arXiv:2304.03440v1 [cs.LG] 7 Apr 2023

利用异质相似度应对分布偏移的监督式对比学习

Takuro Kutsuna*

丰田中央研发实验室株式会社

*kutsuna@mosk.tytlabs.co.jp

摘要

分布偏移是指训练集与测试集之间的数据分布发生变化的问题，这可能会显著降低部署于现实世界中的模型性能。近期研究表明，性能下降的一个原因是某种形式的过拟合，而适当的正则化可以缓解这种性能下降，尤其是在使用神经网络等具有高度代表性的模型时。在本文中，我们提出了一种使用监督式对比学习的新型正则化方法，以防止此类过拟合，并训练出在分布偏移下性能不会下降的模型。我们将对比损失中的余弦相似度扩展为更一般的相似度度量，并提出在比较样本与正例或负例时使用该度量中的不同参数；分析表明，这相当于在对比损失中引入某种间隔。在模拟分布偏移的基准数据集上进行的实验，包括子群体偏移和领域泛化，证明了所提方法优于现有正则化方法。

1 引言

分布偏移是指训练集和测试集之间的数据分布发生变化的问题，这可能会显著降低部署在现实世界中的预测模型的性能。分布偏移可能有多种形式，但目前已确定以下两种类型 [21]：子群体偏移 和领域泛化。子群体偏移是指假设数据中存在不同群体，而这些群体在训练环境和测试环境中的占比发生变化的情况。例如，在行人检测中，收集训练数据时戴帽子的男性比例较低，但部署后趋势发生变化，戴帽子的男性比例显著上升；请注意，此处的群体根据帽子 ×性别定义。领域泛化是指数据收集环境在训练集和测试集之间发生变化的情况。在行人检测的例子中，这对应于使用在亚洲部署的检测器，而该检测器是使用在美国和欧盟收集的数据集训练的。

如 [32] 所报告的，当训练数据中存在少数群体时，子群体偏移会导致问题。具体而言，在使用具有较高表征能力的模型（例如深度神经网络（DNNs））时，有研究报告称，随着训练的进行，某些少数群体的泛化准确率会下降，而总体准确率却会提高。如果这样的少数群体在测试环境中转变为多数群体，就会导致准确率显著下降。这种现象可以看作是对训练数据中多数群体的过拟合。事实上，[32]表明，在训练过程中对深度神经网络应用一般性的正则化，可以防止少数群体的准确率下降。通过使用模拟领域泛化情形的基准数据集进行实验，我们观察到了与子群体偏移相似的现象。也就是说，随着训练的进行，用于训练的领域的准确率会提高，而训练数据中未包含的其他领域数据的准确率则会逐渐下降。如果将这一现象与子群体偏移情形作同样的理解，那么它可以被视为对用于训练的领域的一种过拟合。需要注意的是，在这种情况下，模型并不是以传统意义上的方式发生过拟合，因为通过训练，与训练数据属于同一领域的测试准确率得到了提高。

本文提出了一种新的正则化方法，以防止此类过拟合，并训练出在分布偏移（包括子群体偏移和领域泛化）下性能不会下降的深度神经网络。所提方法利用了在表示学习领域取得巨大成功的对比学习 [4, 13]。其中，我们采用了一种有监督的方法 [17] ，能够直接处理带标签的数据。在对比学习中，会为训练集中的每个样本定义正例和负例，并学习使嵌入接近正例、远离负例。对比损失中通常使用余弦相似度作为相似性度量，以比较嵌入。所提方法将这种余弦相似度扩展为一种更一般的相似性度量。有趣的是，仅改变相似度并不会显著改变使用对比损失得到的学习结果。相反，我们提出在将样本与正样本或负例进行比较时，在度量中使用不同的参数。我们表明，这种操作在对比损失中起到了类似间隔的作用。实验结果表明，与使用基本余弦相似度的方法相比，所提方法能够学习出对分布偏移更加鲁棒的模型。

本文的贡献如下：

- 我们提出一种正则化方法，将监督式对比损失扩展到深度神经网络的训练中，使其能够抵御分布偏移；并证明，在损失函数中调整相似度等同于建立一个间隔。

- 通过使用基准数据集，我们表明，过拟合不仅会出现在领域泛化场景中，也会出现在子群体偏移下的少数群体中。

- 我们通过实验表明，所提方法能够缓解子群体偏移和领域泛化设置下的性能下降。实验结果还表明，所提方法比用于深度神经网络的传统正则化更有效。

在第 2节中，我们回顾有关子群体偏移和领域泛化的现有研究。在第3节中，我们简要介绍对比学习中使用的损失以及一些符号。我们在第 4 节中解释所提方法，并在第 5 节中展示实证结果。第 6 节总结全文。

2 相关工作

分布鲁棒优化与子群体偏移。与假设训练集和测试集数据具有相同分布的经验风险最小化（ERM） [36], 不同，分布鲁棒优化（DRO）假设测试环境中的输入 x 属于一个分布集合 $\mathcal{P}$ ，并优化模型以最小化 $\mathcal{P}$上的最坏情况损失。尽管早期的DRO研究将训练分布周围的一个 ϵ 球视为 $\mathcal{P}$，但后来发现，具有此类 $\mathcal{P}$ 的DRO会陷入与使用训练集分布训练的同一模型 [15]。为了使DRO具有实际意义，Hu等人 [15] 建议为 x 假设一个潜变量 z ，并将 $\mathcal{P}$ 定义为依赖于 z。通过将 x 的群体视为 z，DRO可以成为缓解子群体偏移问题的一种方法。已有多项研究提出了依赖于 z 的 $\mathcal{P}$ 定义，以及最小化此类 $\mathcal{P}$ [32, 23, 31, 39]上的最坏情况损失的算法。另一条研究路线尝试自动寻找 x 的潜在群体，并根据所找到的群体应用 DRO [25, 1]。最近，Zhang等人提出了一种基于对比学习 [38]的该研究路线方法。 [32] 报告的一项重要发现是，在使用深度神经网络等复杂模型时，简单地应用DRO可能无法提升最差群体的泛化性能。为了防止此类问题，在训练深度神经网络时应用适当的正则化至关重要，正如 [32]中所讨论的。

领域泛化与表示学习。在介绍领域泛化之前，我们先简要介绍领域适应。给定源域中的数据集或模型，领域适应 旨在为具有标注数据集的目标域获取一个模型。无监督领域适应是领域适应的一种变体，其中目标域仅提供未标注数据。领域泛化 的目标同样是为目标域获得一个良好的模型，但前提是假设我们没有目标域的数据。从这个意义上说，目标域也称为分布外（OOD），而源域称为分布内（ID）。正如 [40], 中所综述的，已有多种领域泛化方法被提出，例如领域对抗学习 [24]。

表示学习是一种用于学习嵌入模型的方法，这些模型在迁移到下游任务时表现出高性能，并且近年来取得了显著进展。表示学习与领域泛化密切相关。如果表示学习能够为 ID 数据生成良好的嵌入模型，那么这些模型对 OOD 数据同样有效。事实上，许多最初为表示学习开发的方法已被应用于领域泛化任务，例如一种使用拼图游戏的方法 [3]。

对比学习　　arning 是一种表示学习，旨在学习嵌入　　-

这些方法使正样本对保持接近，同时将负样本对分离。Hadsell 等人 [12] 提出了对比学习的早期形式，后来在 [29]中被用于领域泛化。Schroff 等人 [33] 提出了三元组损失，后来在 [10]中被用作领域泛化训练损失的一部分。正如 [18], 中所讨论的，这些对比方法的性能对负样本的选择较为敏感。为缓解这一问题，有人提出了不需要负样本的对比学习方法 [11, 6], ，在此基础上，Kim 等人 [18] 提出了一种领域泛化方法。解决这一问题的另一条研究路线是使用 InfoNCE 损失[1] [30]，其中针对每个锚点样本使用多个负样本。研究表明，通过增加损失中使用的负样本数量，使用 InfoNCE 进行表征学习的性能会得到提升，这主要在自监督设置 [5, 13]中进行了探索。Khosla 等人 [17]展示了 InfoNCE 在普通监督分类问题中的优势，并将其称为监督式对比学习（监督对比学习）。本文提出了一种基于监督对比学习的领域泛化方法，其中相似度函数从通常的余弦相似度进行了扩展。最近，Sun 等人 [35] 提出了一种基于监督对比学习的分布外检测方法，这与我们的研究范围不同；我们的研究旨在防止分布偏移（包括子群体偏移和领域泛化）导致的性能下降。

3 预备知识

在本节中，我们简要回顾采用 InfoNCE 的对比学习，并给出解释所提方法所需的符号表示。

3.1 对比学习

在对比学习中，对于输入 $x \in X$，给定一组在某种意义上与 x 相似的正例 $\mathcal{S}_x^p$ ，以及一组与 x 不相似的负例 $\mathcal{S}_x^n$ ，通过最小化以下损失函数来学习嵌入函数 $f_\theta \colon X \to \mathbb{R}^m$ 。

$$\mathcal{L}_{\mathrm{cl}}(\theta) := -\frac{1}{|\mathcal{S}_x^p|} \sum_{x_p \in \mathcal{S}_x^p} \log \frac{e^{\mathrm{sim}(f_\theta(x), f_\theta(x_p))/\tau}}{e^{\mathrm{sim}(f_\theta(x), f_\theta(x_p))/\tau} + \sum_{x_n \in \mathcal{S}_x^n} e^{\mathrm{sim}(f_\theta(x), f_\theta(x_n))/\tau}}, \tag{1}$$

