对抗性医院无关特征学习用于 WSI 切片分类

Mengliang Zhang
Department of Computer Science and Engineering
The University of Texas at Arlington
Arlington, TX, U.S.A.
mxz3935@mavs.uta.edu

Jacob M. Luber
Department of Computer Science and Engineering
The University of Texas at Arlington
Arlington, TX, U.S.A.
jacob.luber@uta.edu

摘要—病理学基础模型（PFMs）在全幻灯片图像（WSI）诊断中展示了显著的潜力。然而，来自不同医院的病理图像由于扫描硬件和预处理风格的不同而有所差异，这可能导致 PFMs 无意间学习到特定医院的特征，从而对临床部署构成风险。在这项工作中，我们提出了关于因医院来源特性引起的 PFMs 领域偏差的第一个系统性研究。具体来说，我们（1）构建了一个量化 PFMs 中领域偏见的流水线，（2）评估并比较了多个模型的表现，并且（3）提出了一种轻量级对抗框架，在不修改编码器本身的情况下，从冻结表示中去除潜在的特定医院特征。通过引入一个可训练适配器和一个与梯度反转层（GRL）相连的领域分类器，我们的方法学习到了具有任务区分性但无领域偏差的表示。在多中心组织病理学数据集上的实验表明，我们的方法显著降低了领域预测能力，同时保持甚至提高了疾病分类性能，特别是在非领域（未见过医院）场景中。进一步分析，包括医院检测和特征空间可视化，确认了我们方法减轻医院偏见的有效性。我们将根据接受情况提供代码。

***Index Terms*—病理图像，领域偏倚，基础模型。**

I. 介绍

全幻灯片成像（WSI）已成为数字病理学的关键工具，使千兆像素的组织病理切片可进行扩展分析。在大多数下游应用中，WSIs 被分割成小块，深度模型——如 ResNet [3]、视觉变换器（ViT）[16] 或基础编码器如 UNI [5]PLIP [13]——被用于提取包括疾病分类、肿瘤分级和亚型分析等任务的视觉特征。这些块级特征构成了许多计算病理学工作流程的基础。

然而，在这种设置中，一个关键但常常被忽视的挑战是数据中存在的特定领域偏差。从不同医院或扫描仪收集的补丁在染色协议、图像分辨率、扫描仪伪影和组织准备方面经常有所不同。这些差异引入了虚假的相关性到学习表示中，导致模型无意中依赖于特定医院的线索而非真正的疾病相关信号。

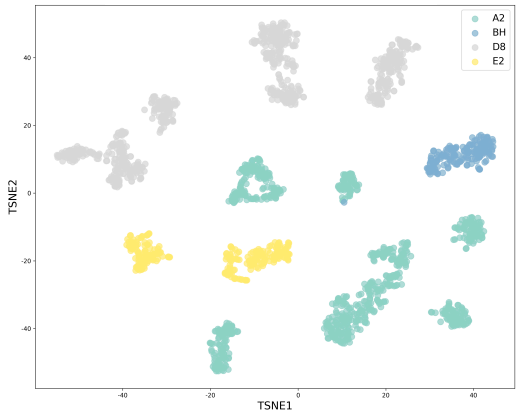

图 1. 使用 Phikon [9] 模型对 TCGA-BRCA 数据集的 patch 图像进行 t-SNE 可视化，展示了来自不同医院的同一疾病样本。颜色表示不同的医院来源。

即使是使用大规模数据集（如 Phikon [9] 和 UNI [5]）通过自监督学习训练的最先进的病理学基础模型 (PFMs) ——也表现出一定程度的领域偏差。如图 1所示，我们使用 Phikon [9] 模型对来自 TCGA-BRCA 数据集的切片嵌入进行了 t-SNE [14] 聚类分析，重点关注了来自不同医院但具有相同疾病的切片。同一来源的切片紧密聚在一起，表明学习到的特征保留了特定于医院的信息。这种现象对临床部署构成了风险：模型可能会依赖潜在的医院线索进行诊断，并且有可能从切片嵌入中泄露医院来源信息。

若干先前的研究探讨了病理学基础模型（PFMs）中的领域偏差。Vaidya 等人 [17] 研究了不同患者群体中 PFMs 的性能差异，而林等人研究了医院间的变化并分析了基于 DINO [20] 的特征学习限制。Edwin 等人 [15] 提出了一个用于来自不同医院来源的切片样本

