arXiv:2305.00650v2 [cs.LG] 5 Jun 2023

# 发现 及其修复：概念感知的虚假相关性缓解

**Shirley Wu** [1] **Mert Yuksekgonul** [1] **Linjun Zhang** [2] **James Zou** [1]


## 摘要

深度神经网络往往依赖伪相关性进行预测，这会阻碍模型在训练环境之外的泛化。例如，将猫与床背景关联起来的模型，在没有床的其他环境中可能无法预测猫的存在。缓解伪相关性对于构建可信赖的模型至关重要。然而，现有研究缺乏透明度，无法为缓解过程提供洞见。在这项工作中，我们提出了一个可解释框架——发现与治愈（DISC），以解决这一问题。借助人类可解释概念，DISC 通过迭代执行以下步骤：1）发现不同环境中不稳定的概念，并将其视为伪属性；2）利用发现的概念干预训练数据，以减少伪相关性。在一系列系统性实验中，与现有方法相比，DISC 展现出更强的泛化能力和可解释性。具体而言，在物体识别任务和皮肤病变分类任务上，其表现分别比最先进的方法高出 7.5% 和 9.6%。此外，我们还提供了理论分析和保证，以阐明由 DISC 训练的模型所具备的优势。代码和数据可在 https://github.com/Wuyxin/DISC.


## 1. 引言

伪相关在现实世界的数据分析中十分常见。伪属性通常与类别标签相关联，但无法泛化（Kaushik 等人，2020 年；Sagawa 等人，2020 年）。例如，如图 1 所示，错误地将猫与床关联起来的神经网络容易在不同场景下失效，例如，狗在床上或猫在书桌上，此时伪相关不再

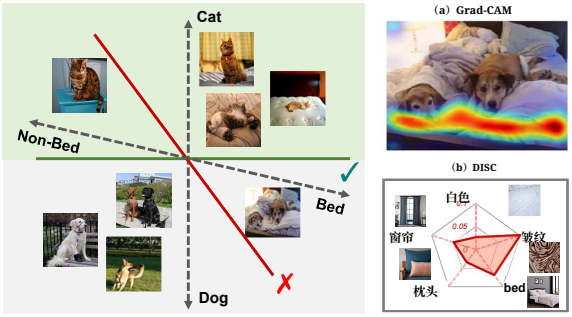



图 *1.* 左：依赖（红色）或不依赖（绿色）伪相关的狗/猫分类器；右：Grad-CAM（Selvaraju 等人，2017）和 DISC 的伪相关发现结果，我们提出了一种类别级指标，用于表示概念与“猫”类别之间伪相关的程度。

成立。这种可靠性不足是关键应用中的一个核心问题，例如，医学诊断（Bissoto 等人，2020 年）。

现有研究已经提出了减轻深度模型中伪相关性的方法。例如，不变学习（Arjovsky 等人，2019；Ahuja 等人，2021年）学习跨环境稳定的表示，以避免受包括伪属性在内的变化因素影响。利用经验风险最小化（ERM）模型容易受到伪属性影响这一弱点，一些研究会对训练好的 ERM 模型中置信度过高（Nam 等人，2020 年）或分类错误（Liu 等人，2021a）的样本赋予更高权重，以抵消伪相关性。

然而，这些研究无法解释模型究竟学到了哪些信息，又忽略了哪些信息，这妨碍了人类理解和模型审计。尽管事后可解释性方法（Selvaraju 等人，2017；Ribeiro 等人，2016；Lundberg 和 Lee，2017）提供的可视化可能包含伪相关区域，但仍难以理解。例如，在图 1（a）中，高亮区域显示了对预测贡献最大的属性，解释了为什么图像被错误地预测为“猫”而非“狗”。然而，哪些属性（例如白色、皱纹纹理或枕头等物品）对伪相关的贡献最大，并不明确。此外，这类实例级解释无法说明类别中整体存在的伪相关性。

在这项工作中，我们采用符合人类理解的概念来发现类别级伪相关属性，并借助概念库作为辅助知识库，

[1]斯坦福大学计算机科学系。[2]罗格斯大学统计学系。通讯作者：Shirley Wu <shirwu@cs.stanford.edu>、James Zou<jamesz@stanford.edu>。

论文集，发表于 $40^{th}$ 国际机器学习会议，美国夏威夷州檀香山。PMLR 202，2023年。版权所有 © 2023，作者所有。

以例如，猫的形状等不变概念在不同数据环境中保持稳定，而例如，“床”等伪概念在同一类别的不同实例中并非始终存在。受此特性的启发，我们提出一种类别级指标，概念敏感度，用于量化概念在不同数据环境中的不稳定性。例如，在图1（b）中，我们根据概念敏感度的较大幅度，将“皱纹”和“床”都识别为“猫”的高度伪相关概念，并将此步骤称为**发现步骤**。

发现伪概念后，我们提出一个干预步骤，即 概念感知干预，以降低模型对伪概念的依赖。其核心思路是对含有伪概念的选定类别进行干预，从而维持伪概念的均衡分布。例如，如图 2 右下角所示，在识别出“皱纹”和“床”是与“猫”类别相关的伪概念后，我们使用它们的概念图像对“狗”类别进行干预。通过使伪概念在不同类别之间均衡分布，我们可以防止模型利用伪概念进行预测，从而消除伪偏差。我们将这一过程称为**治愈步骤**。

**发现与治愈**。最后，我们的算法 DISC 在训练过程中迭代执行发现和治愈步骤。每次迭代中，它都会发现当前模型的伪概念。随后，基于发现的概念，它对训练数据集进行干预，以消除虚假相关性，并据此更新模型。本文聚焦于图像分类任务。实验表明，DISC 发现的伪概念与真实伪属性相符，且其平均表现大幅优于最先进的基线方法。我们的**贡献**如下：

•我们提出了一种新颖且可解释的框架，用于发现伪概念，并有效缓解影响模型泛化的虚假相关性。

• 我们通过实验证实了我们的方法在多样化数据集上的有效性，并揭示了模型如何克服伪相关性的见解。

• 我们从理论上保证了通过 DISC 训练的模型的收敛性和泛化能力。

## 2. 相关工作

我们的工作以人类可解释概念为基础，涉及发现并消除伪相关性。在这里，我们讨论三类相关工作：

**概念**。概念，例如蓝色或条纹，是人类可理解的语义。概念已被用于构建可解释模型（Lampert 等人，2009；Kumar 等人，2009；Koh 等人，2020 年；Y¨uksekg¨on¨ul 等人，2022），也被用于事后方式（Bau 等人，2017；Kim 等人，2018）来解释深度神经网络（DNN）的预测。具体而言，Kim 等人（2018）提出了概念激活向量（CAVs），其中 CAV 表示 DNN 隐空间中对应于某个概念存在的方向，有助于使 DNN 的内部状态与人类预期相一致。

发现虚假相关性。以往研究考察了图像纹理和背景（Geirhos 等人，2019；Sagawa 等人，2020）、领域偏移（Koh 等人，2021；Gulrajani 和 Lopez-Paz，2021；Santurkar 等人，2021；Ye 等人，2023）以及因果不稳定属性（Arjovsky 等人，2019；Wu 等人，2022）等情境中的虚假相关性。检测虚假相关性可以揭示有害于泛化的模型偏差。一些研究利用领域知识获取伪属性（Clark 等人，2019；Kaushik 等人，2020；Nauta 等人，2021），但伪属性的范围可能超出领域知识。例如，Creager 等人（2021）通过学习环境推断伪属性。Sohoni 等人（2020）；Seo 等人（2022）对模型嵌入进行聚类，并利用这些簇揭示伪属性。

近期研究（Plumb 等人，2021年；Hagos 等人，2022；Abid 等，2022）也使用可解释性技术来发现虚假属性，并需要人工检查。与实例级审计不同，我们提出了一种类别级指标，可提供更可靠、易于使用的高层次可解释性。此外，概念级和交互式调试方法（Bontempelli 等人，2022；Bahadori 和 Heckerman，2021；Teso 和 Kersting，2019）利用概念或人工反馈进行调试。综述请参见 Teso 等人（2023）。例如，Bontempelli 等人（2022）提出了 ProtoPDebug，使人工监督者能够就模型的解释向部件原型网络（Chen 等人，2019）（ProtoPNets）提供反馈。与我们的方法不同，这些方法通常适用于类别受限的模型（例如，概念瓶颈模型（Koh 等人，2020）或 ProtoPNets），并且往往需要人工标注来识别概念。所选研究与我们的方法之间的比较见表6。

**消除虚假偏差**。学习伪属性会使模型对伪相关因素及其分布偏移过于敏感，这与不变学习和鲁棒学习有关。

• **不变学习**。 Arjovsky 等人（2019）提出学习一种不变编码器，使下游分类器在不同环境中均达到最优。其他工作则通过相关性对齐（Sun&Saenko，2016）、方差惩罚（Krueger 等人， 2021年；Teney 等人， 2020 年）和跨领域的梯度对齐（Shi 等人， 2021年）来实现不变性。然而，这些方法缺乏可解释性，因此难以深入了解数据偏差。

• **实例重加权**。实例重加权会提高不太可能包含伪相关属性的示例的重要性，以消除偏差（Yaghoobzadeh 等人，2021年；Utama 等人，2020 年；Dagaev 等人，2021年；Zhang 等人，2022b；Nam

等人，2020 年；Li 等人，2022）。尽管这种方法很简单，但当模型完美拟合训练数据时，此类实例可能十分罕见，从而限制其有效性。分布鲁棒优化（DRO）（Ben-Tal 等人，2013；Oren 等人，2019；Sagawa 等人，2020 年；Zhang 等人，2021a）是其中一种特殊情况，它会提高高损失观测的权重（Namkoong &Duchi，2016；Hu 等人，2018；Levy 等人，2020 年）。然而，实例重加权对过参数化 DNN 的影响可能会随着训练轮次的增加而减弱（Byrd& Lipton，2019），最终导致过拟合。

• **数据增强**。其他研究采用对抗性混合（Xu 等人，2020 年）、选择性增强（Yao 等人，2022）和不确定性感知混合（Han 等人）等数据增强方法，以降低对伪相关的依赖（Zhang 等人，2021b）。此外，Pinto 等人(2022) 提出，mixup（Zhang 等人，2018）作为正则化方法，能够进一步提升分布外鲁棒性。然而，这些数据增强方法并未明确考虑多个共存的伪属性，而这在现实世界应用中很常见。DISC 借助文生图生成器生成的概念库检测伪概念，并自适应地将概念图像与选定类别中的实例进行混合。同期研究（Jain 等人，2022）使用图像描述模型捕捉失效模式，并生成合成图像用于微调。然而，生成的数据依赖于图像描述模型；对于皮肤病变图像这类难以描述的数据集，该模型可能无法适用，因而不可行。DISC 则是一种更灵活的解决方案，通过概念图像进行干预。

## 3. 方法

在此，我们介绍问题设定和我们的方法。我们在第 3.1 节对问题设定进行形式化，并在第3.2节介绍概念库的构建。随后，我们在第 3.3节讨论伪概念的发现，并在第3.4节介绍伪相关的消除。为清晰起见，我们在附录 A中总结了主要符号。

### 3.1. 问题形式化

我们考虑一个监督式图像分类问题。具体而言，给定一个训练数据集$\mathcal{D}_{tr}=\{(x_1,\ y_1)...,\ (x_n,\ y_n)\}$。我们将$\mathcal{Y}$定义为标签空间，将 $\mathcal{P}_{tr}$ 定义为训练数据集的分布。

对于任意损失函数 $\ell$，经验风险最小化（ERM）会针对模型 $f$最小化经验损失：

$$\underset{\phi}{\operatorname{argmin}}\mathbb{E}_{(x,y)\sim\mathcal{D}_{tr}}[\ell(f_\phi(x),y)] \tag{1}$$

其中，$f$由 $\phi$参数化。由于伪属性具有不稳定性，测试分布 $\mathcal{P}_{te}$ 通常不同于训练分布，即 $\mathcal{P}_{te}\neq\mathcal{P}_{tr}$。因此，使用 ERM 训练的模型难以泛化到数据集 $\mathcal{D}_{te}\sim\mathcal{P}_{te}$ 中伪相关发生变化或不再成立。因此，我们的目标是在存在伪相关的情况下克服模型的偏差。从因果关系的角度来看，伪属性被定义为与真实标签 $Y$ 不存在因果关系、但由于数据采样偏差或不平衡而在训练数据中与真实标签 $Y$相关的属性$F$。例如，“床”不能决定图像是否被标记为“猫”，但如果猫图像大多是在卧室中拍摄的，那么“床”可能与“猫”这一标签相关。就我们的目的而言，我们将伪属性视为在某些环境中其存在与标签相关、而在其他环境中则不相关的属性。