其中，sim(·, ·)是用于计算嵌入向量之间相似度的函数，通常为余弦相似度，而 τ 是温度参数。在上述损失中， f_θ 经过学习，使得 x 的嵌入与正样本 x_p 的嵌入相比于负样本 x_n 的嵌入更加相似。

3.2 自监督对比学习

自监督对比学习 [5, 13]使用输入 x 来学习 f_θ，其假设是，与 x 对应的标签 y 未被提供。作为 x的正样本， x 本身是

[1]这种损失也称为多类别 N-pair 损失 [34]。

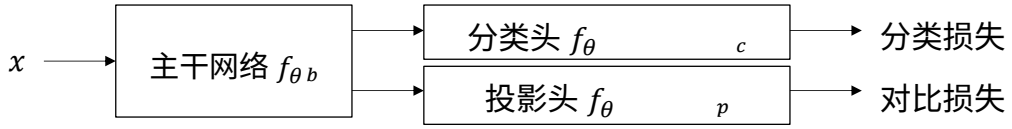


图1：监督式对比学习的模型结构

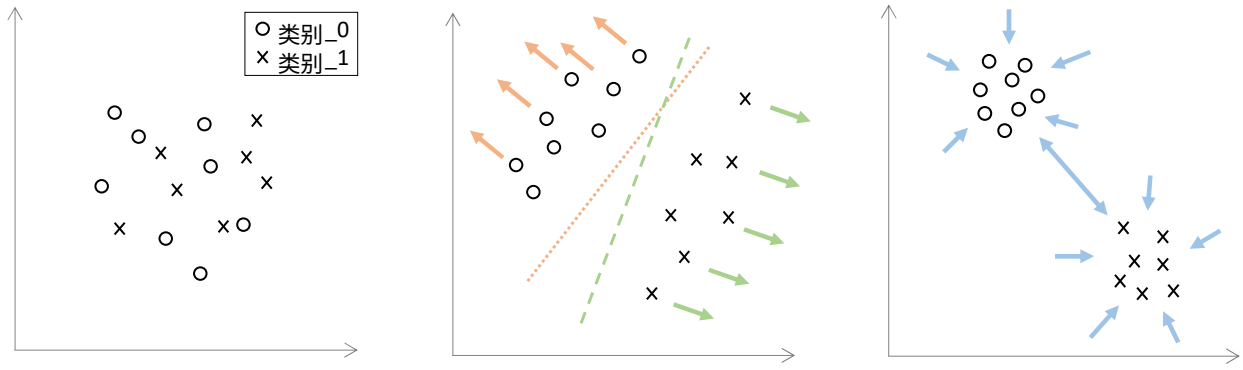
在应用了诸如裁剪或水平翻转等数据增强后使用。自监督对比学习通常会为每个 x($|\mathcal{S}_x^p| = 1$)仅生成一个正样本。除 x (以及对它们进行的任何数据增强)之外的所有图像都被视为负例。文献已经表明，增大 $\mathcal{S}_x^n$ 对于提升对比损失 [5]的性能非常重要。然而，如果将 $\mathcal{S}_x^n$ 的大小设置得过大，计算机将耗尽内存，因此在典型的计算环境中无法进行学习。动量对比 (MoCo) [13]是一种缓解此类问题的方法，它使用由模型计算得到的嵌入队列来有效增大 $\mathcal{S}_x^n$ 的大小，该模型在学习过程中采用移动平均。

3.3 监督式对比学习（监督对比学习）

监督对比学习 [17] 用于处理这样一种情况：输入 x具有给定的离散标签 y 。对于 x而言，数据集 x 中除其自身以外、具有相同标签 y 的样本被视为正例，而标签不同的样本则被视为负例 x。因此，预期具有相同标签的嵌入会在嵌入空间中聚集。为了有效增大 $\mathcal{S}_x^n$，MoCo 等自监督对比学习方法也适用于监督场景。我们在实验中采用了 MoCo，并通过队列增大了 $\mathcal{S}_x^p$ 和 $\mathcal{S}_x^n$ 。

仅使用损失函数（1）进行学习无法得到能够预测标签 y的模型，因此除了用于对比损失的投影头之外，还需要准备一个用于分类的头部模型。图1展示了模型结构概览，其中 f_{θ_b}、 f_{θ_c}和 f_{θ_p} 分别表示主干模型、分类头和投影头。分类头f_{θ_c} 是一个线性层，其输出用于通过分类损失进行训练，分类损失通常为交叉熵损失（CE）。投影头f_{θ_p} 源自自监督对比学习，在监督对比学习 [17], 中也会使用；我们按照 [7]采用了一个三层多层感知机。f_{θ_p} 的输出作为损失函数（1）中的嵌入，其中 $f_\theta := f_{\theta_p} \circ f_{\theta_b}$。

图2展示了训练期间嵌入空间中的样本，体现了交叉熵与监督对比学习之间的差异。使用交叉熵进行训练时，嵌入函数会被学习为使每个样本尽可能远离当前分类平面，如图2b所示。相比之下，使用监督对比学习进行训练会促使同一类别的样本聚集，同时让不同类别的样本彼此分开，如图2c所示。因此，交叉熵与监督对比学习的主要区别之一，在于是否有意将同一类别的样本聚集起来。换言之，监督对比学习试图学习这样的嵌入：它们是



(a) 初始嵌入。 (b) 使用交叉熵进行训练。 (c) 使用监督对比学习进行训练。

图2：使用交叉熵或监督对比学习进行训练的示意图：(a) 初始模型的嵌入空间。(b) 使用交叉熵进行训练，其中虚线表示由最后一个线性层定义的类别投影。(c) 使用监督对比学习进行训练。箭头表示每种损失所产生的作用力。

对于同一类别保持不变量，同时保留区分类别所必需的特征，预计这将产生一个更能体现各类别独有特征的嵌入空间，从而带来更强的泛化能力。正如我们的实验所表明的，对于每个类别，嵌入聚合得有多好取决于该类别的特征。

4 所提方法

我们提出了一种新颖的正则化方法，用于监督对比学习，通过扩展损失中使用的相似度函数来实现；为此，我们使用 t-vMF 相似度 [20]，，它推广了余弦相似度。

4.1 t-vMF 相似度

t-vMF相似度是通过扩展 von Mises-Fisher（vMF）分布得到的。vMF分布是定义在超球面上的正态分布；扩展后的分布具有重尾特性，其命名源于它与正态分布的 student-t 扩展之间的类比。两个 m维向量 a, b 的 t-vMF 相似度定义为

$$\phi_\kappa(a,b) := \frac{1+\cos(a,b)}{1+\kappa\left(1-\cos(a,b)\right)} - 1, \tag{2}$$

其中 $\kappa \in (-1/2, \infty)$ 是 t-vMF 相似度的超参数，cos 是余弦相似度。根据定义可知，当 $\kappa = 0$时，t-vMF 相似度等于余弦相似度。图 3 展示了$\kappa \in \{-0.4, -0.2, 0, 0.3, 2.0\}$的 t-vMF 相似度图示例，其中 x 轴表示输入向量之间的夹角。我们

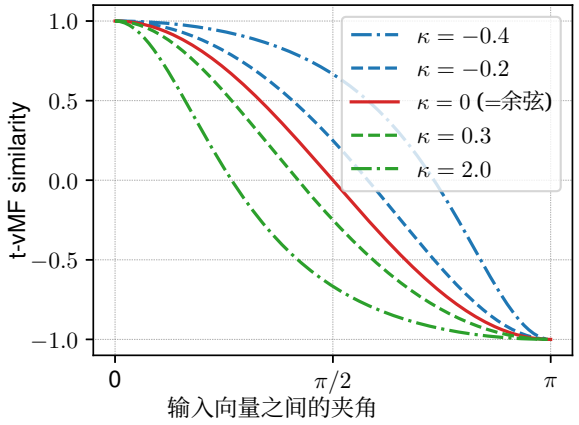


图 3：t-vMF 相似度图。

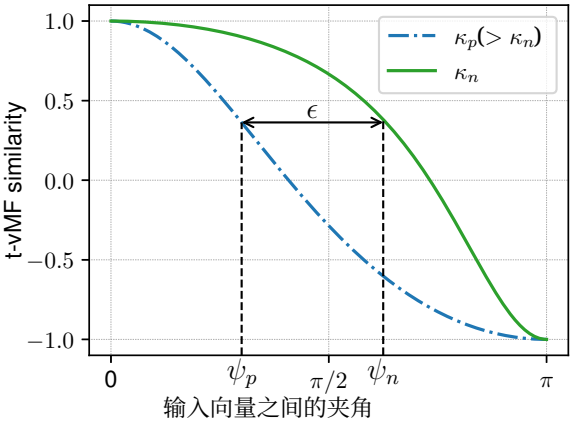


图 4： $\kappa_p > \kappa_n$的间隔 ϵ 。

可以看到，当夹角为 0 时，相似度为 1；当夹角为 π 时，相似度为 -1 ，与 κ 的取值无关。还可以看到，对于 $\kappa = 0$（对应余弦相似度），夹角为 $\pi/2$ 时相似度为 0；而对于 $\kappa < 0$ ，夹角为 $\pi/2$ 时相似度为正，夹角为 $\kappa > 0$ 时则为负。对于 $\kappa_1 < \kappa_2$，有 $\phi_{\kappa_1}(a,\ b) \geq \phi_{\kappa_2}(a,\ b)$，且仅当夹角为 0 或 π 时等号成立。