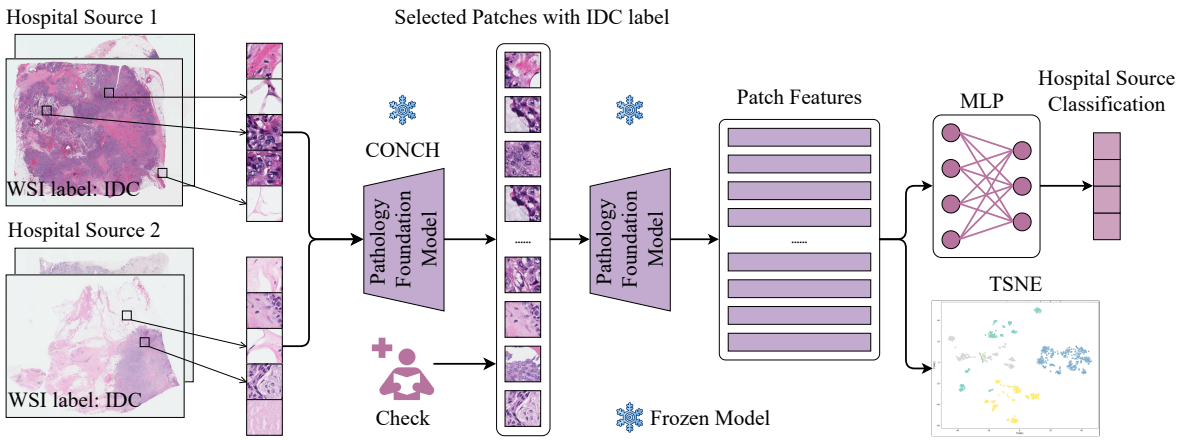

图 2. 评估病理基础模型中医院领域偏差的流程。仅使用与其 WSI 标签一致的切片。训练一个简单的多层感知器 (MLP) 来分类医院来源，其中更高的准确率表示更强的领域偏差。

上 PFMs 鲁棒性的度量标准，采用 KNN 方法来测量具有相同医院起源和疾病标签的相邻切片数量，从而计算出鲁棒性指数。其他研究试图通过染色标准化技术（如 Macenko [19] 标准化）来缓解领域转移问题，在特征提取之前将不同医院中的切片颜色分布标准化。然而，这些方法是非可学习的，对参数调整敏感，并且仅解决低层次的颜色差异，无法解决其他领域的变异来源——例如扫描仪缺陷或标注不一致。

另一条研究路线利用典型相关分析（CCA）将来自不同医院的图像块映射到一个共享特征空间。然而，这些方法存在显著的局限性。埃德温的 [15] 指标无法提供数据集级别的领域偏差评估。基于 CCA 的方法侧重于特征对齐，但忽略了对齐后的表示是否保留了疾病区分能力。瓦伊迪亚的 [17] 方法需要额外的 LoRA [21] 微调 PFMs，并且没有明确确保在保持诊断性能的同时减少领域偏差。

在这项工作中，我们首先建立了一个管道来评估病理基础模型的领域偏见。我们从 TCGA 数据集中提取补丁，并训练两个独立的多层感知器（MLP）用于疾病分类和医院来源分类。通过测量在提取的补丁特征上的医院分类准确率和 AUC，我们量化了模型中领域偏见的程度。

然后，我们提出了一种轻量级的对抗框架来减轻提取特征中的医院特定信息。我们的方法引入了一个可训练的投影头、一个领域分类器和一个疾病分类器，其中领域分类器通过梯度反转层（GRL）与投影头相连。在反向传播过程中，GRL 反转来自领域分类器的梯度，作为对抗反馈来抑制医院相关的线索，同时在整个训练过程中保持疾病的鉴别信息。我们在多个组织病理学数据集上验证了我们的方法。实验结果表明，所提出的对抗训练框架有效抑制了医院特定信号并缓解了领域偏差，同时保持甚至提升了疾病分类性能。

我们的主要贡献如下：

1) 我们建立了一个管道，用于识别和量化由病理学基础模型提取的特征在医院来源方面的领域偏差。
2) 我们提出了一种轻量级的对抗训练框架，该框架通过引入梯度反转层（GRL）来抑制图像特征中的医院相关判别线索，而无需修改编码器。
3) 在 TCGA 数据集上的实验表明，我们的方法在保留疾病分类性能的同时显著抑制了提取特征中的潜在医院信息。