### 3.2. 概念库

为描述伪属性，我们以人类可理解的方式将其视为概念，而不是像素级模式。我们构建了一个全面的概念库，广泛涵盖潜在的伪相关概念候选项。形式上，我们有

**定义 1 (概念库)**。一个包含 $m$个概念的概念库可表示为 $\mathcal{C}:=\{(c_i,\mathcal{P}_{c_i})\mid i=1,...,m\}$，其中每个 $c_i$ 表示一个概念，$\mathcal{P}_{c_i}$ 是包含该概念的图像的分布。

我们在图 7 （附录 D）中展示了示例，其中我们使用文生图生成模型，例如，Stable Diffusion（Rombach 等人，2022）生成代表 $\mathcal{P}_{c_i}$ 的概念图像，并将概念名称用作提示词。

此外，对可解释性的需求要求将概念与深度模型的内部状态对齐。为不失一般性，我们将深度模型记为 $f=h\circ g$，其中 $g$是编码器，而 $h:\ \mathbb{R}^d\to\mathbb{R}^{|\mathcal{Y}|}$是线性层。因此，给定模型 $f$和概念 $c_i$，我们定义一个查询操作，以提取高维概念表示 $v_i$。具体而言，我们通过从 $\mathcal{P}_{c_i}$中采样 $N^p$图像来构建正样本集 $I_i^p$，并通过从 $\{\mathcal{P}_{c_j}\mid j\neq i\}$中随机采样 $N^n$图像来构建负样本集 $I_i^n$。遵循 Kim 等人（2018），我们学习一个线性 SVM，在隐空间 $\mathbb{R}^d$ 中寻找一个超平面，以尽可能区分正图像和负图像。然后，我们计算与该超平面正交的向量 $v_i$，将其作为概念激活向量（CAV）。直观地说，CAV 是隐空间中表示某个概念存在的方向。因此，概念库充当辅助知识库，用于在后续步骤中发现并缓解伪相关。

### 3.3. 关于发现伪概念

借助概念库，我们旨在从概念候选项中识别伪概念。

**关于训练分布的假设**。为保证能够区分伪概念，我们假设训练分布能够代表整体分布。因此

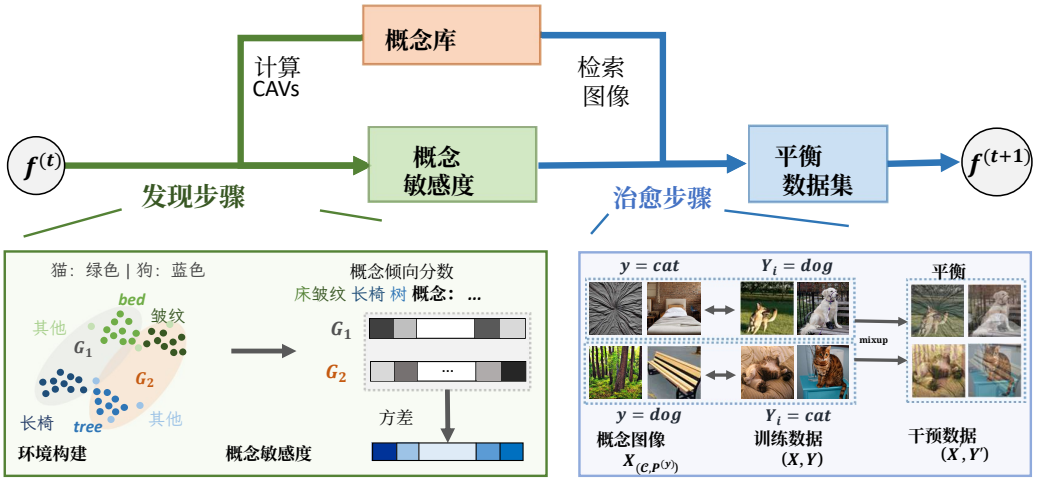



图 2. **DISC 框架**。 在第 $t$ 次迭代中，DISC 根据先前构建的环境和概念库中的 CAV 计算概念敏感度，从而发现“猫”的伪概念为（皱纹、床），而“狗”的伪概念为（长椅、树）。在治愈步骤中，DISC 根据概念敏感度从概念库中检索概念图像，例如“床”，并将其与训练子集（例如 狗图像）混合，以消除伪相关。最后，在平衡数据集上更新模型。

例如，如果“床”在训练数据集中总是与“猫”共存，那么我们就无法区分“猫”是不变概念，而“床”不是。因此，训练分布应当反映整体分布的本质特征，这一点也被先前的工作隐式假设了（Liu 等人，2021a；Nam 等人，2020 年）。我们在定理 1中将这一假设形式化。在代表性假设下，我们进一步提出以下关于伪概念不恒定性的一项观察：

**观察 1(不一致性属性)**。 伪概念往往存在于数据的异质子集中，并且它们与标签之间的相关性也因子集而异。

例如，“床”并非所有“猫”实例都具有的共同特征，它可能只出现在特定的猫图像子集中，而且床的存在与猫标签之间的相关程度也会因不同子集而异。我们希望利用这一属性和洞见来识别伪概念。具体而言，由于有偏模型会将标签与伪概念相关联，伪概念的分布会对模型的决策边界产生重大影响。更重要的是，这种影响在不同环境中往往并不一致。例如，假设我们根据“猫”实例是室内照片还是室外照片，将其聚成不同的簇。在这两种情形下训练的模型，在利用“床”这一概念进行预测时会表现出不同的偏好。因此，伪概念的作用在不同环境中极其脆弱且敏感，伪概念的分布可能发生显著变化。相比之下，不变概念，例如动物形状，在任何数据环境下都更加一致、同质。

在这一性质的启发下，我们提出一种类别级指标——**概念敏感度**，用于判断某个概念是否与特定类别存在伪相关。下面介绍其计算步骤：

**步骤 1（聚类）**。我们首先尝试对训练数据集进行合理分层，以构建数据环境。正如先前研究（Sohoni 等人，2020 年；Seo 等人，2022）所探讨的，即便是有偏模型，也能很好地区分不同特征。因此，我们利用通过 ERM 训练的模型，为训练实例生成表示和预测；与Eyuboglu 等人（2022）的方法类似，该方法通过纳入模型预测来考虑错误类型。基于生成的向量，我们对每个类别中的实例进行一次聚类。对于每个类别 $y \in \mathcal{Y}$，我们得到 $G^{(y)} = \{G_j^{(y)}\}_{j=1}^{k}$，其中 $G_j^{(y)}$ 是类别 $j$中的第 $y$ 个簇， $k$ 是簇的数量。最后，我们将环境构建为

$$G_j = \bigcup_{y=1}^{|\mathcal{Y}|} G_j^{(y)},\ j = 1,\ldots,k, \tag{2}$$

其中，每个类别的簇组合旨在避免各个环境中遗漏标签。需要注意的是，组合可能因簇索引的顺序而异，因此我们随机打乱簇索引，以更新环境并实现更鲁棒的训练。

**步骤 2（计算概念敏感度）**。对于每个 $G_j$，我们将环境梯度矩阵（EGM）定义为

$$M_j = \nabla_\omega [\mathbb{E}_{(x,y)\sim G_j} \ell(f(x), y)] = \frac{\partial [\mathbb{E}_{(x,y)\sim G_j} \ell(f(x), y)]}{\partial \omega} \tag{3}$$

其中， $\omega \in \mathbb{R}^{|\mathcal{Y}| \times d}$ 是 $h$的参数。直观而言， $M_j$表示仅根据 $G_j$ 中的观测所产生的参数流形变化。虽然使用所有训练实例计算公式3 代价高昂，但我们采样小批量数据来近似 $M_j$。此外，为了使当前模型与概念空间对齐，我们查询概念库，提取概念 $c_i$的概念表示 $v_i$。然后，我们

计算 $v_i \cdot M_j^T \in \mathbb{R}^{|\mathcal{Y}|}$，它表示在环境 $G_j$ 下，各类别对概念 $c_i$ 的偏好程度。例如，如果在环境 $G_j$ 中，“床”与“猫”高度相关，那么 $M_j$ 所反映的更新后决策边界将与“床”这一概念的 CAV 对齐，其输出 logits 为“猫”。此外，我们进一步将概念 $c_i$ 的概念敏感度 $S_i$ 定义为

$$S_i = \text{Var}(\{(v_i \cdot M_j^T)_{y_i'} \mid j=1,...,k\}), \tag{4}$$
$$\text{where } y_i' = \underset{y}{\text{argmax}} \left(\sum_{j=1}^{k} v_i \cdot M_j^T\right)_y$$

这里 $y_i'$ 是由概念主导或与概念强相关的类别 $c_i$。Var 表示方差算子。为方便起见，我们将 $(v_i \cdot M_j^T)^{y_i'}$ 称为概念 $c_i$ 在环境 $G_j$ 下的概念倾向分数（CTS）。因此，概念敏感度本质上评估了 CTS 在不同环境中的不一致性。CTS 的方差较大，表明该概念不稳定，且其对最终预测的贡献会因环境不同而变化。由于因果概念在不同环境中应表现出不变性（Arjovsky 等人，2019 年；B¨uhlmann，2020 年），较高的概念敏感度可以被视为证据，表明该概念在模型训练中是伪相关且具有误导性的。在我们之前的例子中，“床”与“猫”类别相关，“床”在不同环境中的 CTS 方差较大，因为在训练数据中，“床”与“猫”并非始终同时出现。另请注意，概念敏感度是按类别计算的，这为理解特定类别中的虚假相关性提供了高层次可解释性。

### 3.4. 通过概念感知干预消除偏差

具体而言，我们认为，类别间伪概念分布不均衡导致了模型偏差。例如，在 Waterbirds（Sagawa 等人，2020 年）中，陆鸟类别中有 95% 的样本具有陆地背景，而水鸟类别中只有 4% 的样本涉及陆地。因此，这种极端的不平衡会促使模型利用伪相关性作为捷径来进行预测。

然而，直接从训练数据集中移除伪属性可能会引入更多噪声，并导致模型过拟合（Khani 和 Liang，2021年）。相反，我们通过数据增强来维持不同类别中伪概念分布的均衡，从而消除这种相关性。

**步骤 3（概念感知干预）**。我们将每个$H^{(y)} \in \mathbb{R}^m$ 表示为一个布尔向量，其中当 $y_i' = y$时为 $H_i^{(y)} = 1$，否则为 $H_i^{(y)} = 0$。然后，我们通过对掩蔽后的概念敏感度进行归一化，计算每个类别上的概念概率，即 $P^{(y)} = S \cdot H^{(y)} / \sum[S \cdot H^{(y)}]$。直观而言，概念概率回答了“与类别 $y$相关的概念有哪些”以及“它们的伪相关性有多强”这两个问题。为了保持伪概念的平衡，我们按概率 $P^{(y)}$采样概念图像，

**算法1**DISC 伪代码

**要求：** 训练数据 $\mathcal{D}$、概念库 $\mathcal{C}$、模型 $f = h \circ g$，学习率 $\alpha$、参数 $\beta_1$，$\beta_2$ 贝塔分布的参数
1：通过对图像嵌入进行聚类获取 $\{\{G_j^{(y)}\}_{j=1}^k\}_{y=1}^{|\mathcal{Y}|}$
2：**当**未收敛时 **执行**
3：随机化簇索引并获得$\{G_j\}_{j=1}^k$（公式 2）
4：$\{P^{(y)}\}_{y=1}^{|\mathcal{Y}|} \leftarrow$概念敏感度($G$，$f$，$\mathcal{C}$)
5：**对**每个类别 $y$ **执行**
6：采样小批量（$X$，$Y$），其中每个 $Y_i \neq y$
7：采样概念图像 $X_{(\mathcal{C},P^{(y)})}$，概率为 $P^{(y)}$
8：执行 mixup 以获得($X'$,$Y'$) (公式 5)
9：$\phi \leftarrow \phi - \alpha \cdot \partial \ell[\mathbb{E}(f(X')\ Y')]/\partial\phi$
10：
11: **函数** 概念敏感度($G$,$f$,$\mathcal{C}$) 12: 查询 CAV 矩阵 $V = [v_1^T,...,v_m^T]$
13: **对于** $j=1,...,k$，**执行**
14：采样小批量 ($X$,$Y$) $\sim G_j$
15：计算 EGM $M_j$（式 3）
16: **对于**每个概念 $c_i$ **执行**
17：计算主导类别 $y_i'$ 和敏感度 $S_i$（式 4）
18：$H_i^{(y)} \leftarrow \mathbb{I}(y_i' = y)$，$y=1,...,|\mathcal{Y}|$
19: **返回** $\{P^{(y)}\}_{y=1}^{|\mathcal{Y}|}$