4.2 所提方法损失

我们提出以下损失函数，将 t-vMF 相似度用作对比损失（1）中的相似度：

$$\tilde{\mathcal{L}}_{\text{cl}}(\theta) := -\frac{1}{|\mathcal{S}_x^p|} \sum_{x_p \in \mathcal{S}_x^p} \log \frac{e^{\phi_{\kappa_p}(f_\theta(x), f_\theta(x_p))/\tau}}{e^{\phi_{\kappa_p}(f_\theta(x), f_\theta(x_p))/\tau} + \sum_{x_n \in \mathcal{S}_x^n} e^{\phi_{\kappa_n}(f_\theta(x), f_\theta(x_n))/\tau}} \tag{3}$$

关键在于对正例和负例分别使用不同的 κ 、 κ_p 和 κ_n。有趣的是，正如我们的实验所表明的，只要 $\kappa_p = \kappa_n$，使用余弦相似度和t-vMF 相似度时的测试准确率就没有显著差异。相反，当 $\kappa_p > \kappa_n$时，结果存在明显差异，在我们的实验中，分布偏移下的准确率得到了提升。造成这种差异的原因可以如下进行分析。首先，我们将损失函数 (3) 变换为

$$\tilde{\mathcal{L}}_{\text{cl}}(\theta) = -\frac{1}{|\mathcal{S}_x^p|} \sum_{x_p \in \mathcal{S}_x^p} \log \frac{1}{1 + \sum_{x_n \in \mathcal{S}_x^n} e^{\left(\phi_{\kappa_n}(f_\theta(x), f_\theta(x_n)) - \phi_{\kappa_p}(f_\theta(x), f_\theta(x_p))\right)/\tau}}. \tag{4}$$

如果我们定义 $\delta := \phi_{\kappa_n}(f_\theta(x)\ f_\theta(x_n)) - \phi_{\kappa_p}(f_\theta(x)\ f_\theta(x_p))$，那么有必要降低 δ 以减小 $\tilde{\mathcal{L}}_{\text{cl}}$。现在，我们分别将 x 与 x_p 和 x_n 所成的角记为 ψ_p 和 ψ_n，

那么 $\kappa_p = \kappa_n$ 满足以下关系：

$$\begin{aligned}\psi_n - \psi_p > 0 \quad &\rightarrow \quad \delta < 0,\\ \psi_n - \psi_p = 0 \quad &\rightarrow \quad \delta = 0,\\ \psi_n - \psi_p < 0 \quad &\rightarrow \quad \delta > 0.\end{aligned}$$

也就是说，$\delta = 0$如果与正样本和负例的夹角相同，并且$\delta < 0$ 如果与正样本的夹角小于与负例的夹角，则表明 x 与正样本的相似度高于与负例的相似度。相比之下，对于 $\kappa_p > \kappa_n$，以下关系成立：

$$\begin{aligned}\psi_n - \psi_p > \epsilon \quad &\rightarrow \quad \delta < 0,\\ \psi_n - \psi_p = \epsilon \quad &\rightarrow \quad \delta = 0,\\ \psi_n - \psi_p < \epsilon \quad &\rightarrow \quad \delta > 0,\end{aligned}$$

其中，ϵ （> 0）是一个非负间隔，其取值取决于 κ_p、κ_n、ψ_p 和 ψ_n。图 4 展示了这种 ϵ 的一个示例。因此，如果 $\kappa_p > \kappa_n$，则 ψ_p 必须比 ψ_n 小至少一个正间隔 ϵ，才能使 $\delta < 0$。这种间隔将在对比损失中起到某种正则化作用，从而在分布偏移下获得更高的准确率。

文献中也提出过类似的间隔，用于采用交叉熵进行训练。ArcFace [9] 是一种用于人脸识别任务的方法，它在损失函数中引入固定间隔。LDAM [2]提出了一种用于解决长尾问题的损失函数，其中加入了依赖于类别的固定间隔。所提方法与这些方法的区别如下：首先，所提方法应用于对比损失，而现有方法均应用于交叉熵。其次，所提方法使用基于 t-vMF 相似度的可变间隔，而现有方法使用固定间隔。

控制 κ_p 和 κ_n 只需一个参数。 在 $\kappa_p > \kappa_n$ 下分别控制 κ_p 和 κ_n 在实践中并不方便。为缓解这一问题，我们提出使用单个参数 $\alpha \in [0, 0.5)$ 来控制 κ_p 和 κ_n，具体公式如下：

$$\kappa_p = \frac{\alpha}{2\alpha + 1}, \quad \kappa_n = -\alpha$$

这些方程是由条件[2] 即与 κ_p 和 κ_n 的相似度在角度 $\pi/2$处变得对称推导而来的。例如，$\alpha = 0.4$ 得到 $(\kappa_p, \kappa_n) = (2.0, -0.4)$，如图3所示。请注意，当我们将 α 设得更大时，$\kappa_p - \kappa_n$ 也会变大，这对应于我们方法中更强的正则化。

5 实验结果

我们使用模拟子群体偏移和领域泛化的基准数据集，展示了所提方法的实用性。

[2] $1/(1+\kappa_p) - 1 = -(1/(1+\kappa_n) - 1)$

表1：实验中使用的数据集的列联表。

(a) CelebA（训练集划分）

	男性	女性
非金发	66874	71629
金发	1387	22880

(b) Camelyon17

中心	$\mathcal{D}_{\text{训练集}}$	$\mathcal{D}_{\text{test}}^{\text{ID}}$	$\mathcal{D}_{\text{test}}^{\text{OOD1}}$	$\mathcal{D}_{\text{test}}^{\text{OOD2}}$
0	53425	6011	0	0
1	0	0	34904	0
2	0	0	0	85054
3	116959	12879	0	0
4	132052	14670	0	0

5.1 数据集

为了评估子群体偏移情况，我们使用CelebA 数据集 [26], ，该数据集由带有多个标注的人脸图像组成。遵循 [32], ，我们考虑预测图像中人物的发色是否为金发的问题[3]。每张训练图像都属于由发色和性别定义的四个群体之一，其列联表见表1a，由此我们可以看出，金发男性占少数。CelebA 数据集的测试集约包含20000 张测试图像，其中所包含数据的比例也与表1a中的比例大致相同。

我们使用 Camelyon17-WILDS 数据集 [21], （为简便起见，以下简称 Camelyon17）来评估领域泛化情境。Camelyon17 由从五家医院收集的组织病理学图像组成。我们的目标是预测标签 y ，该标签表示图像是否包含肿瘤组织。Camelyon17 数据集分为四个子集[4], ，分别称为 $\mathcal{D}_{\text{train}}$、 $\mathcal{D}_{\text{test}}^{\text{ID}}$、 $\mathcal{D}_{\text{test}}^{\text{OOD1}}$ 和 $\mathcal{D}_{\text{test}}^{\text{OOD2}}$。表1b总结了各子集来自哪家医院，其中医院称为中心。众所周知，由于显微镜类型和染色方法不同，显微图像会因医院而呈现不同的分布特征 [22]。请注意，标签 y 经过调整，以确保在每个子集中均匀分布。

5.2 模型和训练设置

我们在图1中使用ResNet18 [14] 作为主干网络模型。应用监督对比学习时，我们采用了 MoCo [13], ，其中队列大小根据 [13]设置为65536。除非另有说明，监督对比学习损失中的温度 τ 设置为1，分类损失和监督对比学习损失的损失系数均设置为1。对于CelebA的分类损失，我们使用了两种群体分布鲁棒优化损失，即Robust-DRO损失 [32] 和CVaR-DRO 损失 [31], ，使用表1a中根据发色和性别定义的四个群组。我们使用了由[5] 实现的

[3]我们使用 CelebA 数据集进行了另一项实验，目标是预测图像中的人物是否在微笑。我们使用“嘴巴微微张开”作为混杂变量。结果见附录 A。[4]这些划分在原始数据中分别称为 train、id val、val 和 test。[5] https://github.com/kohpangwei/group_DRO（07/03/2023）