II. 相关工作

近期视觉基础模型的进展对计算病理学产生了重大影响，特别是在整个幻灯片图像（WSIs）的斑块级特征提取器的设计上。诸如 CLIP [1] 和 DINO [20] 等模型已被广泛用于从图像斑块中提取可迁移的高维特征。最近，UNI [5] 和 CONCH [2] 在零样本和少样本学习设置下的多种病理任务——包括肿瘤分类、亚型分析和分级——中取得了令人印象深刻的结果。这些模型通常在大规模自然图像语料库或多模态数据集上进行预训练，然后作为冻结编码器使用。尽管它们具有强大的语义能力，但研究观察到其特征仍然包含数据集或站点特定的偏差，包括扫描仪伪影和染色变化。

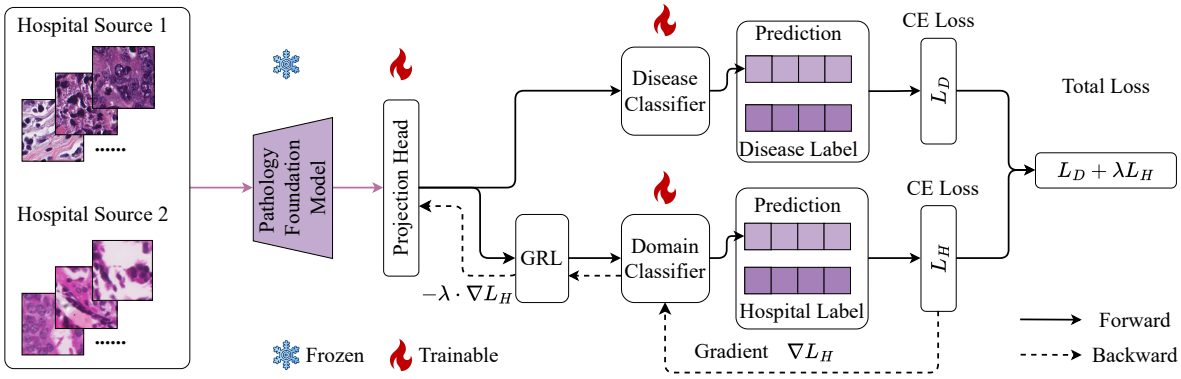

图 3. 对抗训练框架。

几项研究探讨了该背景下医院领域的偏差。Edwin 等人提出了一个基于 KNN 的鲁棒性度量标准，该标准通过计算共享相同医院来源和疾病标签的邻近补丁数量来计算鲁棒性指数。Vaidya 等人进一步引入了基于 LoRA 的微调方法，以使病理基础模型适应分布外（OOD）数据。然而，这些方法缺乏一个系统性的量化与缓解框架，为进一步优化留下了空间。

III. 方法论

A. 补丁选择

对于给定的 WSI 集合，我们首先通过使用可用的 demographic 信息（例如，性别和年龄）过滤 WSI 来构建一个队列，以选择具有可比临床状况的病例。此步骤有助于减少 demographic 因素对 WSI 疾病特征的潜在混淆影响。对于每个医院和每个疾病类别，我们选择指定数量的 WSI，努力在不同医院和疾病类别中保持平衡的样本量。

对于每个 WSI，我们遵循来自 CLAM [12] 的组织分割策略，应用 Otsu 阈值处理自动区分组织区域和背景。这一步生成一个用于后续补丁提取的组织掩模。在识别出的组织区域内，我们使用预定义网格模式的滑动窗口提取固定大小的补丁。每个提取的补丁都会经过质量控制，包括对组织覆盖范围和最小有效面积的检查，确保这些补丁保留足够的诊断价值。

接下来，我们利用 CONCH [2] 模型通过确保补丁级和 WSI 级标签之间的一致性来进一步优化补丁集，如图 2所示。这一选择的动机源于我们的初步观察，即 CONCH 嵌入产生的 t-SNE 分布对于来自不同地点的同一种疾病的补丁具有较少明显的医院聚类。我们使用 CONCH 执行零样本补丁分类，获取前 1 个预测标签及其概率，并仅选择那些 (1) 预测标签概率超过预定义阈值且 (2) 标签与 WSI 级真实标签匹配的补丁。此外，对采样的补丁进行手动检查以确保标签准确性和一致性。

在图 2中，对于提取的补丁，我们训练一个简单的多层感知器（MLP）基于医院来源和疾病标签进行分类。分类性能使用准确率、AUC 和 F1 分数来评估。

对于医院来源分类任务，更高的准确性表明可以从提取的特征中轻易推断出医院来源，反映出相应的病理基础模型存在更强的领域偏差。相比之下，对于疾病分类任务，更高的准确性则表明基础模型实现了更好的疾病区分性能。