并将其与各自非主导类别中的样本进行混合。形式化地，我们有

$$X' = \lambda X + (1-\lambda) X_{(\mathcal{C},P^{(y)})},\ Y' = Y, \tag{5}$$

其中 $\lambda \sim \text{Beta}(\beta_1,\beta_2)$。$X_{(\mathcal{C},P^{(y)})}$ 表示从概念库 $\mathcal{C}$中以概率 $P^{(y)}$ 采样得到的概念图像。($X$，$Y$) 从一个子集中抽取，其中每个 $Y_i \in \mathcal{Y}/\{y\}$。直观而言，敏感度越高的概念，主导类别与其他类别之间的不平衡程度就越大。为补偿这种不平衡，我们设计了这种留一增强：敏感度较高的概念图像会被更频繁地采样并混合，从而实现概念分布平衡。通过这种方式，经过干预的数据集消除了训练数据集中涉及多个伪概念的伪相关性。

**自适应缓解**。然而，模型可能在训练的不同阶段学习到各种伪相关。为彻底纠正模型的决策边界，概念敏感度需要相应调整。因此，如图2所示，我们提出了一个自适应框架 DISC，它通过迭代执行伪概念发现（步骤 2）和概念感知干预（步骤 3）。在每个训练轮次，DISC 都会计算概念敏感度，以指导概念感知干预。随后，模型在经过新一轮干预的数据集上继续训练，其中的伪相关已被消除。因此，DISC 能够逐步缓解模型在此前训练中学到的伪相关。我们在第 5 节的定理 1 中，为概念敏感度和最终模型的收敛提供了保证。通过按照公式5迭代混合图像，DISC 降低了伪概念对最终模型的影响，并提升了模型的泛化能力。

表1。 总体实验结果。最佳结果以**粗体** 显示，次佳结果带下划线。

| | MetaShift | | Waterbirds | | FMoW | | ISIC |
|---|---|---|---|---|---|---|---|
| | 平均准确率 | 最差准确率 | 平均准确率 | 最差准确率 | 平均准确率 | 最差准确率 | 平均 AUROC |
| ERM | 72.9 ± 1.4% | 62.1 ± 4.8% | 97.0 ± 0.2% | 63.7 ± 1.9% | 53.0 ± 0.6% | 32.3 ± 1.3% | 36.4 ± 0.7% |
| ERM+数据增强 | 75.5 ± 1.7% | 65.7 ± 3.3% | 87.4 ± 0.5% | 76.4 ± 2.0% | 55.5 ± 0.4% | 35.7 ± 0.3% | 38.9 ± 1.5% |
| UW | 72.1 ± 0.9% | 60.5 ± 3.8% | 96.3 ± 0.3% | 76.2 ± 1.4% | 52.5 ± 0.5% | 30.7 ± 1.5% | 39.2 ± 0.6% |
| IRM | 73.9 ± 0.8% | 64.7 ± 2.1% | 87.5 ± 0.7% | 75.6 ± 3.1% | 50.8 ± 0.1% | 30.0 ± 1.4% | 45.5 ± 3.6% |
| IB-IRM | 74.8 ± 0.2% | 65.6 ± 1.1% | 88.5 ± 0.9% | 76.5 ± 1.2% | 49.5 ± 0.5% | 28.4 ± 0.9% | 38.6 ± 1.5% |
| V-REx | 72.7 ± 1.7% | 60.8 ± 5.5% | 88.0 ± 1.4% | 73.6 ± 0.2% | 48.0 ± 0.6% | 27.2 ± 0.8% | 24.5 ± 6.4% |
| CORAL | 73.6 ± 0.4% | 62.8 ± 2.7% | 90.3 ± 1.1% | 79.8 ± 1.8% | 50.5 ± 0.4% | 31.7 ± 1.2% | 37.9 ± 0.7% |
| Fish | 64.4 ± 2.0% | 53.2 ± 4.5% | 85.6 ± 0.4% | 64.0 ± 0.3% | 51.8 ± 0.3% | 34.6 ± 0.2% | 42.0 ± 0.8% |
| GroupDRO | 73.6 ± 2.1% | 66.0 ± 3.8% | 91.8 ± 0.3% | **90.6 ± 1.1%** | 52.1 ± 0.5% | 30.8 ± 0.8% | 36.4 ± 0.9% |
| JTT | 74.4 ± 0.6% | 64.6 ± 2.3% | 93.3 ± 0.3% | 86.7 ± 1.5% | 52.5 ± 0.3% | 33.4 ± 0.9% | 33.8 ± 0.0% |
| DM-ADA | 74.0 ± 0.8% | 65.7 ± 1.4% | 76.4 ± 0.3% | 53.0 ± 1.3% | 51.6 ± 0.2% | 34.2 ± 0.8% | 35.8 ± 1.0% |
| LISA | 70.0 ± 0.7% | 59.8 ± 2.3% | 91.8 ± 0.3% | 88.5 ± 0.8% | 52.8 ± 0.9% | 35.5 ± 0.7% | 38.0 ± 1.3% |
| **DISC** | 75.5 ± 1.1% | **73.5 ± 1.4%** | 93.8 ± 0.7% | 88.7 ± 0.4% | 53.9 ± 0.4% | **36.1 ± 1.8%** | **55.1 ± 2.3%** |

## 4. 实验

我们旨在回答以下研究问题：

• **RQ1**：与最先进的基线方法相比，DISC 在具有伪相关的数据集上表现如何?

• **RQ2**：DISC 的训练动态有哪些特点？其中有哪些见解有助于未来的研究?

•**RQ3**：各个组件如何影响 DISC 的性能并促成其改进?

### 4.1. 设置

**数据集**。我们在附录 C 中总结了这些数据集。我们研究具有各种类型伪相关性的图像分类任务。具体而言，Waterbirds（Sagawa 等，2020 年）将每个类别与水或陆地背景相关联；MetaShift（Liang 和 Zou，2022）则为每个类别构建互不相交的伪属性。我们还使用 Wilds Benchmark 中的 FMoW（Koh 等，2021年），其卫星图像采集自不同地理区域，可能由此产生伪相关性。此外，我们还研究一项具有挑战性的任务 ISIC（Codella 等，2019），该任务将皮肤病变的皮肤镜图像分类为良性或黑色素瘤。我们采用 Bissoto 等(2020) 中的训练-测试划分，其中每个训练划分都放大了与 7 种伪属性的相关性。这项任务颇具难度，因为多种伪属性（例如，毛发和凝胶气泡）可能同时存在，并覆盖皮肤病变区域。

**基线方法**。我们将 DISC 与经验风险最小化（ERM）进行比较，包括使用和不使用数据增强的情况；向上加权（UW），即提高少数群体实例的权重；不变学习算法：IRM（Arjovskyet al.，2019）、IB-IRM（Ahuja et al.，2021）；领域泛化/适应方法：V-REx（Kruegeret al.，2021）、CORAL（Sun & Saenko，2016）和 Fish（Shi et al.，2021）；实例重加权方法：GroupDRO（Sagawaet al.，2020）、JTT（Liuet al.，2021a）；数据增强方法：DM-ADA（Xu et al.，2020）、LISA（Yao et al.，2022）。

**概念库**。受先前研究启发（Cimpoiet al.，2014；Fong &Vedaldi，2018），我们构建了一个概念库，其中包含 6 个类别下的 224 个概念。我们使用预训练的文生图模型 Stable Diffusion（Rombachet al.，2022）为每个概念生成 200 张图像。为避免不切实际的干预，我们为每个数据集选择概念类别，如表 4所示。概念库的构建以及概念类别选择的详情见附录 D。

**模型训练**。我们在附录E中总结了超参数，并使用高斯混合模型（GMM）作为聚类算法。在 Waterbirds 中，由于多数组与少数组之间存在极端不平衡，我们提高少数组的权重，以获得更稳定的结果。对于其他数据集的训练，我们不要求提供组信息。

**评估**。对于 ISIC，由于组大小是 $2^7$ 考虑伪属性的组合后，形成了许多较小的组，因此我们按照Bissoto 等人（2020 年）中的惯例，计算跨训练—测试划分的平均 AUROC 分数。对于其他数据集，我们评估平均性能和最差组性能。所有实验均重复三次。

### 4.2. 总体结果（RQ1）

**对基线方法的分析**。在表1中，我们总结了 DISC 与基线方法的总体性能。我们观察到，采用数据增强的 ERM 始终优于 ERM，表明简单的数据增强能够有效防止模型过拟合。此外，我们发现 GroupDRO 在 MetaShift 和 Waterbirds 数据集上表现良好。然而，在 FMoW 和 ISIC 数据集上，其性能接近 ERM 或低于 ERM，这与 Gulrajani 和 Lopez-Paz （2021年）的观察一致；Koh 等人（2021年）指出，GroupDRO 通常无法在野生环境中提升 ERM 的性能。此外，鉴于性能不足，在不变学习框架下训练的模型表现并不理想。

此外，我们还关注 ISIC 数据集。一个有趣的

观察是，JTT 的表现不佳——平均 AUROC 比 ERM 低 3%—— 这也印证了我们的假设：当少数实例较为稀少时，实例重加权的有效性在很大程度上受到限制。在这种情况下，数据混合策略的结果同样不尽如人意。具体而言，LISA 提出的标签内混合可能导致更强的伪相关，因为它增加了多数组的数量。总之，基线方法的局限性使其无法持续取得成功。

**DISC 的有效性**。总体而言，就最差群体准确率而言，DISC 的表现优于大多数基线方法。具体来说，在 MetaShift 和 ISIC 数据集中，我们分别比最佳基线方法提升了 7.5% 和 9.6%。这一证据表明，当伪属性组合存在时，DISC 能够缓解偏差。在 FMoW 中，DISC 在平均准确率和最差群体准确率上均取得最先进水平的结果，表明我们的方法在野外图像数据集中也能表现出色。在 Waterbirds 中，DISC 的表现比 JTT 高 2.7%，但不及 GroupDRO。一种可能的解释是，概念库中的图像未能完全涵盖伪属性，从而限制了缓解效果。尽管如此，DISC 仍能提供可解释性，帮助理解模型偏差，相关讨论见第 4.3 节。

### 4.3. 可解释性与训练动态（RQ2）

除了出色的任务性能外，我们的算法还凭借概念敏感度提供了高度透明性。在此，我们验证概念敏感度能够正确揭示伪相关，并帮助用户理解训练过程中的伪相关。

**对 DISC 解释的验证**。我们考察概念敏感度是否忠实地反映了训练数据中的伪相关。对于每个概念，我们计算其在各训练轮次中的累积概念敏感度，以表明整体伪相关的程度。累积敏感度最高的前 3 个概念是 **“点状”** (0.056)、**“条纹”** (0.032) 和 **“有污渍的”**(0.030)。同时，我们借鉴 Cram´er's V（Cram´er，2016）来衡量 7 个真实伪属性的伪相关程度，其中**凝胶气泡** (0.184)、**尺子** (0.411)、**墨水**(0.215) 是伪属性中相关程度最高的几个。有趣的是，
在图 3中，我们发现伪概念与伪属性之间存在很强的对应关系。例如，尺子与“条纹”具有较高的特征级相似度。

结论有两点。首先，解释结果与真实伪相关属性高度一致，表明其值得信赖。我们还在附录中对 DISC 的解释结果与现有可解释性方法进行了定性比较F ，以进一步展示 DISC 解释结果的高质量。其次，我们展示了当某些

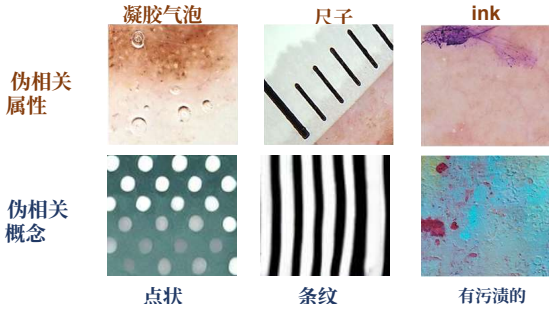



图*3*。伪属性与伪概念的对齐。

概念库中缺少概念时，我们的方法仍然适用，例如，尺子可以由保留相同基本属性的其他概念替代。这种全局且明确的解释清楚地揭示了虚假相关性。

**在缺失真实伪相关概念时，解释结果的鲁棒性与缓解效果**前面的示例表明，即使概念库中未包含真实概念，DISC 仍能找到高度相似的概念。因此，我们旨在探究这种模式是否具有一致性。我们设计了两项实验：（1）在 MetaShift 上移除与“猫”相关的“沙发”和“床”概念；（2）在 Waterbirds 上移除与“陆鸟”相关的“竹子”和“森林”概念。对于每种情况，我们在移除概念后运行 DISC算法，并观察移除前后相应类别的解释结果。