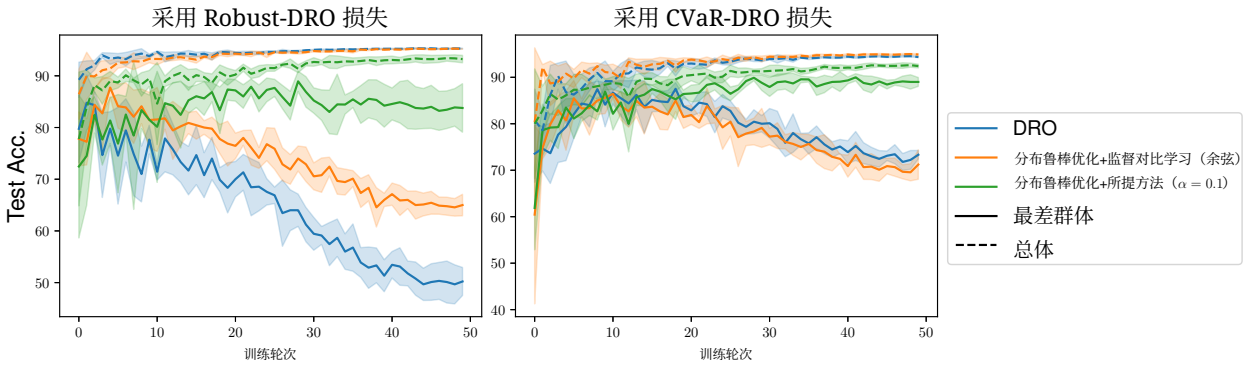


图 5：CelebA 的最差群体（实线）和总体（虚线）测试准确率。

鲁棒分布鲁棒优化的作者以及与CVaR-DRO 相关的论文作者 [39] 所提供的实现[6] 。在使用群体分布鲁棒优化损失进行训练时，我们根据 [32]对样本进行重新采样，以平衡小批次中的群体。对于 Camelyon17，我们使用交叉熵作为分类损失。

训练时使用 Adam 优化器 [19] ，CelebA 训练 50 个训练轮次，Camelyon17 训练 25 个训练轮次，初始学习率为 0.001。除非另有说明，训练过程中使用无重启余弦调度 [27]衰减学习率，权重衰减设为 0。除使用 RandAugment [8]的实验外，数据增强采用随机裁剪和随机水平翻转。实验中未使用预训练模型。每种设置使用不同的随机种子训练 5 次，并以结果的均值和标准差进行评估。所有实验均使用 MMClassification [28]运行。

5.3 子群体偏移的结果

我们首先使用 CelebA 数据集比较三种模型：一个使用群体分布鲁棒优化损失训练的模型、一个使用群体分布鲁棒优化损失和具有余弦相似度的监督对比学习损失训练的模型，以及一个使用第 4节中介绍的具有 $\alpha = 0.1$ 的所提损失和群体分布鲁棒优化损失训练的模型。训练期间的测试准确率曲线如图 5所示，其中实线和虚线分别表示最差群体准确率和总体准确率。请注意，正如第 1节所讨论的，在子群体偏移假设下，提高最差群体准确率非常重要。从图5可以看出，随着训练的进行，即使应用了群体分布鲁棒优化损失，最差群体准确率仍会下降；这一现象与 [32] 中报告的结果一致。相比之下，无论使用哪种分布鲁棒优化损失，通过应用所提损失，我们都可以防止最差群体损失下降。使用普通余弦相似度的监督对比学习损失与鲁棒分布鲁棒优化损失结合时，可以略微缓解这种下降，但与 CVaR-DRO 结合时几乎没有效果。

为了研究在所提损失中使用 κ_p 和 κ_n 使得 $\kappa_p \neq \kappa_n$ 的重要性，我们使用 Robust-DRO 损失以及带有 $\kappa_p = \kappa_n$的损失 (3) 训练了模型。图 6 展示了结果，表明与使用余弦相似度相比，仅将 t-vMF 相似度用作监督对比学习损失的相似度度量

[6]https://github.com/violet-zct/group-conditional-DRO（2023年3月7日）

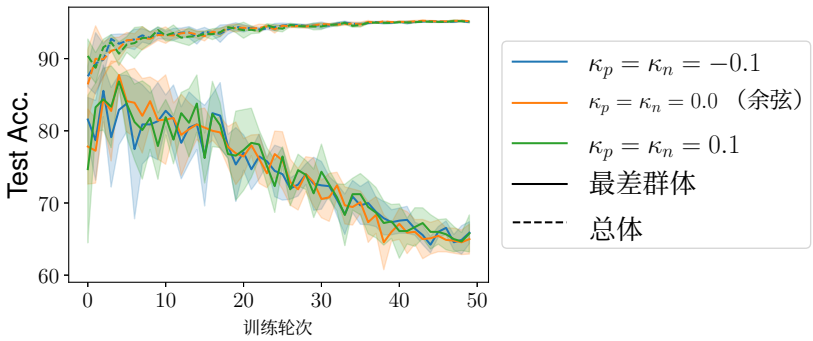


图 6：当 $\kappa_p = \kappa_n$ 时，使用 Robust-DRO 损失训练的 CelebA 测试准确率。

影响很小。正样本和负样本使用不同的 κ至关重要，这会产生一种间隔效应，正如第 4 节所讨论的那样。

我们将所提方法与深度神经网络的标准正则化方法进行了比较：权重衰减（WD）、随机深度（SD）[16], 以及 RandAugment [8]。虽然 SD 仅适用于具有特定结构的深度神经网络模型（如 ResNet），但文献表明它是一种有效的正则化方法 [37]。RandAugment 是一种数据增强方法，而非正则化方法，但预期可通过强数据增强来防止过拟合。对于这些方法，我们使用了 [28] 中的实现。表 2 总结了结果，评估训练结束时最差群体的测试准确率[7]。我们比较了上述两种分布鲁棒优化损失，以及包括固定学习率和余弦衰减在内的学习率调度方式。从表 2 可以看出，与仅使用分布鲁棒优化损失训练的模型相比，应用正则化方法能够提高最差群体的准确率。无论学习率和损失类型如何，所提方法均取得了与现有正则化方法相当或更好的结果。

我们还通过改变多个参数，将所提方法与采用余弦相似度的监督对比学习损失进行了比较。首先，我们增大了与分布鲁棒优化损失结合时监督对比学习的损失权重，以增强监督对比学习在训练中的效果。接着，我们改变了温度 τ。我们还实现了带有固定边际的监督对比学习，其形式类似于 [9], ，用 cos(θ_p)替换cos($\theta_p + m$)，其中 θ_p 是式(1) 中 $f_\theta(x)$ 与 $f_\theta(x_p)$ 之间的角度，而 m 是固定边际。注意， $m = 0.5$ 用于 [9]。结果汇总于表 3中，其中训练采用了分布鲁棒优化损失。可以看出，对于 $\alpha \geq 0.1$，所提方法的表现优于其他设置。

5.4 领域泛化的结果

我们使用 $\mathcal{D}_{\text{train}}$ 从 Camelyon17 数据集训练了三个模型：一个使用交叉熵训练的模型、一个使用交叉熵和余弦相似度监督对比学习训练的模型，以及一个使用交叉熵和所提损失函数（ $\alpha = 0.1$）训练的模型。我们使用 $\mathcal{D}_{\text{test}}^{\text{OOD1}}$、 $\mathcal{D}_{\text{test}}^{\text{OOD2}}$和 $\mathcal{D}_{\text{test}}^{\text{ID}}$分别评估训练好的模型。图 7 展示了训练期间的测试准确率曲线。有趣的是，使用交叉熵训练的模型取得了

[7]总体准确率汇总于附录 A。

表 2：CelebA 上最差群体的测试准确率。

	固定学习率		衰减学习率		
	鲁棒分布鲁棒优化	CVaR-DRO	鲁棒分布鲁棒优化	CVaR-DRO	平均值
DRO损失	54.44 (±3.82)	67.56 (±8.04)	50.22(±2.71)	73.33 (±0.93)	61.39
+ 权重衰减(1e-4)	78.89(±6.97)	79.62 (±6.85)	65.44(±2.49)	85.56 (±1.05)	77.38
+ 权重衰减(1e-3)	79.71 (±6.73)	75.50 (±14.56)	81.33(±3.12)	88.33 (±0.86)	81.22
+ 权重衰减(1e-2)	82.26 (±4.97)	61.28 (±13.84)	87.33(±2.47)	88.97 (±1.44)	79.96
+ 标准差(0.1)	60.44 (±2.29)	76.89 (±6.01)	58.56 (±1.88)	77.78 (±2.25)	68.42
+ 标准差(0.3)	73.22 (±4.06)	84.00 (±3.40)	74.44(±1.22)	85.22 (±1.25)	79.22
+ 标准差(0.5)	82.78 (±5.01)	85.78 (±5.15)	82.67(±0.42)	88.56 (±0.75)	84.95
+ RandAug	78.33(±4.47)	87.27 (±1.57)	76.56(±0.96)	90.11 (±1.12)	83.07
+ 监督对比学习（余弦）	74.89 (±5.36)	73.44 (±4.19)	65.00(±2.08)	71.22 (±3.17)	71.14
+ 所提方法(0.05)	83.44(±2.03)	84.78 (±3.31)	79.56(±2.75)	80.56 (±0.99)	82.09
+ 所提方法(0.1)	86.56 (±0.96)	82.90 (±4.14)	83.78(±4.69)	89.00 (±0.96)	85.56
+ 所提方法(0.15)	84.67(±2.91)	85.35 (±4.81)	88.44(±0.96)	89.24 (±1.47)	86.93
+ 所提方法(0.2)	85.46 (±2.24)	80.74 (±1.27)	85.78(±1.47)	88.66 (±1.94)	85.16

在 $\mathcal{D}_{\text{test}}^{\text{ID}}$上较高的准确率，但随着训练的进行，它在 $\mathcal{D}_{\text{test}}^{\text{OOD1}}$和 $\mathcal{D}_{\text{test}}^{\text{OOD2}}$上的准确率却逐渐下降。这种表现与使用 CelebA 训练时的最差群体准确率相似。需要注意的是，这种性能下降不同于通常所说的过拟合，因为模型在 $\mathcal{D}_{\text{test}}^{\text{ID}}$上取得了较高的准确率。从图 7 可以看出，所提方法能够防止模型在分布外数据集上的性能下降，但使用所提损失函数训练时，模型在 $\mathcal{D}_{\text{test}}^{\text{ID}}$ 上的准确率略有下降。