B. 对抗方法

我们提出了一种轻量级的对抗训练框架，以抑制病理基础模型提取的特征中潜在的特定医院信息，同时保留疾病区分信号。该框架包含一个可训练的投影头、一个领域分类器和一个疾病分类器，其中领域分类器通过梯度反转层（GRL）与投影头相连。在反向传播过程中，GRL 反转来自领域分类器的梯度，提供对抗反馈以抑制学习表示中的特定医院线索。

a) 问题设定。： 令 $\mathcal{D} = \{(x_i, y_i, d_i)\}_{i=1}^{N}$ 表示一组 WSI 切片数据集，其中 $x_i \in \mathbb{R}^{H\times W\times 3}$ 是一个图像切片，y_i 是疾病标签，而 d_i 是域标签，表示来源医院。我们假设可以访问冻结的编码器 $E(\cdot)$（例如，UNI 或 CONCH），它将 x_i 映射到一个特征向量

$$\boldsymbol{f}_i = E(x_i) \in \mathbb{R}^D. \tag{1}$$

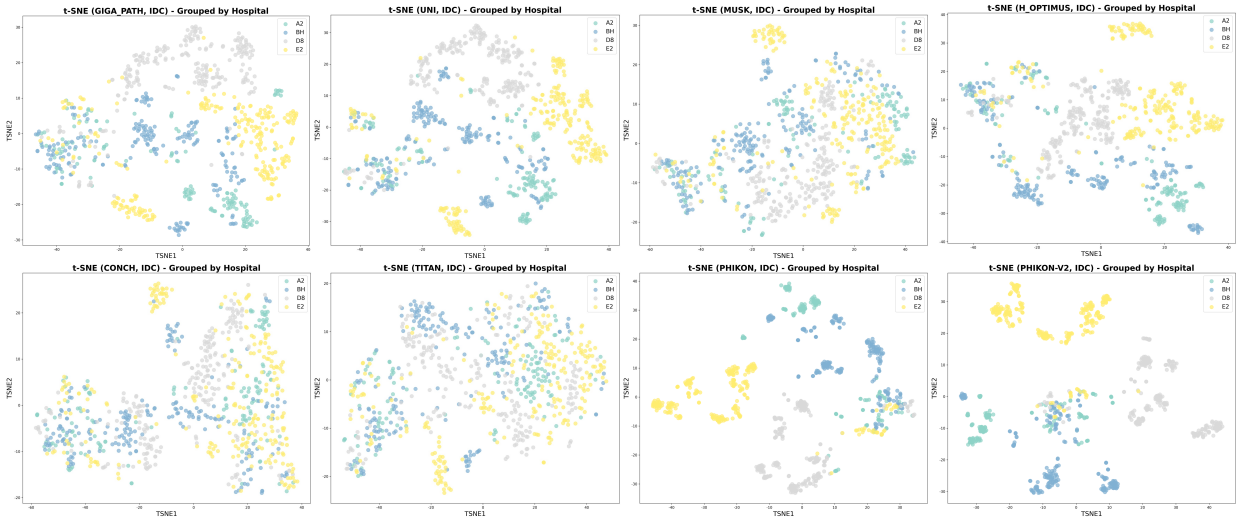

图 4. TCGA-BRCA 数据集中的不同模型的 IDC 补丁特征的 t-SNE 可视化。

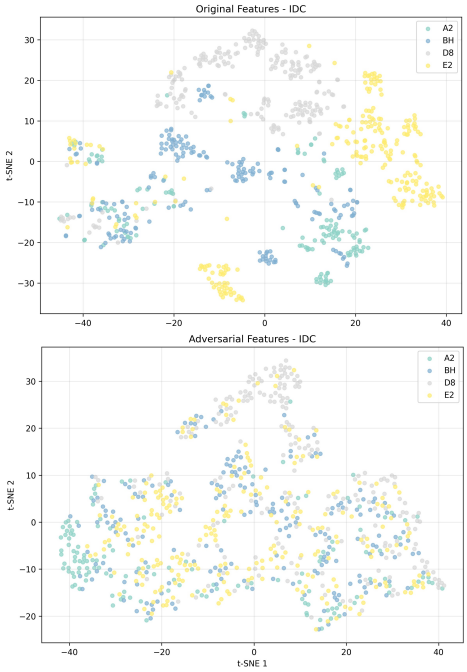

图 5. 使用 UNI 模型的原始 patch 特征的 t-SNE 聚类结果（顶部）和对抗训练后特征的 t-SNE 聚类结果（底部）。所有 patch 具有相同的标签，而不同颜色对应的点代表不同的医院。