移除前后的前 3 个解释分别是（**皱纹、床、窗帘**）→（**壁炉、卧室、佩斯利纹**）（MetaShift）和（**竹子、森林、花盆**）→（**树冠、地面、植物**）（Waterbirds）。有趣的是，我们发现移除前后的解释存在一些概念重叠（例如，**“床”**→ **“卧室”**（MetaShift），以及**“森林”** → **“树冠”**和**“花盆”**→**“植物”**（Waterbirds））。我们进一步研究了移除概念对最差群体表现的影响。具体而言，MetaShift 上的表现下降了 0.2%，Waterbirds 上则下降了 0.9%。缺失的概念对 MetaShift 的影响较小。尽管 Waterbirds 数据集上的表现对缺失概念更为敏感，但移除后的表现仍优于大多数基线方法。

当缺失伪相关概念的 CAV 与其他概念的概念激活向量相似时，DISC 能够发现伪相关。其直觉在于，伪相关概念（例如，森林）与其他概念（例如，树冠）可能共享部分关键属性（例如，叶子或绿色），这些属性在一定程度上导致了伪相关，因此它们在嵌入空间中的 CAV 相似。因此，正如我们的实验结果所支持的那样，即使缺少真实标注，DISC 的解释和性能依然稳健。

**训练动态如何降低概念敏感度**。概念敏感度反映了模型受伪偏差影响的程度，可用于探查模型当前的

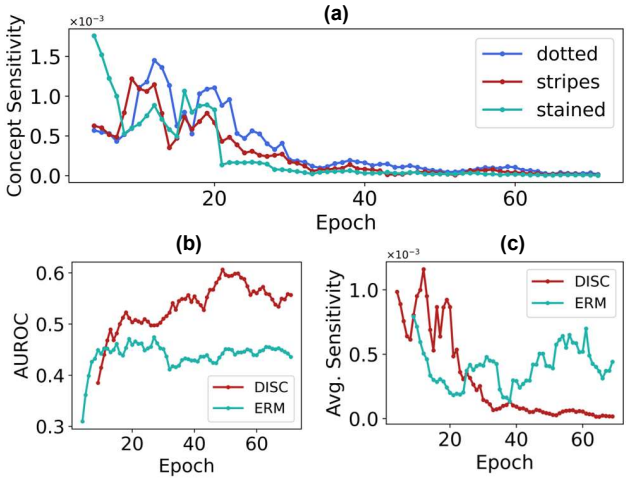



图 4.ISIC 上的训练动态。（a）单个概念敏感度与训练轮次的关系。（b）测试 AUROC，以及（c）DISC 和 ERM 在训练过程中的平均概念敏感度。

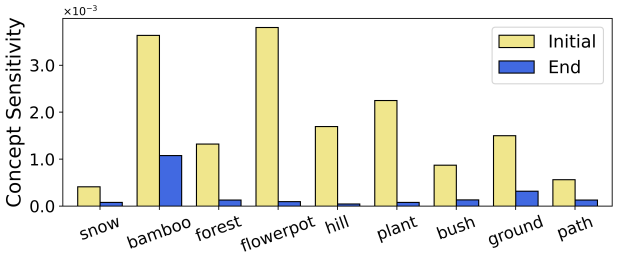



图 5.Waterbirds 上陆鸟类别的伪概念在训练开始和结束时的概念敏感度。

状态。如图4(a)所示，我们观察了训练过程中各个概念的敏感度。在权重随机初始化时，模型往往会在训练早期从伪属性中学习。相应地，排名前三的概念在训练初期的敏感度相对较高。令人欣慰的是，我们通过概念感知干预，在不同环境之间维持了伪概念的平衡，使平均敏感度逐渐降至接近于零。

此外，我们比较了 DISC 和 ERM 的三个概念的平均敏感度以及任务性能。在图 4 （b）和（c）中，我们发现 ERM 的平均概念敏感度有所上升，并在最后保持较高水平。我们认为，与标签形成错误关联并被模型记住的伪概念，导致了 ERM 的性能不佳。相比之下，DISC 最终达到了较低的概念敏感度。这一模式在所使用的数据集中具有一致性。作为另一个例子，在图 5 中，我们展示了在 Waterbirds 上训练前后概念敏感度的对比。平均敏感度的下降表明，模型权重已经达到一个“最佳点”，此时模型不受伪概念的影响。

### 4.4. 消融与敏感性研究（RQ3）

在此，我们通过实证分析剖析了算法中（1）概念敏感度、（2）概念感知干预和（3）自适应缓解的贡献。我们分别提出了三个消融模型：

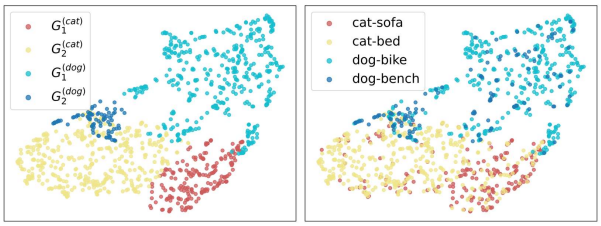
• **DISC-Randint**，该方法舍弃概念敏感度，并随机采样概念图像用于治愈步骤。

•**DISC-Reweight**，该方法将治愈步骤替换为对不太可能包含伪概念的实例进行重加权。形式上，实例（$x_j$, $y_j$）的权重为 $\exp\{-\sum_{i=1}^{m} P(y_j)_i \cdot \max\{0, g(x_j)^T v_i\} / |g_{x_j}|\}$，与其表示 $g(x_j)$ 和敏感概念的概念激活向量之间的对齐程度成负相关。

• **DISC-Inadaptive**并非每个训练轮次都更新概念敏感度，而是根据预训练的 ERM 模型生成概念敏感度，并将其固定下来，在训练期间进行干预。

**消融结果**。在表7（附录G）中，我们报告了消融结果。通过比较 DISC-Randint 和 DISC，我们发现，有效降低伪相关的不仅是干预本身，恰当的干预同样至关重要。这里的“恰当”是指“应选择哪些概念图像”以及“应针对特定概念干预多少训练数据”，而概念敏感度正是实现这一点的关键。此外，DISC-Reweight 与 DISC 的比较表明，概念感知干预能够促进伪概念之间的平衡，进一步减轻伪偏差，这是 DISC 成功的关键。DISC-Inadaptive 的表现始终逊于 DISC，在 F MoW 和 ISIC 上也逊于 DISC-Randint。具体而言，在这两个数据集上，我们发现，使用固定的概念敏感度分数虽然消除了敏感度较高的概念所产生的伪相关，但其他概念的伪相关程度可能会加剧，这表明我们的自适应机制对于更彻底地消除伪相关至关重要。总的来说，这三种消融模型验证了我们框架的有效性。

**无监督聚类**。在步骤 1 中，我们对训练实例进行无监督聚类，以便在伪属性上找到良好的分层方式。这里，我们关注 DISC 对聚类结果的依赖。为搜索簇数 $k$，我们遵循 Sohoni 等人（2020 年）的做法，采用轮廓系数作为启发式指标。由于篇幅限制，我们将结果列于表 9（附录 G ）中。我们展示了，在大多数数据集上，DISC 在很宽的簇数范围内都优于最佳基线。我们进一步在图 6中对 MetaShift 上的聚类结果进行了可视化，其中各簇与真实分组分配高度吻合。附录 G中还提供了更多可视化结果。因此，我们验证了由训练好的 ERM 得到的聚类对于构建数据环境具有信息价值。




(a) 聚类结果　　(b) 真实分组

图6。聚类与分组分配的可视化。

## 5. 理论分析

本节提供理论层面的洞见，以理解 DISC 在消除伪相关方面的优势。

一般假设。我们将高斯混合模型视为数据生成机制，该机制被机器学习理论广泛采用，以帮助理解复杂现象（Montanari 等，2019；Liu 等，2021b；Ji 等，2021；Deng 等，2021；Zhang 等，2022a）。我们考虑概念均已得到良好学习的情形，并在概念层面构建模型。具体而言，因果（不变）概念被建模为 $x_{inv} = y \cdot \mu + \epsilon_{inv}$，其中 $y \in \{-1, 1\}$ 表示类别索引，$\mu \in \mathrm{R}^{p_1}$ 表示因果概念的信号，其维度数量为 $p_1$，而 $\epsilon_{inv}$ 是均值为0且协方差矩阵为$\Sigma_1$的高斯噪声项。我们进一步假设类别是平衡的，即$\mathbb{P}(Y = 1) = \mathbb{P}(Y = -1) = 1/2$。假设因果概念的分布在训练环境和测试环境中保持不变。此外，伪概念被建模为 $x_{spu} = \gamma_y^{(i)} + \epsilon_{spu}$，其中 $\gamma_y^{(i)} \in \mathbb{R}^{p_2}$控制伪相关性，并会根据不同环境 $i \in [k]$而变化。由于 $x_{spu}$ 是伪相关的，因此在测试分布 $\mathcal{P}_{te}$中有 $\gamma = 0$。与 $\epsilon_{inv}$类似，$\epsilon_{spu}$ 是噪声项，服从均值为0且协方差矩阵为 $I$的高斯分布。由于$\mu$和$\gamma$的每个坐标都表示一个概念，我们假设 $\gamma_y^{(i)}$ 的取值为0或1，其中1表示类别 $y$的第 $i$个环境中存在相应的概念。

我们考虑最小化 $\ell_2$ 分类问题的损失，这在深度学习理论领域中得到了广泛应用（Ma& Belkin，2017；Shankaretal.，2020 年；Liang & Recht，2021年）。这里，我们将提出的 DISC 方法与标准 ERM 方法进行比较

$$(\hat{\mu}_{ERM}, \hat{\gamma}_{ERM}) = \underset{\mu,\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \mu^\top x_{inv,i} - \gamma^\top x_{spu,i})^2,$$

其中分类器构造为$\hat{}\ C_{ERM}(x) = sgn(\hat{}\mu_{ERM}^\top x_{inv} + \hat{}\gamma_{ERM}^\top x_{spu})$。类似地，我们将 DISC 生成的分类器定义为$\hat{}\ C_{DISC}(x) = sgn(\hat{}\mu_{DISC}^\top x_{inv} + \hat{}\gamma_{DISC}^\top x_{spu})$，其中$\hat{}\mu_{DISC}$和$\hat{}\gamma_{DISC}$是算法1生成的解。

**定理1**。假设 *(1)*。 $supp(\gamma_y^{(i)})$ 对于不同的 互不相交，且 $Var(\{\gamma_{y,j}^{(i)}\}_{i=1}^k) > K_0$ 对于 $j \in [p_2]$ 和某个常数 $K_0 > 0$，*(2)*。 $\|\mu\|_\infty \to 0$ 当 $p_1 \to \infty$ 时，且 $K_1 \le \lambda_{\min}(\Sigma_1) \le \lambda_{\max}(\Sigma_1) \le K_2$ 对于某些常数 $K_1$、$K_2 > 0$，*(3)*。 $p_1/n \to 0$ 且 $p_2$ 固定。当训练样本量 $n$ 足够大时，算法*1*以指数速度收敛。此外，以至少 $1-o(1)$ 的概率，解 $(\hat{}\mu_{DISC}, \hat{}\gamma_{DISC})$ 满足

$$\mathbb{P}_{\mathcal{P}_{te}}[\hat{C}_{DISC}(x) \neq y] < \mathbb{P}_{\mathcal{P}_{te}}[\hat{C}_{ERM}(x) \neq y].$$

我们在附录B中阐明假设并给出详细证明。该定理表明，通过反复发现并进行干预，DISC 减轻了伪概念对最终模型贡献的变化。因此，DISC 减少了最终模型中的虚假相关性，并且优于 ERM。

## 6. 结论与讨论

我们提出了 DISC，这是一种有原则的方法，可以便捷地发现伪相关，再通过数据增强来缓解这些相关性。DISC 基于一项经验观察：在许多情况下，伪属性在数据的不同子集中存在异质性。系统性的实验表明，DISC 能显著提升模型泛化性。此外，它还能帮助我们了解哪些概念具有敏感度，以及这种敏感度在训练过程中如何变化。

**局限性**。虽然 CAV 将嵌入空间与概念空间联系起来，但学习 CAV 需要在训练过程中进行额外计算。另一个局限是，使用生成模型构建的概念库可能存在自身偏差，从而限制缓解方法的效果。

**未来工作**。有趣的是，我们在 CIFAR-10-C 上进行了实验，发现 DISC 在四种损坏类型上的平均表现比 ERM 高出 13.1%，这表明 DISC 在**分布外泛化**方面具有潜力。此外，未来的工作还可以构建**更完善的概念库和用于自动概念类别选择的工具**，详见附录D。还可以将**DISC 的适用范围**扩展到多目标视觉数据集和自然语言处理任务，或使用 DISC 分析 SENN 等技术生成的概念（Alvarez-Melis 和 Jaakkola，2018）。

## 致谢

我们感谢斯坦福大学的 Serina Chang、Zhi Huang、Lingjiao Chen、Boyang Deng、Ruocheng Wang 和 Yang Zheng，他们为改进我们的工作提出了很好的建议。我们也感谢 ICML 2023 会议的匿名审稿人提出的深刻意见。Linjun Zhang 的研究部分得到 NSF DMS-2015378 的支持。James Zou 的研究部分得到 NSF CAREER 和斯隆奖学金的资助。