我们使用 Came- lyon17 数据集，将所提方法与其他正则化方法进行了比较。表4汇总了结果，其中每个测试集划分上的准确率均在训练结束时进行评估。可以看到，

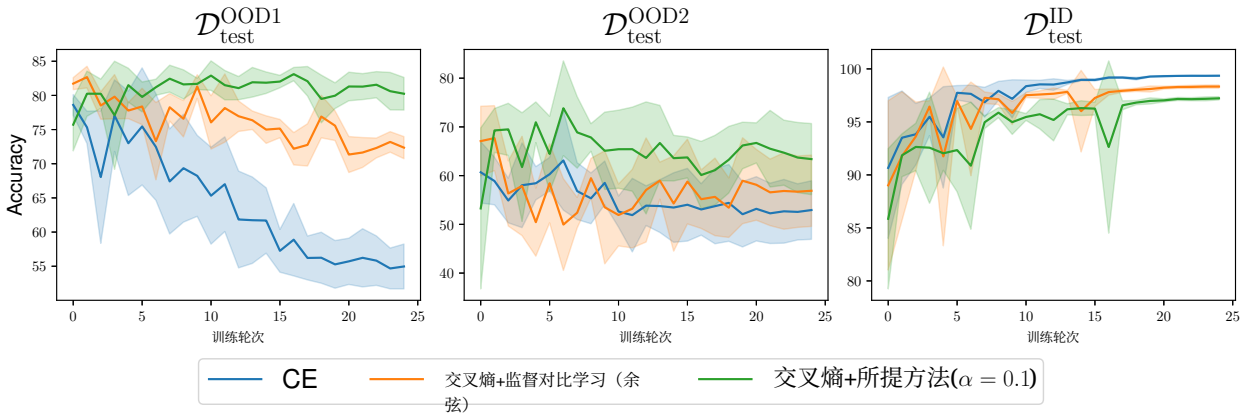


图 7：Camelyon17 数据集测试集划分上的准确率。

表 3：与其他监督对比学习设置的比较

	最差群体
监督对比学习（余弦）	65.00（±2.08）
+损失权重(5)	73.89 (±1.27)
+损失权重(10)	76.11 (±2.25)
+ 温度(0.1)	43.56 (±0.97)
+ 温度(10)	76.78 (±4.24)
+固定边际(0.1)	81.00（±1.08）
+固定边际（0.3）	81.78 (±5.34)
+固定边际(0.5)	83.11 (±4.41)
所提方法(0.05)	79.56 (±2.75)
所提方法(0.1)	83.78 (±4.69)
所提方法(0.15)	88.44 (±0.96)
所提方法(0.2)	85.78（±1.47）

应用正则化后，OOD 划分上的准确率有所提升，这与 CelebA 数据集上最差群体准确率的提升类似。所提方法在 OOD 划分上的表现与现有正则化方法相当或更好，尤其是在 $\mathcal{D}_{\text{test}}^{\text{OOD1}}$ 上；不使用正则化进行训练时，其表现会显著下降。

为了研究 SupCon 与 CE 结合进行训练的效果，我们使用主成分分析（PCA）对主干网络模型获得的嵌入进行了可视化。图8展示了 $\mathcal{D}_{\text{train}}$中样本的嵌入，其中点的颜色对应样本的来源医院，“o”和“x”标记分别表示 $y = 0$ （无肿瘤）和 $y = 1$ （肿瘤）[8]。我们可以看到，仅使用 CE 训练的模型所获得的嵌入在嵌入空间中彼此分离，以区分标签。相比之下，使用 CE 和 SupCon，或 CE 和所提方法获得的嵌入，在相同类别内聚集得更紧密；但来自中心-4的 $y = 1$ 样本除外，这可能是因为中心-4的肿瘤样本特征与其他中心的不同。图8中的结果与图2中的讨论一致。

所提方法的局限性。我们还使用 iWildCam 数据集 [21] 作为另一个数据集进行实验，以评估领域泛化能力。与 Camelyon17 的结果相比，我们发现，使用交叉熵、交叉熵和监督对比学习，或交叉熵和所提方法训练的模型，其分布外性能差异很小。与 Camelyon17 不同，我们没有观察到训练期间 iWildCam 的分布外划分出现任何性能下降。原因尚不明确，但 iWildCam 的训练集划分包含来自许多地点的样本，这些地点与 Camelyon17 中的医院相对应，这或许避免了过拟合。这些结果表明，所提方法能够缓解分布外划分上的性能下降，但在没有发生性能下降时作用有限，

[8]测试集划分中的嵌入可视化结果见附录 C。

表4：Camelyon17 数据集测试集划分上的准确率比较。

	$\mathcal{D}_{\text{test}}^{\text{OOD1}}$	$\mathcal{D}_{\text{test}}^{\text{OOD2}}$	$\mathcal{D}_{\text{test}}^{\text{ID}}$
CE	54.97 (±3.28)	52.94 (±6.01)	99.36 (±0.03)
+ 权重衰减(1e-4)	60.50 (±1.61)	68.13 (±8.56)	99.00 (±0.06)
+ 权重衰减(1e-3)	62.17 (±7.29)	59.23 (±5.42)	98.57 (±0.06)
+ 权重衰减(1e-2)	79.97 (±3.06)	59.91 (±9.10)	97.87 (±0.10)
+ 标准差(0.1)	55.22 (±2.63)	58.91 (±2.88)	99.31 (±0.01)
+ 标准差(0.3)	58.32 (±2.54)	63.13 (±3.96)	99.20 (±0.04)
+ 标准差(0.5)	67.15 (±4.89)	64.17 (±2.59)	98.94 (±0.06)
+ RandAug	72.69 (±4.27)	66.85（±2.58）	98.92（±0.03）
+ 监督对比学习（余弦）	72.35 (±1.62)	56.89（±7.32）	98.33（±0.15）
+ 所提方法(0.05)	74.63（±3.49）	55.21 (±6.84)	97.40 (±0.21)
+ 所提方法(0.1)	80.24 (±2.39)	63.39 (±7.30)	97.22 (±0.16)
+ 所提方法(0.15)	80.18 (±1.87)	59.45 (±10.02)	97.01 (±0.39)
+ 所提方法(0.2)	79.66 (±1.27)	53.70 (±11.89)	97.21 (±0.31)

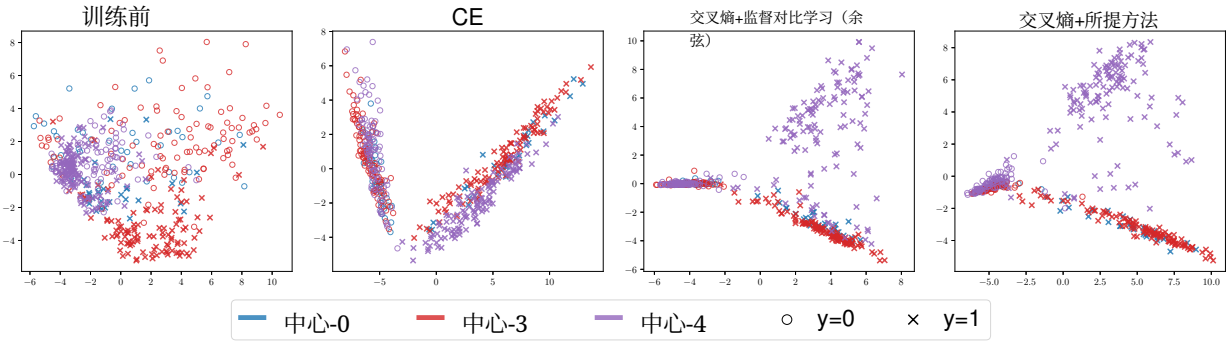


图 8： $\mathcal{D}_{\text{train}}$中样本嵌入的主成分分析可视化。

也不会产生不良副作用。详情请参见附录 B。

6 结论

我们提出了一种新颖的方法，用于对深度神经网络训练进行正则化，以提升其在分布偏移下的泛化能力。所提方法使用t-vMF相似度扩展了监督式对比损失，而 t-vMF相似度是余弦相似度的推广形式。我们通过解析研究发现，通过为正例和负例使用不同的相似度参数，可以在对比损失中建立某种间隔。实验结果表明，当使用t-vMF 等广义相似度扩展对比损失时，采用这种不同的相似度至关重要。实验证明，在使用基准数据集的深度神经网络中，所提方法的性能优于标准正则化方法，这些