我们的目标是学习一个变换后的表示 $\boldsymbol{z}_i = A(\boldsymbol{f}_i)$，使得：

- $\boldsymbol{z}_i$ 保留了用于预测 y_i 的疾病区分信息；
- $\boldsymbol{z}_i$ 抑制特定医院的信息，阻止可靠预测 d_i。

C. 模型架构

模型包含四个组件：

a) 冻结编码器 $E(x)$: 一个预训练的视觉编码器，用于提取基于补丁级别的特征。其参数在训练过程中保持不变。

b) 投影头 $A(\boldsymbol{f})$: 一个轻量级的 MLP，将编码器特征 $\boldsymbol{f}_i$ 投影到低维表示 $\boldsymbol{z}_i$:

$$\boldsymbol{z}_i = A(E(x_i)) \in \mathbb{R}^{D'}. \tag{2}$$

c) 疾病分类器 $C_y(\boldsymbol{z})$: 预测从投影特征中的疾病标签:

$$\hat{y}_i = C_y(\boldsymbol{z}_i). \tag{3}$$

d) 域分类器 $C_d(\cdot)$ 带梯度反转层: 预测来自经过 GRL 的 $\boldsymbol{z}_i$ 的医院来源:

$$\hat{d}_i = C_d\big(\mathrm{GRL}(\boldsymbol{z}_i)\big). \tag{4}$$

D. 梯度反转层（GRL）

在图 3中，GRL 在前向传递时充当恒等函数，$\mathrm{GRL}(\boldsymbol{z}_i) = \boldsymbol{z}_i$，但在反向传递时反转并缩放梯度:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{z}_i} \leftarrow -\lambda \cdot \frac{\partial \mathcal{L}}{\partial \boldsymbol{z}_i}, \tag{5}$$

其中 $\lambda \geq 0$ 控制对抗强度。这迫使投影头生成对于疾病预测仍然具有信息性的特征而对医院来源分类则变得不具信息性。

E. 损失函数

总损失是疾病损失和领域损失的加权和:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{D}} + \lambda \cdot \mathcal{L}_{\text{H}}, \tag{6}$$

其中 λ 与 GRL 缩放因子对齐。

表 I
不同模型在医院来源分类上的表现，在对抗训练前后的性能。我们报告了准确率、AUC 和 F1 的均值 ± 标准差。

模型	**方法**	**准确率 (±std)**	**AUC（± 标准差）**	**F1(± 标准差)**
混凝土	MLP	0.6049±0.0573	0.8442±0.0299	0.5459±0.0787
	Adversarial	0.2250±0.1675	0.5900±0.1309	0.1678±0.1175
GIGA_路径	MLP	0.6843±0.1178	0.9104±0.0388	0.6440±0.1217
	Adversarial	0.2427±0.1888	0.5719±0.1104	0.1411±0.1276
H_OPTIMUS	MLP	0.8064±0.0875	0.9652±0.0167	0.7753±0.1214
	Adversarial	0.3384±0.1570	0.6736±0.1333	0.2288±0.1040
马斯克	MLP	0.6886±0.0943	0.8980±0.0448	0.6283±0.1002
	Adversarial	0.2883±0.1620	0.5000±0.0000	0.1051±0.0545
PHIKON	MLP	0.8463±0.0735	0.9805±0.0105	0.8105±0.0969
	Adversarial	0.4533±0.1168	0.7349±0.1062	0.3590±0.1449
PHIKON-V2	MLP	0.9161±0.0282	0.9918±0.0036	0.8874±0.0684
	Adversarial	0.4760±0.1861	0.7629±0.1707	0.4372±0.2146
RESNET50	MLP	0.6019±0.0848	0.8305±0.0537	0.5460±0.0898
	Adversarial	0.2544±0.1611	0.5124±0.0248	0.0947±0.0531
泰坦	MLP	0.5667±0.0699	0.8226±0.0395	0.5134±0.0669
	Adversarial	0.2289±0.1124	0.5446±0.0891	0.0924±0.0450
联合国	MLP	0.7836±0.0730	0.9483±0.0317	0.7416±0.0954
	Adversarial	0.2166±0.1130	0.5788±0.0630	0.0875±0.0377
UNI2-H	MLP	0.7860±0.0951	0.9571±0.0199	0.7569±0.1072
	Adversarial	0.1928±0.1079	0.5142±0.0497	0.1045±0.0681
魏尔肖	MLP	0.7133±0.1507	0.9095±0.0675	0.6794±0.1574
	Adversarial	0.1335±0.1062	0.4938±0.0311	0.0558±0.0405

a) 疾病损失。: 我们采用标准交叉熵 (CE) 损失函数进行疾病预测：

$$\mathcal{L}_{\mathrm{D}} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{CE}\big(C_y(\boldsymbol{z}_i), y_i\big). \tag{7}$$

b) 域损失。: 域分类器被训练以使用交叉熵损失预测医院来源：

$$\mathcal{L}_{\mathrm{H}} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{CE}\big(C_d(\mathrm{GRL}(\boldsymbol{z}_i)), d_i\big). \tag{8}$$