## 参考文献


Abid, A., Y¨uksekg¨on¨ul, M., and Zou, J. 使用概念反事实解释有意义地调试模型错误。发表于 *ICML*，2022。

Ahmed, F.、Bengio, Y.、van Seijen, H. 和 Courville, A. C. 基于组不变预测的系统性泛化。发表于*ICLR*，2021年。

Ahuja, K., Caballero, E., Zhang, D., Bengio, Y., Mitliagkas, I. 和 Rish, I. 不变性原理与信息瓶颈相结合，实现分布外泛化。2021年。

Alvarez-Melis, D. 和 Jaakkola, T. S. 通过自解释神经网络实现稳健可解释性。发表于*NeurIPS*，2018年。

Arjovsky、M.、Bottou、L.、Gulrajani、I. 和 Lopez-Paz、D. 不变风险最小化。*arXiv* 预印本 *arXiv:1907.02893*，2019 年。Bahadori、M. T. 和 Heckerman、D. 通过因果分析消除基于概念的解释中的偏差。发表于*ICLR*，2021年。Bau、D.、Zhou、B.、Khosla、A.、Oliva、A. 和 Torralba、A. 网络解剖：量化深度视觉表征的可解释性。发表于*CVPR*，2017 年。Ben-Tal、A.、den Hertog、D.、Waegenaere、A. D.、Melenberg、B. 和 Rennen、G. 受不确定概率影响的优化问题的鲁棒解。管理科学，2013 年。Bissoto、A.、Valle、E. 和 Avila、S. 皮肤病变数据集和模型中的偏差消除？没那么简单。发表于*CVPR Workshops*，2020 年。Bontempelli、A.、Teso、S.、Giunchiglia、F. 和 Passerini、A. 部件原型网络的概念级调试。*ICLR*，2022。B¨uhlmann、P. 不变性、因果关系与鲁棒性。统计科学，35(3):404–426，2020 年。Byrd、J. 和 Lipton、Z. C. 重要性加权在深度学习中有何影响？发表于*ICML*，2019 年。Cai、T. T. 和 Zhang、L. 高维线性判别分析：最优性、自适应算法与缺失数据。英国皇家统计学会杂志 *B* 辑：统计方法学，81(4):675–705，2019 年。Cai、T. T. 和 Zhang、L. 一种针对高维稀疏二次判别分析的凸优化方法。统计年鉴，49(3):1537–1568，2021年。Cai、T. T.、Wang、Y. 和 Zhang、L. 隐私的代价：差分隐私下参数估计的最优收敛速率。统计年鉴，49(5):2825–2850，2021年。

Chen, C.、Li, O.、Tao, D.、Barnett, A.、Rudin, C. 和 Su, J. 这看起来像那样：用于可解释图像识别的深度学习。发表于*NeurIPS*，2019年。Cimpoi, M.、Maji, S.、Kokkinos, I.、Mohamed, S.、，以及 Vedaldi, A. 野外纹理描述。发表于 *CVPR*，2014年。Clark, C.、Yatskar, M. 和 Zettlemoyer, L. 不要走捷径：用于避免已知数据集偏差的基于集成的方法。发表于 Inui, K.、Jiang, J.、Ng, V. 和 Wan, X.（编），*EMNLP*，2019年。Codella, N. C. F.、Rotemberg, V.、Tschandl, P.、Celebi, M. E.、Dusza, S. W.、Gutman, D. A.、Helba, B.、Kalloo, A.、Liopyris, K.、Marchetti, M. A.、Kittler, H. 和 Halpern, A. 面向黑色素瘤检测的皮肤病变分析 2018：由国际皮肤成像协作组织（ISIC）主办的挑战赛。2019年。Cra m´er, H. 统计学数学方法（*PMS-9*），第*9*卷。普林斯顿大学出版社，2016年。Creager, E.、Jacobsen, J.-H. 和 Zemel, R. 用于不变学习的环境推断。发表于国际机器学习会议，第2189–2200页。PMLR，2021年。Dagaev, N.、Roads, B. D.、Luo, X.、Barry, D. N.、Patil, K. R. 和 Love, B. C. 一个好得令人难以置信的先验，用于减少捷径依赖。2021年。Deng, Z.、Zhang, L.、Vodrahalli, K.、Kawaguchi, K. 和 Zou, J. Y. 对抗训练通过更好的表征助力迁移学习。神经信息处理系统进展，34:25179–25191，2021年。Eyuboglu, S.、Varma, M.、Saab, K. K.、Delbrouck, J.、Lee-Messer, C.、Dunnmon, J.、Zou, J. 和 R´e. Domino：利用跨模态嵌入发现系统性错误。发表于*ICLR*，2022年。Fong, R. 和 Vedaldi, A. Net2vec：量化并解释概念如何由深度神经网络中的滤波器编码。发表于*CVPR*，2018年。Geirhos, R.、Rubisch, P.、Michaelis, C.、Bethge, M.、Wichmann, F. A. 和 Brendel, W. 在 ImageNet 上训练的卷积神经网络偏向纹理；提高形状偏差可改善准确率和鲁棒性。发表于 *ICLR*，2019年。Gulrajani, I. 和 Lopez-Paz, D. 寻找失落的领域泛化。发表于 *ICLR*，2021年。Hagos, M. T.、Curran, K. M. 和 Namee, B. M. 识别虚假相关性，并通过基于解释的学习加以纠正。2022年。

Han, Z.，Liang, Z.，Yang, F.，Liu, L.，Li, L.，Bian, Y.，Zhao, P.，Wu, B.，Zhang, C.，和 Yao, J. UMIX：通过不确定性感知混合改进针对子群体偏移的重要性加权。

He, K., Zhang, X., Ren, S., 和 Sun, J. 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。

Hu, W., Niu, G., Sato, I., 和 Sugiyama, M. 分布鲁棒监督学习能否带来鲁棒分类器？载于*ICML*，2018年。

Huang, G., Liu, Z., Van Der Maaten, L. 和 Weinberger, K. Q. 密集连接卷积网络。载于*IEEE* 计算机视觉与模式识别会议论文集，第 4700–4708 页，2017。Jain, S., Lawrence, H., Moitra, A. 和 Madry, A. 将模型失效提炼为潜在空间中的方向。2022。Ji, W., Deng, Z., Nakada, R., Zou, J. 和 Zhang, L. 对比对特征学习的力量：理论分析。*arXiv* 预印本 *arXiv:2110.02473*，2021。Kaushik, D., Hovy, E. H. 和 Lipton, Z. C. 使用反事实数据增强学习造成差异的关键因素。载于*ICLR*，2020。Khani, F. 和 Liang, P. 移除伪特征可能损害准确率，并对不同组造成不成比例的影响。载于 Elish, M. C., Isaac, W. 和 Zemel, R. S.（编），*FAccT*，2021。Kim, B., Wattenberg, M., Gilmer, J., Cai, C. J., Wexler, J., Vi´egas, F. B. 和 Sayres, R. 超越特征归因的可解释性：使用概念激活向量（TCAV）进行定量测试。载于*ICML*，2018。Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B. 和 Liang, P. 概念瓶颈模型。载于*ICML*，2020。Koh, P. W., Sagawa, S., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T. 等。WILDS：野外分布偏移基准。载于国际机器学习会议，第 5637–5664 页。PMLR，2021。Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Le Priol, R. 和 Courville, A. 通过风险外推（REX）实现分布外泛化。载于国际机器学习会议，第 5815–5826 页。PMLR，2021。Kumar, N., Berg, A. C., Belhumeur, P. N. 和 Nayar, S. K. 用于人脸验证的属性和明喻分类器。载于*ICCV*，2009。

Lampert, C. H., Nickisch, H. 和 Harmeling, S. 通过类别间属性迁移学习检测未见过的物体类别。载于*CVPR*，2009。

Lang, O., Gandelsman, Y., Yarom, M., Wald, Y., Elidan, G., Hassidim, A., Freeman, W. T., Isola, P., Globerson, A., Irani, M., 和 Mosseri, I. 以风格解释：训练一个 GAN，在风格空间中解释分类器。载于*ICCV*，2021年。

Levy, D., Carmon, Y., Duchi, J. C., 和 Sidford, A. 用于分布鲁棒优化的大规模方法。载于*NeurIPS*，2020年。Li, Z. 和 Xu, C. 发现图像分类器中未知的偏差属性。载于*ICCV*，2021年。Li, Z., Hoogs, A., 和 Xu, C. 使用去偏交替网络发现并缓解未知偏差。载于*ECCV*，2022。Liang, T. 和 Recht, B. 插值分类器很少出错。*arXiv* 预印本 *arXiv:2101.11815*，2021年。Liang, W. 和 Zou, J. MetaShift：用于评估上下文分布偏移和训练冲突的数据集集合。载于国际学习表征会议，2022。Liu, E. Z., Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., 和 Finn, C. 只需训练两次：无需组信息即可提升组鲁棒性。载于*ICML*，第 6781–6792 页。PMLR，2021a。Liu, H., HaoChen, J. Z., Gaidon, A., 和 Ma, T. 自监督学习对数据集不平衡更具鲁棒性。*arXiv* 预印本 *arXiv:2110.05025*，2021b。Lundberg, S. M. 和 Lee, S.-I. 一种统一的模型预测解释方法。载于*NeurIPS*。2017。Ma, S. 和 Belkin, M. 深入浅层：大规模浅层学习的计算视角。神经信息处理系统进展，30，2017。Montanari, A., Ruan, F., Sohn, Y., 和 Yan, J. 最大间隔线性分类器的泛化误差：过参数化情形下的高维渐近分析。*arXiv* 预印本 *arXiv:1911.01544*，2019。Nakada, R., Gulluk, H. I., Deng, Z., Ji, W., Zou, J., 和 Zhang, L. 理解多模态对比学习并纳入非配对数据。载于国际人工智能与统计学会议，第 4348–4380 页。PMLR，2023。Nam, J., Cha, H., Ahn, S.-S., Lee, J., 和 Shin, J. 从失败中学习：从有偏分类器中去除分类器的偏差。神经信息处理系统进展，33，2020。

Namkoong, H. 和 Duchi, J. C. 基于 f-散度的分布鲁棒优化随机梯度方法。载于*NeurIPS*，2016。

Nauta, M., Walsh, R., Dubowski, A., 和 Seifert, C. 揭示并纠正皮肤癌诊断机器学习模型中的捷径学习。*Diagnostics*，2021年。

Oren, Y., Sagawa, S., Hashimoto, T. B., 和 Liang, P. 分布鲁棒语言建模。发表于*EMNLP-IJCNLP*，2019年。

Pinto, F., Yang, H., Lim, S., Torr, P. H. S., 和 Dokania, P. K. Regmixup：作为正则化器的 Mixup 能够出人意料地提升准确率和分布外鲁棒性。2022。

Plumb, G., Ribeiro, M. T., 和 Talwalkar, A. 借助解释发现并修复虚假模式。2021年。

Ribeiro, M. T., Singh, S., 和 Guestrin, C. "我为什么应该相信你？"：解释任意分类器的预测。发表于*KDD*，第1135–1144页，2016年。

Rombach, R., Blattmann, A., Lorenz, D., Esser, P. 和 Ommer, B. 使用潜在扩散模型进行高分辨率图像合成。发表于*CVPR*，2022。

Sagawa, S., Koh, P. W., Hashimoto, T. B. 和 Liang, P. 面向群体偏移的分布鲁棒神经网络：正则化对于最坏情况泛化的重要性。发表于*ICLR*，2020 年。

Santurkar, S., Tsipras, D. 和 Madry, A. BREEDS：子群体偏移基准测试。发表于*ICLR*，2021年。

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. 和 Batra, D. Grad-CAM：通过基于梯度的定位从深度网络生成视觉解释。发表于*ICCV*，第 618–626 页，2017。

Seo, S., Lee, J. 和 Han, B. 利用伪属性无监督学习去偏表征。发表于*CVPR*，2022。

Shankar, V., Fang, A., Guo, W., Fridovich-Keil, S., Ragan-Kelley, J., Schmidt, L., 和 Recht, B. 无切线的神经核。发表于国际机器学习会议，第 8614–8623 页。PMLR，2020 年。

Shi, Y., Seely, J., Torr, P. H., Siddharth, N., Hannun, A., Usunier, N., 和 Synnaeve, G. 用于领域泛化的梯度匹配。*arXiv*预印本*arXiv:2104.09937*，2021年。

Singla, S. 和 Feizi, S. 显著的 ImageNet：如何在深度学习中发现伪特征？发表于*ICLR*，2022。

Singla, S., Moayeri, M. 和 Feizi, S. 使用显著性 ImageNet 进行核心风险最小化。abs/2203.15566，2022。