模拟分布偏移。未来工作包括研究所提出的方法对自监督对比学习的适用性。

参考文献

[1] Bao,Y., Chang, S., Barzilay, R.:预测后插值：一种学习稳定类别的简单算法。载于：国际机器学习会议。第640–650页（2021年）[2] Cao, K., Wei, C., Gaidon,A.,Arechiga,N., Ma, T.: 使用标签分布感知间隔损失学习不平衡数据集。神经信息处理系统进展**32** （2019年）[3] Carlucci, F.M., D'Innocente, A., Bucci, S., Caputo,B., Tommasi,T.: 通过解决拼图游戏实现领域泛化。载于：计算机视觉与模式识别会议（2019年）[4] Chen, T., Kornblith, S.,Norouzi,M., Hinton,G.: 用于视觉表示对比学习的简单框架。载于：国际机器学习会议。第1597–1607页（2020年）[5] Chen, T., Kornblith, S.,Norouzi,M., Hinton,G.: 用于视觉表示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）[6] Chen, X.,He,K.: 探索简单的孪生表示学习。载于：计算机视觉与模式识别会议。第15750–15758页（2021年）[7] Chen,X., Xie, S.,He, K.: 自监督视觉变换器训练的实证研究。载于：国际计算机视觉会议。第9640–9649页（2021年）[8]Cubuk, E.D., Zoph, B., Shlens, J.,Le,Q.V.:RandAugment：在缩减搜索空间下实用的自动化数据增强。载于：计算机视觉与模式识别会议。第702–703页（2020年）[9] Deng,J., Guo, J.,Xue,N.,Zafeiriou,S.:ArcFace：用于深度人脸识别的加性角度间隔损失。载于：计算机视觉与模式识别会议。第4690–4699页（2019年）[10] Dou, Q.,Coelho de Castro,D.,Kamnitsas, K.,Glocker, B.: 通过语义特征的模型无关学习实现领域泛化。神经信息处理系统进展**32** （2019年）[11] 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——一种自监督学习的新方法。神经信息处理系统进展**33**，21271–21284（2020年）。

[12] Hadsell,R., Chopra, S.,LeCun,Y.：通过学习不变映射进行降维。载于：计算机视觉与模式识别会议。第 1735–1742 页（2006 年）

[13] He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.：用于无监督视觉表示学习的动量对比。在：计算机视觉与模式识别会议。第9729–9738页（2020）

[14] He,K.,Zhang,X.,Ren,S., Sun, J.：用于图像识别的深度残差学习。载于：计算机视觉与模式识别会议。第770–778页（2016年）

[15] Hu,W.,Niu, G.,Sato, I.,Sugiyama, M.：分布鲁棒监督学习能否产生鲁棒分类器？载于：国际机器学习会议。第 2029–2037 页（2018）

[16] Huang, G., Sun, Y., Liu,Z., Sedra, D.,Weinberger, K.Q.：具有随机深度的深度网络。载于：欧洲计算机视觉会议。第646–661页（2016）

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

[18] Kim, D., Yoo, Y., Park, S., Kim, J., Lee, J.：SelfReg：用于领域泛化的自监督对比正则化。载于：国际计算机视觉会议。第9619–9628页（2021年）

[19] Kingma, D.P., Ba, J.：Adam：一种随机优化方法。载于：国际学习表征会议（2015）

[20] Kobayashi, T.：用于类内特征分布正则化的 t-vMF 相似度。载于：计算机视觉与模式识别会议。第6612–6621页（2021年）

[21] Koh, P.W., Sagawa, S., Marklund, H., Xie, S.M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R.L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B., Haque, I., Beery, S.M., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., Liang, P.: Wilds：真实环境中分布偏移的基准。在：国际机器学习会议。第 5637–5664 页（2021年）[22] Komura, D., Ishikawa, S.: 组织病理图像分析的机器学习方法。《计算与结构生物技术杂志》**16**，34–42（2018）[23] Krueger, D., Caballero, E., Jacobsen, J.H., Zhang, A., Binas, J., Zhang, D., Le Priol, R., Courville, A.: 通过风险外推（REx）实现分布外泛化。在：国际机器学习会议。第 5815–5826 页（2021年）[24] Li, Y., Tian, X., Gong, M., Liu, Y., Liu, T., Zhang, K., Tao, D.: 通过条件不变量对抗网络实现深度领域泛化。在：欧洲计算机视觉会议。第 624–639 页（2018）

[25] Liu, E.Z., Haghgoo, B., Chen, A.S., Raghunathan, A., Koh, P.W., Sagawa, S., Liang, P., Finn, C.: 只训练两次：无需训练组信息即可提升群体鲁棒性。在：国际机器学习会议。第 6781–6792 页（2021年）[26] Liu, Z., Luo, P., Wang, X., Tang, X.: 野外人脸属性的深度学习。在：国际计算机视觉会议（2015年12月）[27] Loshchilov, I., Hutter, F.: SGDR：带热重启的随机梯度下降。在：国际学习表征会议（2017）[28] MMClassification 贡献者：OpenMMLab 的图像分类工具箱与基准。 https://github.com/open-mmlab/mmclassification （2020）[29] Motiian, S., Piccirilli, M., Adjeroh, D.A., Doretto, G.: 统一的深度监督式领域自适应与泛化。在：国际计算机视觉会议。第 5715–5725 页（2017）[30] Oord, A.v.d., Li, Y., Vinyals, O.: 基于对比预测编码的表示学习。arXiv 预印本 arXiv:1807.03748（2018）[31] Oren, Y., Sagawa, S., Hashimoto, T.B., Liang, P.: 分布鲁棒语言建模。在：自然语言处理实证方法会议暨国际自然语言处理联合会议。第 4227–4237 页（2019）[32] Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: 分布鲁棒神经网络。在：国际学习表征会议（2019）[33] Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet：用于人脸识别与聚类的统一嵌入。在：计算机视觉与模式识别会议。第 815–823 页（2015）[34] Sohn, K.: 基于多类 N 对损失目标改进深度度量学习。《神经信息处理系统进展》**29** （2016）[35] Sun, Y., Ming, Y., Zhu, X., Li, Y.: 利用深度近邻进行分布外检测。在：国际机器学习会议。第 20827–20840 页（2022年）[36] Vapnik, V.N.: 统计学习理论综述。《IEEE 神经网络汇刊》10(5)，988–999（1999）[37] Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: 使用带噪学生进行自训练可提升 ImageNet 分类性能。在：计算机视觉与模式识别会议。第 10687–10698 页（2020）[38] Zhang, M., Sohoni, N.S., Zhang, H.R., Finn, C., Re, C.: Correct-n-contrast：一种通过对比提升对虚假相关性鲁棒性的方法。在：国际机器学习会议。第 26484–26516 页（2022年）

[39] Zhou, C.,Ma,X.,Michel,P., Neubig, G.：分布偏移下伪特征的研究与对抗。载于：国际机器学习会议。第12857– 12867页（2021年）[40] Zhou, K., Liu, Z., Qiao,Y., Xiang, T.,Loy, C.C.：领域泛化：综述。IEEE模式分析与机器智能汇刊（2022年）

A 使用 CelebA 数据集的额外实验结果。

A.1 金发/性别设置

表 5 展示了主稿第 5.3 节中讨论的使用 CelebA 数据集进行的实验的总体准确率。

表 5：CelebA 数据集测试集的总体准确率（金发/性别设置）。

	固定学习率		采用余弦衰减的学习率	
	鲁棒分布鲁棒优化	CVaR-DRO	鲁棒分布鲁棒优化	CVaR-DRO
DRO损失	95.10 (±0.16)	94.49 (±0.47)	95.28 (±0.07)	94.36 (±0.12)
+权重衰减(1e-4)	93.53 (±0.79)	89.46 (±5.13)	94.76 (±0.05)	93.68 (±0.11)
+权重衰减(1e-3)	92.55 (±2.08)	80.35 (±11.12)	93.91 (±0.17)	92.38 (±0.18)
+权重衰减(1e-2)	88.80 (±3.95)	68.56 (±10.87)	93.00 (±0.37)	91.34 (±0.31)
+标准差(0.1)	94.93 (±0.16)	94.22 (±0.52)	95.13（±0.04）	94.29 (±0.09)
+标准差(0.3)	94.51 (±0.36)	93.70（±0.68）	94.56（±0.10）	93.65（±0.09）
+标准差（0.5）	93.80（±0.60）	93.36（±0.54）	94.07 (±0.17)	93.00 (±0.28)
+RandAug	93.99 (±0.25)	92.06 (±0.75)	94.20 (±0.13)	92.11 (±0.18)
+监督对比学习（余弦）	94.08 (±0.54)	94.60 (±0.52)	95.18 (±0.06)	94.90 (±0.19)
+所提方法(0.05)	92.58（±0.86）	92.57 (±0.29)	94.14 (±0.18)	93.86 (±0.25)
+所提方法(0.1)	92.13 (±0.43)	91.57 (±0.78)	93.23 (±0.72)	92.41 (±0.41)
+所提方法(0.15)	90.57（±1.46）	89.50 (±1.78)	91.49 (±0.53)	91.35 (±0.89)
+所提方法(0.2)	89.55 (±2.37)	88.86（±2.29）	91.72 (±0.93)	90.93 (±0.61)

A.2 微笑/嘴巴微微张开设置

除了主论文中所解释的金发/性别设置外，我们还使用 CelebA 数据集进行了另一项实验。我们考虑预测图像中的人物是否在微笑这一问题，并将“嘴巴微张”这一属性视为混杂变量，我们将其称为微笑/嘴巴微张设置。每张训练图像属于由微笑和嘴巴微张情况定义的四个组别之一，其列联表见表 6。