在训练过程中，仅更新投影头和两个分类器，而编码器保持不变。在推理时，丢弃 GRL 和域分类器，最终预测结果为：

$$\hat{y}_i = C_y\big(A(E(x_i))\big) = C_y(\boldsymbol{z}_i). \tag{9}$$

IV. 实验

A. 设置

表 II
按疾病类型和医院来源统计 TCGA-BRCA 中的过滤补丁。

类别	**名称**	**过滤计数**
Disease	IDC	2,146
Disease	ILC	1,883
Hospital	E2	911
Hospital	A2	1,221
Hospital	D8	1,220
Hospital	BH	677
Total	All	4,029

表 III
不同模型与 MLP 和对抗方法的性能对比。我们报告了准确率、AUC 和 F1 的 ± 标准差均值。CONCH 模型的疾病分类结果仅供参考，因为 CONCH 用于补丁预过滤。

模型	方法	准确率 (±std)	AUC（± 标准差）	F1(± 标准差)
海螺壳	MLP	1.0000±0.0000	1.0000±0.0000	1.0000±0.0000
	Adversarial	1.0000±0.0000	1.0000±0.0000	1.0000±0.0000
GIGA_路径	MLP	0.9148±0.0307	0.9746±0.0063	0.9091±0.0336
	Adversarial	0.9227±0.0220	0.9611±0.0120	0.9169±0.0271
H_OPTIMUS	MLP	0.9062±0.0536	0.9731±0.0229	0.9020±0.0542
	Adversarial	0.9173±0.0418	0.9546±0.0255	0.9136±0.0428
马斯克	MLP	0.9391±0.0121	0.9877±0.0029	0.9359±0.0128
	Adversarial	0.9398±0.0127	0.9748±0.0073	0.9366±0.0127
PHIKON	MLP	0.8536±0.0915	0.9412±0.0440	0.8403±0.0955
	Adversarial	0.8763±0.0921	0.9259±0.0579	0.8695±0.0919
PHIKON-V2	MLP	0.8701±0.0755	0.9470±0.0433	0.8561±0.0847
	Adversarial	0.8596±0.0856	0.9365±0.0414	0.8459±0.0908
RESNET50	MLP	0.8512±0.0282	0.9307±0.0202	0.8420±0.0331
	Adversarial	0.8580±0.0472	0.9312±0.0327	0.8499±0.0501
泰坦	MLP	0.9389±0.0108	0.9831±0.0067	0.9349±0.0136
	Adversarial	0.9294±0.0155	0.9817±0.0079	0.9249±0.0180
统一	MLP	0.9337±0.0338	0.9822±0.0138	0.9291±0.0369
	Adversarial	0.9278±0.0368	0.9613±0.0363	0.9230±0.0400
UNI2-H	MLP	0.9219±0.0360	0.9823±0.0083	0.9182±0.0369
	Adversarial	0.9187±0.0432	0.9700±0.0199	0.9151±0.0438
魏尔肖	MLP	0.9100±0.0505	0.9829±0.0076	0.9054±0.0518
	Adversarial	0.9049±0.0487	0.9733±0.0134	0.8993±0.0500

作为一个示例，我们重点关注来自 TCGA-BRCA 数据集的全幻灯片图像（WSI）样本 [22]。我们首先确定贡献最多 WSI 的四家医院，并从每家随机选择 20 张 WSI，限制在满足白人、女性和年龄在 60 至 79 岁的人群标准的患者范围内。从每个选定的 WSI 中，我们以均匀的方式采集大小为 256 × 256 的图像块共 500 个，在 40× 放大倍数下进行。这些初始图像块随后使用 CONCH 基础模型进行过滤：仅保留那些零样本预测标签与 WSI 级别真实标签匹配且置信度得分至少为 0.8 的图像块。经过过滤后，总共剩下 4,029 张高置信度图像块，对应两类疾病：浸润性导管癌（IDC）和浸润性小叶癌（ILC）。

由于随机采样的补丁可能无法始终反映 WSI 级别的诊断（例如，标记为 IDC 的 WSI 可能包含正常组织），我们采用提供强大零样本疾病分类的 CONCH 基础模型 [2] 来过滤这些补丁。我们仅保留那些预测的补丁级别标签与 WSI 标签匹配且预测置信度超过 0.8 的补丁。