Sohoni, N. S., Dunnmon, J., Angus, G., Gu, A. 和 R´e, C. 不让任何子类掉队：粗粒度分类问题中的细粒度鲁棒性。发表于*NeurIPS 2020*，2020。Sun, B. 和 Saenko, K. Deep coral：深度领域自适应的相关性对齐。发表于欧洲计算机视觉会议，第 443–450 页。施普林格，2016。Teney, D., Abbasnejad, E. 和 van den Hengel, A. 通过打乱数据来提升泛化能力。*arXiv*，2002.11894，2020。Teso, S. 和 Kersting, K. 解释性交互式机器学习。发表于*AIES*，2019。

¨Teso, S., Alkan, O., Stammer, W., 和 Daly, E. 在交互式机器学习中利用解释：综述。人工智能前沿，2023。

Utama, P. A.、Moosavi, N. S. 和 Gurevych, I. 朝着消除自然语言理解模型中的未知偏差迈进。在 *EMNLP*，2020 年。

Wu, Y.、Wang, X.、Zhang, A.、He, X. 和 Chua, T. 发现图神经网络的不变性依据。在*ICLR*，2022。

Xu, M.、Zhang, J.、Ni, B.、Li, T.、Wang, C.、Tian, Q. 和 Zhang, W. 使用领域混合进行对抗式领域自适应。在 *AAAI*人工智能大会论文集，第 34 卷，第 6502–6509 页，2020 年。

Yaghoobzadeh, Y.、Mehri, S.、des Combes, R. T.、Hazen, T. J. 和 Sordoni, A. 使用可遗忘样本提高对虚假相关性的鲁棒性。在*EACL*，2021年。

Yao, H., Wang, Y., Li, S., Zhang, L., Liang, W., Zou, J. 和 Finn, C. 通过选择性增强提升分布外鲁棒性。发表于*ICML*，2022。

Ye, H., Zou, J. 和 Zhang, L. 先冻结后训练：迈向伪相关和特征噪声条件下可证明的表示学习。发表于国际人工智能与统计学会议，第 8968–8990 页。PMLR，2023。

Y¨uksekg¨on¨ul, M., Wang, M. 和 Zou, J. 事后概念瓶颈模型。*CoRR*，2022。

Zhang, H., Ciss´e, M., Dauphin, Y. N. 和 Lopez-Paz, D. mixup：超越经验风险最小化。发表于 *ICLR*，2018。

Zhang, J., Menon, A., Veit, A., Bhojanapalli, S., Kumar, S. 和 Sra, S. 通过分布鲁棒优化应对标签偏移。发表于*ICLR*，2021a。

Zhang, L.、Deng, Z.、Kawaguchi, K.、Ghorbani, A. 和 Zou, J. mixup 如何提升鲁棒性和泛化能力？发表于*ICLR*，2021b。

Zhang, L.、Deng, Z.、Kawaguchi, K. 和 Zou, J. mixup 何时以及如何改善校准。发表于 国际机器学习会议，第 26135–26160 页。PMLR，2022a。

Zhang, M., Sohoni, N. S., Zhang, H. R., Finn, C. 和 R´e, C. Correct-n-contrast：一种提升对虚假相关性鲁棒性的对比方法。发表于*ICML*，2022b。

# A. 符号表

表2. 方法部分中使用的主要符号。点击此处返回论文正文。

| 符号 | 含义 |
|---|---|
| $\mathcal{D}_{tr}/\mathcal{D}_{te}$ | 训练/测试数据集 |
| $\mathcal{P}_{tr}/\mathcal{P}_{te}$ | 训练/测试分布 |
| $\mathcal{Y}$ | 标签空间 |
| $f$ | 一个深度模型 |
| $g/h$ | $f$的编码器/最后一层线性层 |
| $\phi/\omega$ | $f/h$的参数 |
| $k$ | 每个类别的簇数 |
| $m$ | 概念库中的概念数量 |
| $d$ | 隐藏维度的数量 |
| $\mathcal{C}$ | 概念库 |
| $c_i$ | 概念库中的第 $i$ 个概念 $\mathcal{C}$ |
| $v_i$ | 概念 $c_i$的概念激活向量 |
| $y_i'$ | 概念 $c_i$的主导类别 |
| $\mathcal{P}_{c_i}$ | 第 $i$个概念的图像分布 |
| $S_i$ | 概念 $c_i$的概念敏感度 |
| $I_i^p/I_i^n$ | 概念 $c_i$的正/负图像集 |
| $N^p/N^n$ | 正/负图像集中的图像数量 |
| $G_j^{(y)}$ | 类别 $y$中的第 $j$个簇 |
| $G_j$ | 第 $j$个环境 |
| $M_j$ | 与 $G_j$对应的环境梯度矩阵 |
| $H^{(y)}$ | 主导类别的概念布尔掩码 $y$ |
| $P(y)$ | 类别 $y$的概念采样分布 |
| $X_{(\mathcal{C},P^{(y)})}$ | 从概念库 $\mathcal{C}$ 中以概率 $P^{(y)}$采样的图像 |

# B. 理论分析

**定理 (定理 1的重述。假设 (1) 成立。** $supp$ $y$ $>$ $K_0$ $j\in[p_2]$ $K_0>0$ 2 $\|\mu\|_\infty\to 0$ $p_1\to\infty$ $K_1\le\lambda_{\min}\ \Sigma_1\ \le\lambda_{\max}\ \Sigma_1\ \le K_2$ $(\gamma_y^{(i)})$对于不同的，彼此不相交，并且 方差$(\{\gamma_{y,j}^{(i)}\}_{i=1}^k)$对于 和某个常数，*( )*。当 时，且 () () 对于某些常数 $K_1,K_2>0$, *(3)*。 $p_1/n\to 0$且 $p_2$固定。则当训练样本量 $n$ 足够大时，算法 *1* 以指数速度收敛。此外，以至少 $1-o(1)$的概率，解 $(\hat{}\mu_{DISC},)$ 满足

$$\mathbb{P}_{\mathcal{P}_{te}}[\hat{C}_{DISC}(x)\neq y]<\mathbb{P}_{\mathcal{P}_{te}}[\hat{C}_{ERM}(x)\neq y].$$

**关于假设的说明**。我们对每个假设作如下直观说明：

• **(1)** 支撑集运算 $supp(\cdot)$ 是由输入向量中所有非零项对应的索引组成的集合。支撑集互不相交这一条件假设不同类别中的伪概念互不相交，这一点得到了我们实验观测的支持，例如，图 6 见第 4节。方差条件假设伪相关的强度（以伪概念贡献的方差衡量）不会太小。该条件通过假设存在一定程度的可区分性，对于检测伪概念是必要的。

• **(2)** $\lambda_{\text{最小值}}(\cdot)$ 和 $\lambda_{\text{最大值}}(\cdot)$ 分别表示矩阵的最小和最大特征值。关于 $\|\mu\|_\infty$ 的条件假设因果概念的贡献应当分散开来。关于 $\Sigma_1$ 特征值上下界的假设是统计学和机器学习文献中的标准假设（ Cai 和 Zhang，2019；2021；Cai 等人，2021；Nakada等人，2023）。

• **(3)** 我们假设伪概念的数量有限。此外，训练数据的数量 $n$ 需要大幅超过因果概念的数量，以便模型能够很好地学习不变概念，从而进行分类。

证明。 我们首先分析每次迭代中的两个主要步骤：概念敏感度计算，以及通过 mixup 进行梯度更新。

将小批量大小记为 $B$。我们注意到，假设所有概念都已得到充分学习，并且只分析在这些充分学习的概念之上的拟合。整个证明过程中都不会出现 CAV 矩阵乘法。在我们的模型设定下，在迭代 $t$时，对于 $j\in[k]$， $M_j$在（3）中的值等于

$$M_{j,y}=X^\top(y\mathbf{1}_B-X(\mu_t;\gamma_t))/B.$$

因果特征和伪特征对应的部分分别为

$$M_{j,y;inv}=X_{inv}^\top(y\mathbf{1}_B-X(\mu_t;\gamma_t))/B=X_{inv}^\top y\mathbf{1}_B/B-X_{inv}^\top X_{inv}\mu_t/B-X_{inv}^\top X_{spu}\gamma_t/B,$$

and

$$M_{j,y;spu}=X_{spu}^\top(y\mathbf{1}_B-X(\mu_t;\gamma_t))/B=X_{spu}^\top y\mathbf{1}_B/B-X_{spu}^\top X_{inv}\mu_t/B-X_{spu}^\top X_{spu}\gamma_t/B.$$

由于假设 $X$为次高斯分布，我们有

$$\begin{aligned}M_{j,y;inv}=&\mathbb{E}[X_{inv}^\top y\mathbf{1}_B/B-X_{inv}^\top X_{inv}\mu_t/B-X_{inv}^\top X_{spu}\gamma_t/B]+O(\sqrt{\frac{p_1}{n}})\\=&\mu-(\mu\mu^\top+\Sigma_1)\mu_t-\gamma_t^\top\gamma_y^{(i)}\cdot\mu+O(\sqrt{\frac{p_1}{n}}),\end{aligned}$$

and

$$\begin{aligned}M_{j,y;spu}=&\mathbb{E}[X_{spu}^\top y\mathbf{1}_B/B-X_{spu}^\top X_{inv}\mu_t/B-X_{spu}^\top X_{spu}\gamma_t/B]+O(\sqrt{\frac{p_2}{n}})\\=&y\cdot\gamma_y^{(i)}-\mu^\top\mu_t\cdot\gamma_y^{(i)}-(\gamma_y^{(i)}(\gamma_y^{(i)})^\top+I)\gamma_t+O(\sqrt{\frac{p_2}{n}})\\=&y\cdot\gamma_y^{(i)}-\mu^\top\mu_t\cdot\gamma_y^{(i)}-(\gamma_y^{(i)})^\top\gamma_t\cdot\gamma_y^{(i)}+I\gamma_t+O(\sqrt{\frac{p_2}{n}}).\end{aligned}$$

接下来，由于现在考虑的是二元分类设定， $S_i$在（4）中的值现在等于

$$S_{i,j}^{(y)}=\mathrm{Var}(\{\gamma_{y,j}^{(i)}\}_{i=1}^k).$$

现在，对于不变部分，由于 $\|\mu\|_\infty=O(1)$，且固定的 $p_2$意味着 $|\gamma_t^\top\gamma_y^{(i)}|=O(1)$，我们对所有 $j\in[p_1]$都有：

$$S_{i,j}^{(y)}=o(1).$$

此外，根据假设，我们有 $\|\gamma_y^{(i)}\|>1$ and $\mathrm{Var}(\{\gamma_{y,j}^{(i)}\}_{i=1}^k)>K_0$ for $j\in[p_2]$，因此对于所有 $j\in[p_2]$ 以及某个常数 $K_0>0$，有

$$S_{i,j}^{(y)}=\Omega(1).$$

因此，以至少 $1-o(1)$ 的概率，按照 $P^{(y)}$ 进行采样时，总会从 $\cup_{i=1}^k\mathrm{supp}(\gamma_y^{(i)})$ 中抽取伪概念。我们将此类事件记为 $E$，其中 $\mathbb{P}(E)\geq1-o(1)$。

现在我们分析事件 $E$ 上的 mixup 部分。根据我们的模型设定，对于 $j\in[p_2]$，概念图像被建模为基向量 $e_j$，其第 $j$ 个分量等于 1，表示该概念的存在。

令 $\tilde{\gamma}_y=\frac{1}{k}\sum_{i=1}^k\gamma_y^{(i)}$。那么，经过 mixup 后，对于伪概念，存在一个向量 $c_{y,\ spu}$ 其支撑属于 $\cup_{i=1}^k\mathrm{supp}(\gamma_y^{(i)})$，且非零元素位于 (0，1) 中，使得梯度更新变为

$$S_{spu}^{(i)}=\sum_{y\in\{-1,1\}}y\cdot(\tilde{\gamma}_y+c_{-y,spu})-\mu^\top\mu_t\cdot(\sum_{y\in\{-1,1\}}\tilde{\gamma}_y+c_{-y,spu})-(\sum_{y\in\{-1,1\}}(\tilde{\gamma}_y\tilde{\gamma}_y^\top+c_{-y,spu}c_{-y,spu}^\top)+I)\gamma_t+O(\sqrt{\frac{p_2}{n}}),$$

请注意 我们假设 $y\in\{-1,1\}$。利用 $\tilde{\gamma}_y$ 和 $c_{-y,\ spu}$ 的支撑集互不相交这一事实，可得 th at

$$S^t_{spu}=\sum_{y\in\{-1,1\}} y\cdot(\tilde{\gamma}_y+c_{-y,spu})-\mu^\top\mu_t\cdot(\sum_{y\in\{-1,1\}}\tilde{\gamma}_y+c_{-y,spu})-(\sum_{y\in\{-1,1\}}(\tilde{\gamma}_y+c_{-y,spu})(\tilde{\gamma}_y+c_{-y,spu})^\top+I)\gamma_t+O(\sqrt{\frac{p_2}{n}}).$$

此外，不变（因果）部分的梯度

$$S^t_{inv}=\mu-(\mu\mu^\top+\Sigma_1)\mu_t-\gamma_t^\top\gamma_y^{(i)}\cdot\mu+O(\sqrt{\frac{p_1}{n}}).$$

因此，每次迭代中的更新 $t$ 的效果等同于通过梯度下降最小化损失函数 $\ell(\hat{}\mu,\ \hat{}\gamma)=$
$(\hat{}\mu;\hat{}\gamma)^\top\begin{pmatrix}\Sigma_1+\mu\mu^\top & 0\\ 0 & \sum_{y\in\{-1,1\}}(\tilde{\gamma}_y+c_{-y,spu})(\tilde{\gamma}_y+c_{-y,spu})^\top+I\end{pmatrix}(\hat{\mu};\hat{\gamma})+(\hat{\mu};\hat{\gamma})^\top(\mu;\sum_{y\in\{-1,1\}}y\cdot(\tilde{\gamma}_y+c_{-y,spu}))$.

