	CLIP	24.3	1.7	8.1	9.5	10.9
	Im.Aug	**26.8**	2.0	8.9	9.0	11.7
	CLAP	25.4	**2.9**	**10.2**	**10.3**	**12.2**
CW-20	CLIP	24.9	1.2	6.3	7.1	9.9
	Im.Aug	26.8	1.5	6.1	6.3	10.2
	CLAP	**27.5**	**2.0**	**7.2**	**7.9**	**11.1**

表 2. 图像增强和 CLAP 都增强了 CLIP 的带有对抗性攻击的零样本识别的鲁棒性，其中 CLAP 表现出更大的改进。

在我们的 t-SNE 可视化中，我们检查了 PACS 数据集的艺术绘画领域内所有图像的 CLIP、Im.Aug 和 CLAP 的表示。图 5 显示 CLAP 的图像表示比 CLIP 以及 Im.Aug 的图像表示显示出明显的类间分离和更紧密的类内聚类。这一观察结果表明，与其他两者相比，CLAP 的表示与内容信息的联系更紧密，受风格信息的影响较小。



(a) CLIP (b) Im.Aug (c) CLAP

图 6. PACS 艺术绘画数据集中所有图像的 t-SNE 可视化显示 CLAP 优于原始 CLIP 和 Im.Aug，具有更清晰的类间区别和更紧密的类内聚类。

6 总结与展望

为了增强预先训练的类 CLIP 模型，从图片中提取出更加存粹的内容特征，减少风格特征的影响，该文深入研究了潜在内容变量的解缠。通过对视觉语言数据的潜在生成过程的因果分析，我们发现以一种模态训练解缠结网络可以有效地解开两种模态的内容。鉴于文本数据的高语义性质，该文发现通过文本增强干预在语言模态中更容易实现解缠结。基于这些见解，该文引入了 CLAP（增强提示对比学习）来获取解开的视觉语言内容特征。综合实验验证

了 CLAP 的有效性，证明了零样本和少样本性能的显着改进，并增强了抗扰动的鲁棒性。不过该文存在着一些局限性，该方法仅仅只针对那种图像中只存在一个类别的例子当中，并不适用于多类别图像当中，该方法也无法完全做到将图像当中的内容特征与风格特征做到完全解纠缠，未来在解纠缠这一方向上仍有着不小的探索可能。

参考文献

[1] Isabela Albuquerque, João Monteiro, Mohammad Darvishi, Tiago H Falk, and Ioannis Mitliagkas. Generalizing to unseen domains via distribution matching. *arXiv preprint arXiv:1911.00804*, 2019.

[2] Nicholas Carlini and Andreas Terzis. Poisoning and backdooring contrastive learning. *arXiv preprint arXiv:2106.09667*, 2021.

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[5] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

[6] Junhyeong Cho, Gilhyun Nam, Sungyeon Kim, Hunmin Yang, and Suha Kwak. Promptstyler: Prompt-driven style generation for source-free domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15702–15712, 2023.

[7] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[8] Stanislav Fort. Adversarial vulnerability of powerful near out-of-distribution detection. *arXiv preprint arXiv:2201.07012*, 2022.

[9] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024.

[10] Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[11] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[12] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[13] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[14] Tao Hong, Xiangyang Guo, and Jinwen Ma. Itmix: Image-text mix augmentation for transferring clip to image classification. In *2022 16th IEEE International Conference on Signal Processing (ICSP)*, volume 1, pages 129–133. IEEE, 2022.

[15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[16] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023.

[17] Lingjing Kong, Shaoan Xie, Weiran Yao, Yujia Zheng, Guangyi Chen, Petar Stojanov, Victor Akinwande, and Kun Zhang. Partial disentanglement for domain adaptation. In *International conference on machine learning*, pages 11455–11472. PMLR, 2022.

[18] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[19] Haoyang Li, Xin Wang, Ziwei Zhang, Zehuan Yuan, Hang Li, and Wenwu Zhu. Disentangled contrastive learning on graphs. *Advances in Neural Information Processing Systems*, 34:21872–21884, 2021.

[20] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021.

[21] Yuhang Liu, Zhen Zhang, Dong Gong, Mingming Gong, Biwei Huang, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Identifying weight-variant latent causal models. *arXiv preprint arXiv:2208.14153*, 2022.

[22] Yuhang Liu, Zhen Zhang, Dong Gong, Mingming Gong, Biwei Huang, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Identifiable latent polynomial causal models through the lens of change. *arXiv preprint arXiv:2310.15580*, 2023.

[23] Yuhang Liu, Zhen Zhang, Dong Gong, Mingming Gong, Biwei Huang, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Identifiable latent neural causal models. *arXiv preprint arXiv:2403.15711*, 2024.

[24] Aleksander Mądry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *stat*, 1050(9), 2017.

[25] Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. *arXiv preprint arXiv:2212.07016*, 2022.

[26] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[27] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[29] Mohammad Mahfujur Rahman, Clinton Fookes, Mahsa Baktashmotlagh, and Sridha Sridharan. Multi-component image translation for deep domain generalization. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 579–588. IEEE, 2019.

[30] Eduardo Hugo Sanchez, Mathieu Serrurier, and Mathias Ortner. Learning disentangled representations via mutual information estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 205–221. Springer, 2020.

[31] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14453–14463, 2023.

[32] Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021.

[33] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7959–7971, 2022.

[34] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9593–9602, 2021.

[35] Wenhan Yang, Jingdong Gao, and Baharan Mirzasoleiman. Robust contrastive language-image pretraining against data poisoning and backdoor attacks. *Advances in Neural Information Processing Systems*, 36, 2024.

[36] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[37] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022.

[38] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[39] Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990. PMLR, 2021.