过滤后的补丁然后用于从一组具有代表性的病理基础模型中提取特征，包括 ResNet-50 [3], Giga-Path [4], UNI [5], UNI2-H [5], CONCH [2], TITAN [6], MUSK [7], H-Optimus-0 [8], Phikon [9], Phikon-v2 [10] 和 Virchow [11]。得到的补丁级特征使用 t-SNE 进行可视化以评估潜在的领域偏差。

对于模型训练，我们使用 30 个 epoch，批量大小为 64，将对抗强度 λ 设置为 0.5，并使用一个隐藏维度为 512 的投影头。采用五折交叉验证，确保来自同一 WSI

的切片仅出现在一个 fold 中。

对于 TCGA-LUAD [23] 和 TCGA-LUSC [24] 数据集，我们应用相同的 demographic criteria 和 patch extraction strategy。疾病标签分别为肺腺癌（LUAD）和肺鳞状细胞癌（LUSC）。

B. 结果

我们在 TCGA-BRCA、TCGA-LUAD 和 TCGA-LUSC 数据集上进行了实验，从多个角度全面评估了病理基础模型，包括 t-SNE 特征可视化、疾病分类性能以及医院来源分类性能。

图 4展示了 TCGA-BRCA 数据集中标记为 IDC 的补丁特征的 t-SNE 可视化，其中不同颜色的点对应不同的医院来源。值得注意的是，在 Phikon 和 Phikon-v2 的 t-SNE 嵌入中，来自同一医院的补丁表现出明显的聚类模式。这一观察结果表明，这些模型提取的特征保留了大量特定于医院的信息，反映了它们表示中的领域偏差存在。

表 I进一步总结了各种模型的医院来源分类性能。在*方法*列中，“MLP” 表示在对抗训练之前使用原始补丁特征训练 MLP 所获得的分类性能，而 “Adversarial” 代表应用我们的对抗训练框架后的性能。为了参考，我们还包含了 ResNet-50 这个未在病理图像上进行预训练的基础模型。有趣的是，Phikon 和 Phikon-v2 在医院来源分类中达到了最高的准确率和 AUC，这与 t-SNE 图中的强集群现象一致。即使 ResNet-50 基础模型也表现出非平凡的区分医院来源的能力，表明潜在的医院特定线索嵌入在补丁特征中，并且可以在没有病理特定预训练的情况下被利用。相比之下，在我们的管道中用于补丁过滤的 CONCH 模型显示出相对较低的医院来源分类准确率，暗示它对与医院相关的领域偏差具有更强的鲁棒性。经过对抗训练后，所有模型的医院来源分类能力显著降低，AUC 接近 0.5，这对应于随机猜测。这一结果证实了我们的对抗框架有效抑制了学习到表示中的潜在医院特定线索。

表 III展示了评估模型的疾病分类性能。在*方法*列中，“MLP” 表示通过在对抗训练之前使用原始补丁特征训练 MLP 所获得的分类性能，而 “Adversarial” 表示应用我们的对抗训练框架后的性能。由于补丁选择是使用 CONCH 模型进行的，其报告的性能达到 1.0，并不意味着 CONCH 在一般情况下实现了完美的疾病分类。

其他模型也展示了强大的疾病分类性能。例如，ResNet-50 达到了约 0.9 的 AUC，这在一定程度上归因于所选的补丁包含了高度可区分的病理特征。此外，在所有模型中，对抗训练后的疾病分类性能与预训练 MLPs 相当，表明我们的对抗框架有效地抑制了特定医院的信息而没有降低疾病的辨别能力。

我们比较了颜色失真和一种基于 CCA 的方法，使用 UNI 模型。如表 IV所示，颜色失真仅轻微减少了领域偏差。虽然 CCA 通过对齐特征空间有效去除了医院特定的特征，但它极大地损害了疾病分类性能，使其不适合用于此任务。

我们还在对抗训练前后对 UNI 模型在 TCGA-LUAD 和 TCGA-LUSC 数据集上进行了评估（表 V）。训练后，医院来源分类指标显著下降，而疾病分类保持稳定，这展示了我们的对抗框架的鲁棒性和泛化能力。

C. 参数研究

在对抗训练中，参数 λ 控制抑制医院特定特征与保留疾病区分信息之间的权衡。在图 6中，一个小的 λ（例如，0.1）仅弱抑制领域线索，而一个适中的值（例如，1.0）有效移除医院特定信息而不损害疾病分类。过大的 λ（例如，5.0）会降低疾病的性能，表明抑制了有用的特征。根据结果，我们选择一个适度的 $\lambda = 1.0$ 达到平衡。