由于 $\lambda_{\min}(\Sigma_1)$ $\lambda_{\min}(I)$，$>K_1$，$\ell$是一个强凸函数，这意味着算法1 以指数速度收敛。

最后，我们比较 DISC 和 ERM 的性能。

由于 $(\hat{}\mu_{DISC},\hat{}\gamma_{DISC})$ 是 $\ell$的极小值，我们可以写出其解析解:

$$\hat{\mu}_{DISC}=(\Sigma_1+\mu\mu^\top)\mu+O(\sqrt{\frac{p_1}{n}}),$$

and

$$\begin{aligned}\hat{\gamma}_{DISC}&=(\sum_{y\in\{-1,1\}}(\tilde{\gamma}_y+c_{-y,spu})(\tilde{\gamma}_y+c_{-y,spu})^\top+I)^{-1}\sum_{y\in\{-1,1\}}y\cdot(\tilde{\gamma}_y+c_{-y,spu})\\&=(\sum_{y\in\{-1,1\}}(\tilde{\gamma}_y)(\tilde{\gamma}_y)^\top+\sum_{y\in\{-1,1\}}(c_{-y,spu})(c_{-y,spu})^\top+I)^{-1}\sum_{y\in\{-1,1\}}y\cdot(\tilde{\gamma}_y+c_{-y,spu})\end{aligned}$$

同样，我们有

$$\hat{\mu}_{ERM}=(\Sigma_1+\mu\mu^\top)\mu+O(\sqrt{\frac{p_1}{n}}),$$

and

$$\hat{\gamma}_{ERM}=(\sum_{y\in\{-1,1\}}(\tilde{\gamma}_y)(\tilde{\gamma}_y)^\top+I)^{-1}\sum_{y\in\{-1,1\}}y\cdot\tilde{\gamma}_y.$$

由于 $\tilde{\gamma}_y$ 的所有元素均为 0 或 1，$c_{y,spu}$ 的所有元素均介于 0 和 1 之间，并且 $\tilde{\gamma}_y$ 和 $c_{-y,spu}$的支撑集互不相交，因此有

$$\|\hat{\gamma}_{DISC}\|<\|\hat{\gamma}_{ERM}\|.$$

现在，我们分析测试域中的误分类误差。对于任意的 $\hat{}\mu$和 $\hat{}\gamma$，有

$$\begin{aligned}\mathbb{P}_{\mathcal{D}_{te}}(sgn(\hat{\mu}^\top x_1+\hat{\gamma}^\top x_2)\neq y)&=\frac{1}{2}\mathbb{P}_{\mathcal{D}_{te}}(\hat{\mu}^\top\mu+\hat{\mu}^\top\epsilon_1+\hat{\gamma}^\top\epsilon_2>0)+\frac{1}{2}\mathbb{P}_{\mathcal{D}_{te}}(-\hat{\mu}^\top\mu+\hat{\mu}^\top\epsilon_1+\hat{\gamma}^\top\epsilon_2<0)\\&=\frac{1}{2}\mathbb{E}[\mathbb{P}(\hat{\mu}^\top\mu+\hat{\mu}^\top\epsilon_1+\hat{\gamma}^\top\epsilon_2>0\,|\,\epsilon_1)]+\frac{1}{2}\mathbb{E}[\mathbb{P}(-\hat{\mu}^\top\mu+\hat{\mu}^\top\epsilon_1+\hat{\gamma}^\top\epsilon_2<0\,|\,\epsilon_1)]\\&=\mathbb{E}[\Phi(-\frac{\hat{\mu}^\top\mu}{\sqrt{\|\hat{\mu}\|^2+\|\hat{\gamma}\|^2}})],\end{aligned}$$

其中，$\Phi$ 是标准正态分布的累积分布函数。因此，使用$\hat{}\mu^\top\mu=c\|\mu\|^2+O(\sqrt{p_1/n})>0$和 $\|\hat{}\gamma_{DISC}\|<\|\hat{}\gamma_{ERM}\|$，我们有

$$\mathbb{P}_{\mathcal{D}_{te}}[\hat{C}_{DISC}(x)\neq y]<\mathbb{P}_{\mathcal{D}_{te}}[\hat{C}_{ERM}(x)\neq y].$$

□

## C. 数据集

表2。 (*a*) **MetaShift 数据集**。

**目标**：将图像分类为猫/狗。

**伪特征**：物体/背景；沙发、床（猫）；长椅、自行车（狗）。

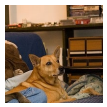
| **图像：** |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| **组** $g$**:** | 0 | 1 | 2 | 3 | 4 | 5 |
| **目标** $y \in \{0,1\}$**:** | 0（猫） | 0（猫） | 1（狗） | 1（狗） | 0（猫） | 1（狗） |
| **伪相关** $s$**:** | 0（沙发） | 1（床） | 2（长椅） | 3（自行车） | 4（架子） | 4（架子） |
| **# 训练数据：** | 231 | 380 | 145 | 367 | - | - |
| **# 验证数据（分布外）** | - | - | - | - | 34 | 47 |
| **# 测试数据：** | - | - | - | - | 201 | 259 |

表2. (*b*)**水鸟数据集**。

**目标**：鸟类类型；**伪特征**：背景类型。

| **图像：** |  |  |  |  |
|---|---|---|---|---|
| **组** $g$**:** | 0 | 1 | 2 | 3 |
| **目标** $y \in \{0,1\}$**:** | 0（陆鸟） | 0（陆鸟） | 1（水鸟） | 1（水鸟） |
| **伪相关** $s$**:** | 0（陆地） | 1（水） | 0（陆地） | 1（水） |
| **# 训练数据：** | 3,498（73%） | 184（4%） | 56（1%） | 1,057（22%） |
| **# 验证数据：** | 467 | 466 | 133 | 133 |
| **# 测试数据：** | 2,255 | 2,255 | 642 | 642 |

表2。 (*c*) **FMoW数据集**。

**目标**：62种建筑物或土地利用类别之一，例如公园、购物中心、水坝、体育场、机场。

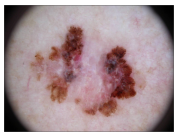
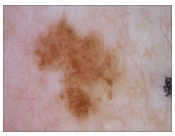
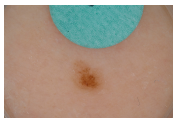
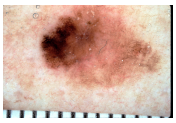
**伪特征**：未知（数据源中未明确给出）。

| **图像：** |  |  |  |  |  |
|---|---|---|---|---|---|
| **组** $g$**:** | 欧洲 | Asia | 美洲 | 非洲 | 大洋洲 |
| **# 训练数据：** | 34,816 | 17,809 | 20,973 | 1,582 | 1,641 |
| **# 验证数据：** | 7,732 | 4,121 | 6,562 | 803 | 693 |
| **# 测试数据：** | 5,858 | 4,963 | 8,024 | 2,593 | 666 |

表2. *(d)* **ISIC 数据集**。对于需要领域信息的方法，我们使用是否存在毛发作为领域标签。每个训练划分都会强化不同的相关性，相应的测试集则提供相反的相关性。

**目标**：良性 / 黑色素瘤皮损

**伪特征**：暗角、毛发、凝胶边缘、凝胶气泡、尺子、墨水标记/染色、贴片。

| | | | | | |
|---|---|---|---|---|---|
| **图像：** |  |  | … |  |  |
| **目标** $y \in \{0,1\}$**：** | 0（良性） | 0（良性） | … | 1（恶性） | 1（恶性） |
| **伪相关** $s$**：** | 补丁，凝胶边界 | ink，hair | … | 暗角，凝胶气泡 | 尺子，暗角 |
| **# 训练数据：** | 1,826 | | | | |
| **# 验证数据：** | 154 | | | | |
| **# 测试数据：** | 618 | | | | |

## D. 概念库

**概念类别**。表3列出了概念库中按 6 个类别划分的全部 224 个概念，分别为*(*颜色、纹理、自然、城市、家居、其他*)*。请注意，借助文生图生成模型，可以轻松扩展概念库中的用户自定义概念，因为获取概念图像的成本很低。

表3. 本文概念库的完整概念列表。

| 概念类别 | 概念 |
|---|---|
| **颜色** | [黑色、蓝色、绿色、红色、白色] |
| **纹理** | [混凝土、花岗岩、皮革、层压材料、金属、斑驳的、模糊感、条纹、波点、针织的、开裂的、褶边的、薄纱状的——有斑点的、鳞片状的、蕾丝状的、有沟槽的、层状的、薄纱状的、大理石纹的、有斑点的、有污渍的、编织的、缠结的、网状的、蛛网状的，螺旋状的、点状、交叉排线的、起皱的、编织的、坑洼的、结晶状的、佩斯利纹、有脉纹的、纤维状的、镶嵌的、气泡状的，褶皱的、网格、穿孔的、多孔的、交织的、涂抹状的、蜂窝状的、点缀的、方格状的、线条状的、带状的，凹凸不平的、锯齿状的、旋涡状的、有凹坑的、有雀斑状斑点的] |
| **自然** | [竹子、海滩、桥、灌木、树冠、泥土、田野、花、花盆、荧光灯、森林、草、地面、港口，山丘、湖、山、口鼻部、棕榈树、小路、植物、河流、沙子、海、雪、树、水] |
| **City** | [遮阳篷、底座、长椅、建筑物、泥土、围栏、田野、地面、房屋、窨井、小路、雪、街道] |
| **家居** | [空调、围裙、扶手椅、靠垫、阳台、楼梯扶手、浴室、浴缸、床、床上用品、床-间、柜子、地毯、天花板、椅子、枝形吊灯、五斗柜、台面、窗帘、坐垫、书桌、餐饮-间、门、门框、双开门、抽屉、水杯、抽油烟机、小雕像、壁炉、地板、花、花盆、荧光灯、地面、把手、横杆、床头板、车头灯、房屋、罐子、灯、灯光、微波炉、镜子、脚凳、烤箱、枕头、盘子、冰箱、沙发、楼梯、马桶] |
| **其他** | [鸟、猫、牛、狗、马、老鼠、爪子、手臂、背部、身体、耳朵、眼睛、眉毛、女性面部、腿、男性面部、脚、头发，手、头、内侧手臂、旋钮、嘴、脖子、鼻子、外侧手臂、垃圾桶、飞机、包、公共汽车、喙、自行车、百叶窗，木板、书、书柜、瓶子、碗、盒子、砖、篮子、桶、保险杠、罐头、烛台、帽子、汽车、纸板，陶瓷、链轮、烟囱、时钟、长途客车、咖啡桌、柱子、电脑、柜台、杯子、书桌、发动机、织物，风扇、水龙头、旗帜、地板、食物、脚踏板、框架、玻璃、键盘、盖子、扬声器、小型摩托车、摩托车、餐巾，包裹、彩绘、绘画、窗格、纸、底座、人、支柱、管道] |

**概念图像生成与示例**。所有概念图像均为合成图像，由使用预训练权重“stable-diffusion-v1-4”的 Stable Diffusion 模型生成，提示词采用概念名称或其复数形式。概念库的生成代码已公开于此链接。如图7所示，我们展示了概念库中精选的概念图像作为示例。




图7。 概念库中的概念图像示例。

**概念图像可能存在的偏差**。生成模型未必能完美地生成概念图像。虽然它们确实可能存在自身偏差，但由于其训练所用的数据集规模大得多，因此对于简单概念，它们不太可能包含更严重的伪相关性。根据实证结果，我们发现，与真实图像相比，合成概念图像的噪声或偏差更少。例如，我们观察到，BRODEN 数据集（Fong 与 Vedaldi，2018）中的“树”视觉概念图像往往与“人类”高度共现（例如，徒步时），而如图7所示，我们概念库中的合成图像则很少包含此类偏差。此外，先前的研究（Abid 等人，2022）也表明，学习 CAV 并不需要大量概念图像，因此可以对概念库进行简单筛选，进一步确保其可信度。