图 9展示了三个模型的测试准确率曲线：一个使用群体DRO 损失训练的模型、一个使用群体 DRO 损失和采用余弦相似度的SupCon 损失训练的模型，以及一个使用群体 DRO损失和所提方法损失（其中包含 $\alpha = 0.1$）训练的模型，其中实线和虚线分别表示最差群体准确率和总体准确率。我们可以看到，当模型仅使用 DRO损失进行训练时，随着训练的进行，最差群体准确率会下降。然而，与主文稿中报告的金发/性别设置相比，其下降速率较为缓和，这可能是因为在微笑/嘴巴微张设置中不存在极端少数群体。

表 6：CelebA 数据集关于两个属性的列联表：微笑和嘴巴微张。

	未打开	Open
不微笑	81339	23591
微笑	23318	74351

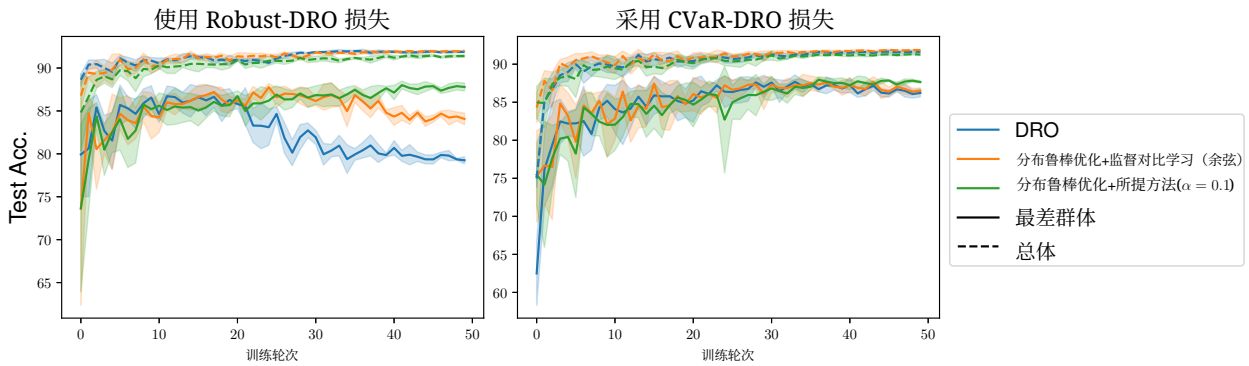


图 9：CelebA（微笑/嘴巴微张设置）中最差群体（实线）和总体（虚线）的测试准确率。

我们将所提方法与其他 DNN 正则化方法进行了比较。表 7 和表 8 分别汇总了最差群体准确率和总体准确率。可以看到，与其他正则化方法相比，所提方法取得了更高或相当的最差群体准确率。

B 使用 iWildCam 数据集的实验

除在 Camelyon17 数据集上进行实验外，我们还在 iWildCam2020-WILDS（v2.0）数据集上进行了实验 [21], 为简洁起见，以下简称 iWildCam，以评估所提方法在领域泛化情境下的表现。iWildCam 数据集由安装在野外、用于监测动物种群的相机陷阱拍摄的照片组成。每张照片都被标记为 182 个类别中的一个，用于表示动物物种，其中包括表示照片中没有动物的标签。每个相机陷阱的位置被视为一个领域，iWildCam 数据集包含来自 323 个不同领域的照片。iWildCam 数据集包含五个划分[9] 包括 $\bar{\mathcal{D}}_{\text{train}}$、$\bar{\mathcal{D}}_{\text{val}}^{\text{ID}}$、$\bar{\mathcal{D}}_{\text{test}}^{\text{ID}}$、$\bar{\mathcal{D}}_{\text{test}}^{\text{OOD1}}$和 $\bar{\mathcal{D}}_{\text{test}}^{\text{OOD2}}$。$\bar{\mathcal{D}}_{\text{train}}$、$\bar{\mathcal{D}}_{\text{val}}^{\text{ID}}$和 $\bar{\mathcal{D}}_{\text{test}}^{\text{ID}}$ 的领域相同，而 $\bar{\mathcal{D}}_{\text{test}}^{\text{OOD1}}, \bar{\mathcal{D}}_{\text{test}}^{\text{OOD2}}$的数据来自不同的领域。

我们使用 $\bar{\mathcal{D}}_{\text{train}}$训练了一个模型。除了训练轮次为 20 轮外，模型和训练设置均与 Camelyon17 数据集相同。图10展示了在每个$\bar{\mathcal{D}}_{\text{test}}^{\text{OOD1}}$、$\bar{\mathcal{D}}_{\text{test}}^{\text{OOD2}}$和 $\bar{\mathcal{D}}_{\text{test}}^{\text{ID}}$上训练期间的测试准确率曲线。与

[9]这些划分在原始数据中分别命名为训练集、id 验证集、id 测试集、验证集和测试集。

表 7：CelebA（微笑/嘴巴微张设置）中最差群体的测试准确率。

	固定学习率		衰减学习率		
	鲁棒分布鲁棒优化	CVaR-DRO	鲁棒分布鲁棒优化	CVaR-DRO	平均值
DRO损失	79.74 (±1.31)	87.96 (±0.31)	79.25(±0.41)	86.22 (±0.54)	83.29
+ 权重衰减(1e-4)	86.65 (±0.90)	83.78 (±3.36)	83.97(±0.52)	87.93 (±0.25)	85.58
+ 权重衰减(1e-3)	83.01 (±4.56)	81.20 (±6.12)	88.16(±0.29)	87.31 (±0.25)	84.92
+ 权重衰减(1e-2)	43.89(±33.69)	56.37 (±17.30)	87.67 (±0.43)	68.76 (±34.38)	64.17
+ 标准差(0.1)	79.90 (±2.00)	85.09 (±1.31)	82.90(±0.52)	87.32 (±0.53)	83.80
+ 标准差(0.3)	84.78 (±1.83)	87.08 (±1.02)	85.25(±0.53)	87.61 (±0.35)	86.18
+ 标准差(0.5)	86.96（±1.84）	87.43（±0.94）	86.59(±0.57)	88.39 (±0.22)	87.34
+ RandAug	86.88 (±1.25)	34.03(±31.60)	88.37 (±0.24)	64.99 (±1.92)	68.57
+ 监督对比学习（余弦）	84.39 (±0.84)	86.14 (±1.49)	84.07(±0.69)	86.49 (±0.39)	85.27
+所提方法(0.05)	87.38 (±1.21)	86.49 (±0.95)	88.09（±0.31）	88.22（±0.34）	87.55
+ 所提方法(0.1)	84.84 (±1.57)	86.93 (±0.62)	87.79(±0.47)	87.66 (±0.15)	86.81
+所提方法(0.15)	86.03 (±0.69)	84.74 (±2.11)	87.07(±0.14)	87.42 (±0.46)	86.32
+ 所提方法(0.2)	85.10 (±1.48)	83.01 (±3.02)	86.12(±0.63)	86.26 (±0.69)	85.12

表 8：CelebA 数据集测试集的总体准确率（微笑/嘴巴略微张开的设置）。

	固定学习率		采用余弦衰减的学习率	
	鲁棒分布鲁棒优化	CVaR-DRO	鲁棒分布鲁棒优化	CVaR-DRO
DRO损失	91.74 (±0.13)	91.12 (±0.21)	91.88 (±0.07)	91.63 (±0.13)
+权重衰减(1e-4)	90.80 (±0.23)	89.78 (±1.18)	90.72（±0.07）	91.03（±0.08）
+权重衰减(1e-3)	89.90 (±1.91)	88.45 (±2.52)	91.34 (±0.14)	90.89 (±0.06)
+权重衰减(1e-2)	73.62 (±14.37)	80.77 (±5.82)	90.88 (±0.21)	81.93 (±15.98)
+标准差(0.1)	91.70 (±0.18)	90.82 (±0.56)	92.21 (±0.07)	91.56 (±0.12)
+标准差(0.3)	90.96 (±0.23)	91.18 (±0.40)	91.28 (±0.11)	91.71 (±0.07)
+标准差(0.5)	91.72 (±0.23)	91.25 (±0.25)	92.12 (±0.05)	91.65 (±0.05)
+RandAug	91.81 (±0.27)	60.62 (±15.32)	91.81 (±0.07)	68.73 (±4.82)
+监督对比学习（余弦）	91.71 (±0.13)	91.23 (±0.45)	92.00 (±0.08)	91.85（±0.11）
+所提方法（0.05）	91.15（±0.30）	91.26 (±0.37)	91.71 (±0.06)	91.61 (±0.08)
+所提方法(0.1)	90.91 (±0.70)	90.84 (±0.23)	91.40 (±0.15)	91.27 (±0.09)
+所提方法(0.15)	90.21 (±0.38)	90.21 (±0.79)	90.88 (±0.17)	91.07 (±0.31)
+所提方法(0.2)	89.80 (±0.63)	88.97 (±1.40)	90.39 (±0.57)	90.34 (±0.80)

与 Camelyon17 数据集的结果不同，iWildCam 数据集上的 OOD 划分性能并未下降。原因尚不明确，但如正文所述，与 Camelyon17 相比，iWildCam 的训练集划分包含来自许多不同领域的样本，这或许避免了过拟合。表 9 显示了训练结束时各测试集划分上的准确率。可以看到，与使用交叉熵进行训练相比，所提方法在 OOD 划分上取得了略好的性能。