表 4。 每个数据集选定的概念类别

| | 概念类别 | | | | | |
|---|---|---|---|---|---|---|
| | 颜色 | 纹理 | 自然 | City | 家居 | 其他 |
| MetaShift | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |
| Waterbirds | ✔ | ✔ | ✔ | | | |
| FMoW | ✔ | ✔ | ✔ | ✔ | | |
| ISIC | ✔ | ✔ | | | | |

**概念类别选择与筛选**。如表 4所示，我们为每个数据集选择概念类别。选择的总体原则是根据对数据集上下文的先验知识，纳入数据集中出现的对象。以下是一些示例：

- 我们为所有数据集纳入颜色和纹理，因为这两个概念类别普遍存在。
- 根据 FMoW 是卫星图像数据集这一先验知识，我们纳入自然和城市类别，因为它们可能出现在数据集中，因此可能包含伪概念候选项。
- 根据 ISIC 是皮肤病数据集这一先验知识，我们将自然、城市等在该数据集中不存在的类别排除在概念候选项之外，仅纳入纹理和颜色概念。

在现实世界应用中，这种关于数据集上下文的知识是下游任务的根本需求，并且**具有泛化性**，可应用于其他数据集。此外，由于各数据集共享大型概念库，从业者可以选择类别而非单个概念，从而**减少工作量**。

此外，在我们的实现中，我们借鉴 Abid 等人(2022) 的方法，使用过滤模块筛选数据集中的相关概念。概念类别选择和过滤的好处包括：(1) 避免不现实的干预，例如，将动物图像与卫星图像进行 mixup；(2) 降低训练过程中计算 CAV 的计算成本。

**自动概念类别选择**。作为未来的研究方向，为了进一步避免针对未知下游任务进行概念类别选择，可以采用自动选择合适概念类别的协议。

- 利用图像识别模型识别数据集中已有的对象。
- 然后，提取概　　念或概念类别，并从我们与数据集无关的概念库中选取；该概念库定义于表　　3.

**学习概念激活向量**。为了学习 CAVs，我们对所有概念使用 $N^p = N^n = 150$。另一个未来的研究方向是利用困难负样本学习更准确的概念表征。例如，我们可以使用与树图像相似的概念图像，为树概念构建负样本集，例如草和花。为简便起见，本文使用随机采样来构建负样本集。

## E. 模型与优化细节

我们在 FMoW 上采用 DenseNet121（Huang 等人，2017），并在其他数据集上采用 ResNet-50（He 等人，2016）。超参数汇总于表 5。对于贝塔分布，我们在所有数据集中均使用 $\alpha = \mu = 2$。请注意，我们使用轮廓系数搜索每个类别的簇数，详情见附录 G。

表 *5.* DISC 训练期间的超参数。

| | 学习率 | 批量大小 | 权重衰减 | 每类簇数 |
|---|---|---|---|---|
| MetaShift | $5e$-4 | 16 | $1e$-4 | 2 |
| Waterbirds | $1e$-4 | 32 | $1e$-4 | 3 |
| FMoW | $1e$-4 | 10 | 0.0 | 3 |
| ISIC | $5e$-4 | 16 | $1e$-5 | 3 |

## F. 解释比较结果

这里，我们首先分析 DISC 生成的解释相较于现有识别伪相关的基线方法所具有的优势。我们从可解释性的三个维度展开研究：

- **类别/组级**：解释是否针对某个类别或组，这种解释相比实例级解释具有优势，因为它能够从多个实例中获得共同洞见。
- **概念/描述文本级**：解释是否基于描述文本或概念，而非特征图；前者更符合人类习惯且含义明确。
- **自适应**：解释在训练过程中是否具有自适应性或内在性，从而支持动态检查，而非事后解释。

我们考虑了不同类型的解释，包括现有的基于显著性的方法和基于概念的方法。值得强调的是，DISC 是唯一兼具这三项优势的方法。

在图 8 中，我们进一步对 DISC 及另外三种类型的解释进行了定性评估：(1) Grad-CAM（基于显著性的方法）。(2) Failure-Direction（Jain 等人，2022）（基于描述文本的方法）。(3) CCE（Abid 等人，2022）（基于概念的方法）。对于 Grad-CAM，与先前的观察类似，逐实例的显著性图可能难以解读，也难以从中获得全局洞见，以理解某个类别的预测。对于 Failure-Direction，我们遵循原论文计算描述文本分数，并通过聚合得到词语分数。具体而言，我们发现描述文本模型有时关注前景而非背景，导致部分描述文本无法为调试提供有效信息。此外，受图像描述模型的限制，这些解释缺乏多样性。对于 CCE，我们发现 DISC 和 CCE 的解释相似。这符合我们的预期，因为 DISC 和 CCE 都利用概念激活向量生成解释。此外，DISC 在模型训练过程中支持更动态的检查。

表6。 DISC 与现有方法的解释对比。

| | 解释类型 | 类别/组级 | 概念/描述文本级 | 自适应 |
|---|---|---|---|---|
| Singla & Feizi (2022)<br>Selvaraju 等人 （2017）<br>Singla 等人 （2022） | 基于显著性 | ✔（部分） | ✔（部分） | ✗ |
| Sohoni 等人 （2020 年）<br>Seo 等人 （2022） | 基于聚类 | ✔ | ✗ | ✗ |
| Creager 等人 （2021年）<br>Liu 等人 （2021a）<br>Li 等人（2022）<br>Ahmed 等人(2021年) | 基于划分 | ✔ | ✗ | ✗ |
| Abid 等人 (2022)<br>Bontempelli 等人 (2022) | 基于概念 | ✔ | ✔ | ✗ |
| Jain 等人 (2022)<br>Eyuboglu 等人 (2022) | 基于标题 | ✗ | ✔ | ✗ |
| Lang 等人 (2021年)<br>Li &Xu (2021年) | 生成式反事实样本 | ✗<br>✔ | ✗<br>✗ | ✗<br>✗ |
| **DISC** | 自适应概念方法 | ✔ | ✔ | ✔ |

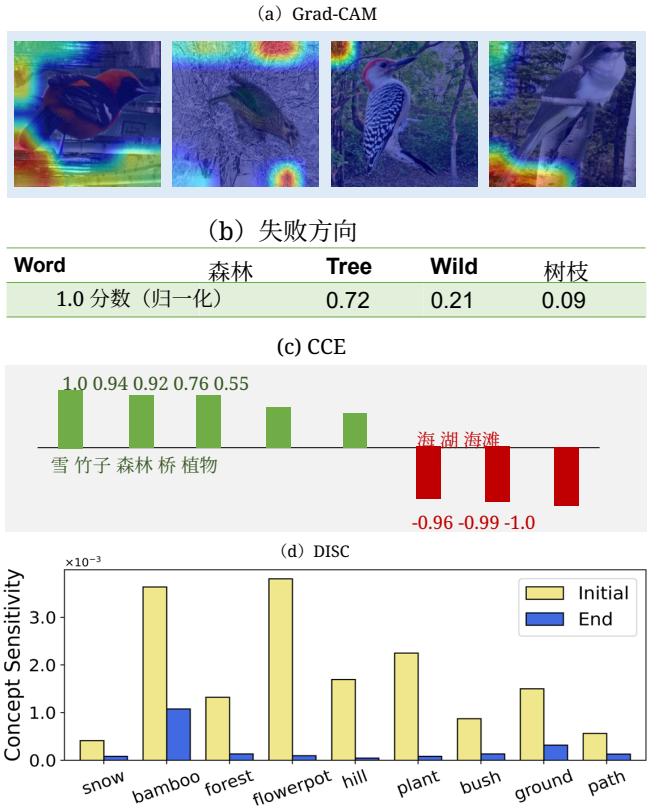



图 *8.* Waterbirds 中用于解释“**为何图像被预测为陆地鸟类**？”。(a) Grad-CAM 可视化。(b) Jain 等人 (2022) 生成的单词分数。(c) 使用 CCE 生成反事实样本时的平均概念分数（Abid 等人，2022）。(d) DISC 训练前后陆鸟 类别中伪概念的概念敏感度。

## G. 消融与敏感性研究结果

表 7。 模型设计选择的消融实验结果。

| | MetaShift | | Waterbirds | | FMoW | | ISIC |
|---|---|---|---|---|---|---|---|
| | 平均准确率 | 最差准确率 | 平均准确率 | 最差准确率 | 平均准确率 | 最差准确率 | 平均 AUROC |
| DISC-Randint | 71.7% | 64.5% | 91.0% | 85.9% | 53.0% | 32.1% | 49.3% |
| DISC-Reweight | 72.8% | 62.5% | 88.9% | 81.4% | 51.0% | 32.0% | 35.9% |
| DISC-Inadaptive | 73.0% | 68.3% | 89.6% | 86.5% | 51.9% | 31.8% | 47.1% |
| **DISC** | 75.4% | 72.6% | 93.8% | 88.7% | 53.9% | 36.1% | 55.1% |

**消融结果**。表 7 汇报了所有数据集上的消融结果。结论与正文中的陈述一致。具体而言，DISC 的表现大幅优于消融模型，从实证角度验证了我们的算法设计。

**无监督聚类**。我们使用轮廓系数作为启发式指标，来搜索每类簇数这一超参数。如图 9 所示，有趣的是，我们发现这一指标与大多数数据集上的测试表现高度一致。根据实证结果，每个类别采用较少的簇数（例如，3）通常能取得最佳效果。一种可能的解释是，随着簇数增加，概念敏感度可能会因识别出无关紧要的伪概念而变得消极且任意。我们认为，未来研究概念敏感度，或更广义地研究环境构建时，这也是一个值得探讨的视角。

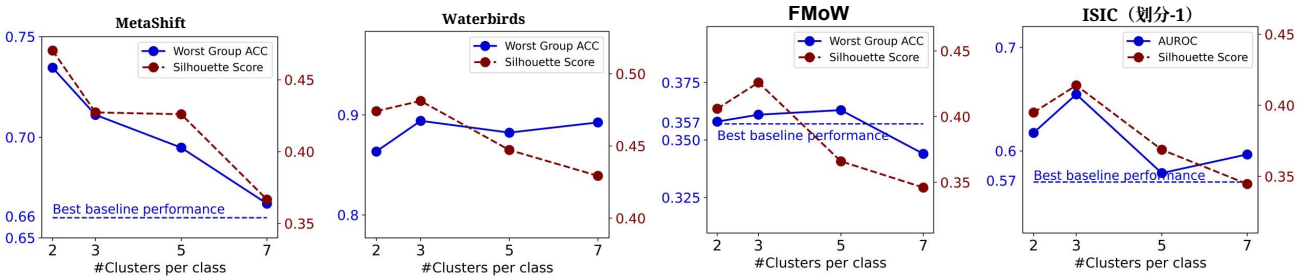
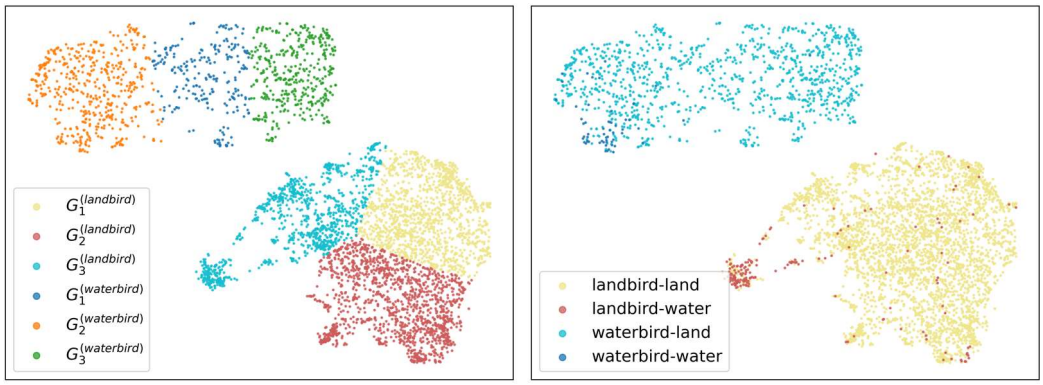



图 *9*。最差群体准确率和轮廓系数关于每个类别的簇数。对于 ISIC 数据集，我们报告了其中一个训练-测试划分上的敏感度结果。




**（a）聚类结果** **（b）真实分组**

图 *10.* Waterbirds 上聚类与分组分配的对比。

除了图 6 中 MetaShift 的聚类结果外，我们还在图 10 中展示了 Waterbirds 上的聚类结果。我们发现，聚类算法能够捕捉部分伪属性。然而，当各组极不均衡时，可能难以找到合适的数据环境。尽管 DISC 也优于大多数基线方法，但这些结果表明，即使环境“不精确”，DISC 也更加稳健。