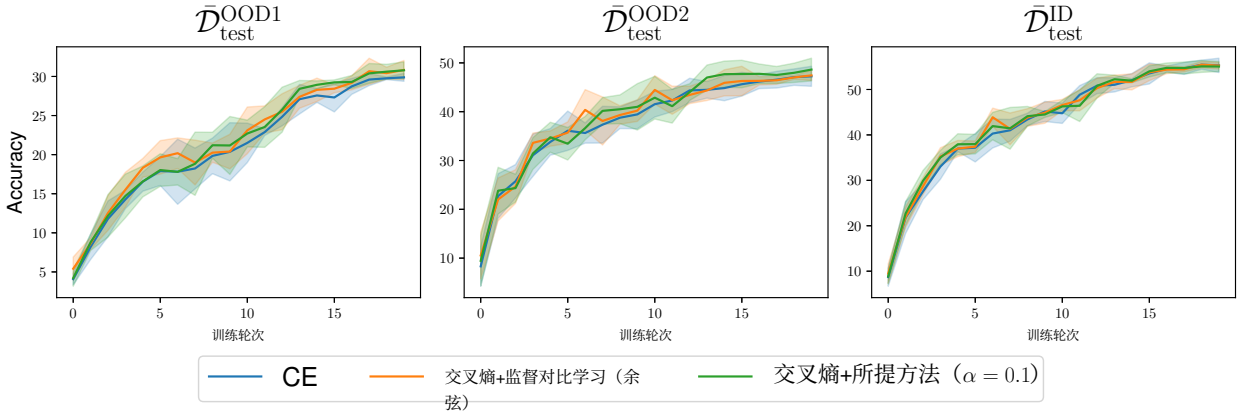


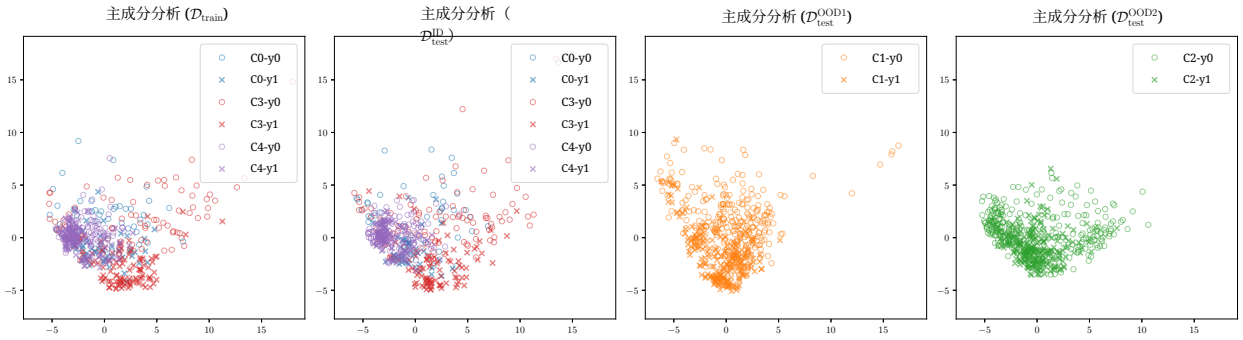
图10：iWildCam 数据集测试集划分上的准确率。

表9：iWildCam 数据集测试集上的准确率。

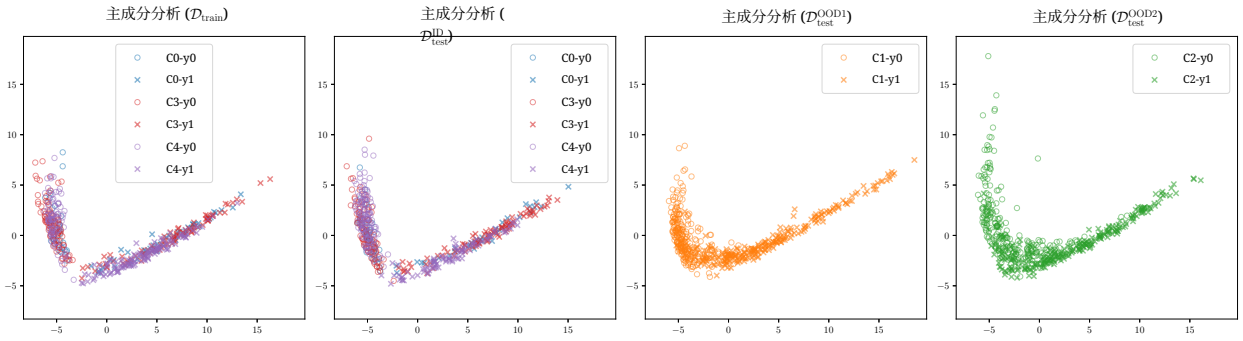
	$\bar{\mathcal{D}}_{\text{test}}^{\text{OOD1}}$	$\bar{\mathcal{D}}_{\text{test}}^{\text{OOD2}}$	$\bar{\mathcal{D}}_{\text{test}}^{\text{ID}}$
CE	29.89（±0.51）	47.25（±2.05）	55.31（±1.58）
交叉熵+监督对比学习（余弦）	30.83 (±1.22)	47.44 (±1.04)	55.25 (±0.74)
交叉熵+所提方法($\alpha = 0.1$)	30.79 (±1.07)	48.63 (±2.33)	55.05（±1.10）

C 使用 Camelyon17 数据集的其他实验结果。

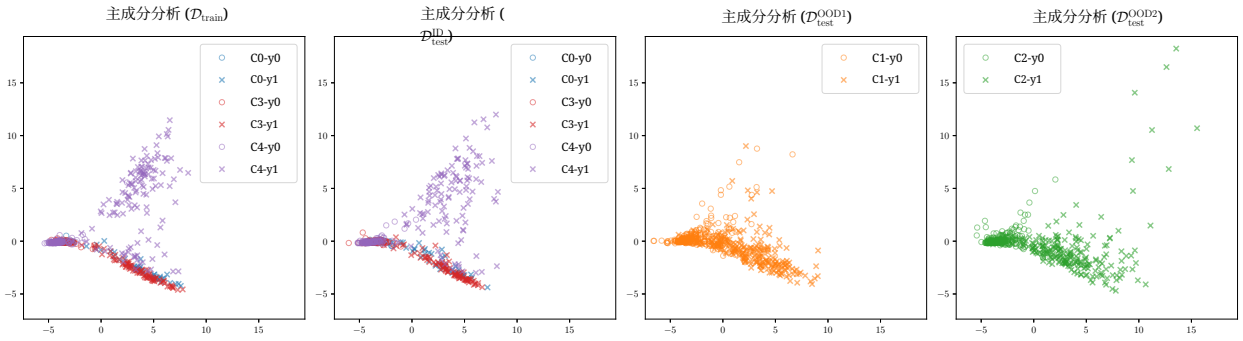
图11展示了 Camelyon17 数据集中 $\mathcal{D}_{\text{train}}$、$\mathcal{D}_{\text{test}}^{\text{ID}}$、$\mathcal{D}_{\text{test}}^{\text{OOD1}}$和 $\mathcal{D}_{\text{test}}^{\text{OOD2}}$样本嵌入的主成分分析可视化结果。嵌入分别由四种模型得到：初始模型、使用交叉熵训练的模型、使用交叉熵和余弦相似度监督对比学习训练的模型，以及使用交叉熵和所提损失函数（ $\alpha = 0.1$）训练的模型。点的颜色表示样本来源医院，“o”和“x”标记分别表示 $y = 0$ （无肿瘤）和 $y = 1$ （肿瘤）。



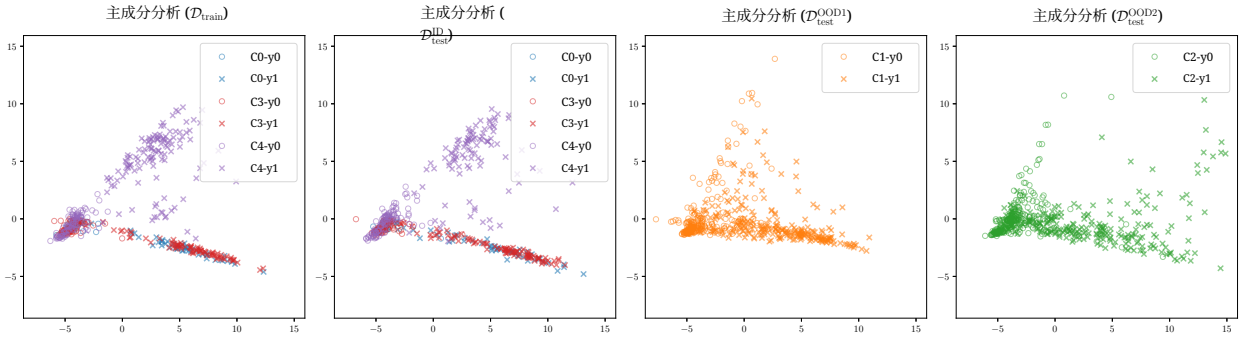
(a) 初始模型。



(b) 使用交叉熵训练的模型。



(c) 使用交叉熵+监督对比学习（余弦）训练的模型。



(d) 使用交叉熵+所提方法($\alpha = 0.1$)训练的模型。

图11：Camelyon17 数据集的嵌入的主成分分析可视化。图例中的“C”和“y”分别表示中心和标签。