V. 结论

在这项工作中，我们研究了病理基础模型在应用于病理图像时存在的领域偏差问题。我们建立了一个涵盖 WSI 收集和分割、补丁过滤、MLP 训练以及 t-SNE 可视化的流程，以评估不同模型中的领域偏差严重程度。此外，我们提出了一种轻量级对抗训练框架，该框架利用梯度反转层去除潜在的医院特定特征，同时保持疾病分类能力。在几个 TCGA 数据集上的实验结果表明，我们的流程有效评估了领域偏差，并且我们的对抗训练框架成功消除了潜在的医院特定特征。我们认为，明确建模并移除隐藏的领域偏差对于构建稳健、可泛化和公平的医疗 AI 系统至关重要。我们的工作为此提供了实用蓝图，并为未来向现实世界临床环境扩展打开了大门。

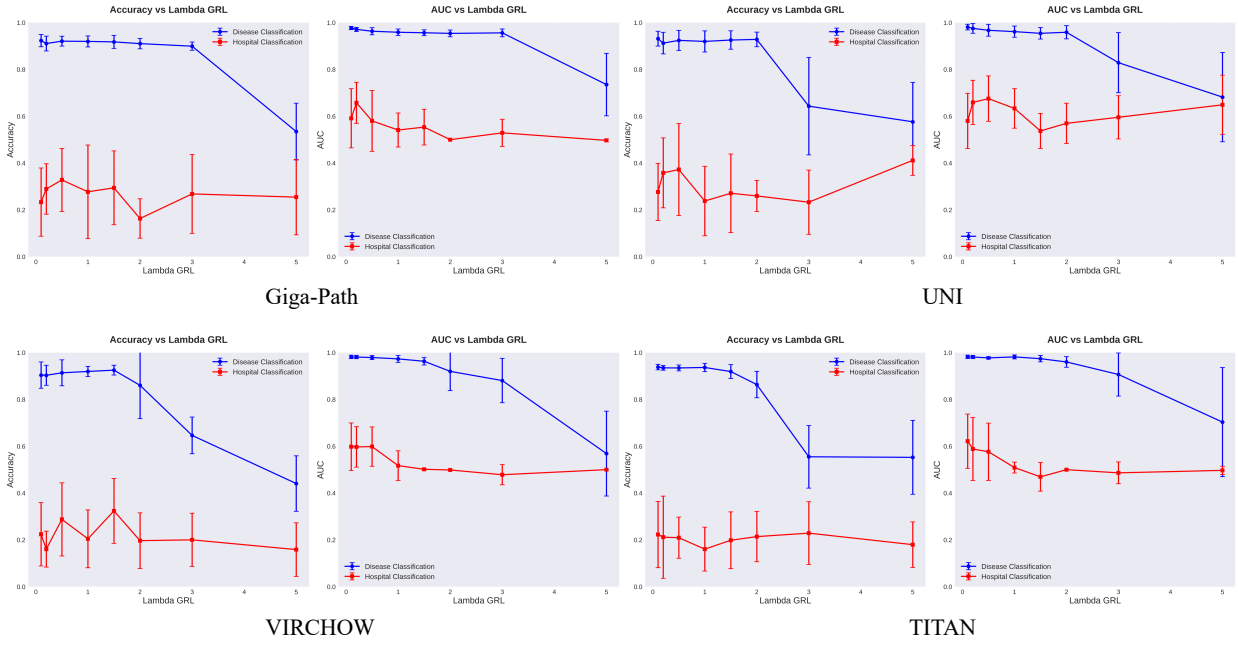

图 6. 疾病分类和医院来源分类在不同 λ 值下的准确率和 AUC 变化。蓝色曲线代表疾病分类，而红色曲线代表医院来源分类。垂直线表示标准差误差棒。

表 IV

不同方法在使用 UNI 模型的医院来源和疾病分类任务上的表现。我们报告了准确率、AUC 和 F1 的平均值 ± 标准差。

任务	**方法**	**准确率 (± 标准差)**	**AUC（± 标准差）**	**F1(± 标准差)**
医院	MLP	0.7836±0.0730	0.9483±0.0317	0.7416±0.0954
	Color Distortion	0.7561±0.0875	0.9142±0.0309	0.7229±0.0875
	CCA	0.2289±0.0419	0.4504±0.0388	0.1440±0.0217
	Adversarial	0.2166±0.1130	0.5788±0.0630	0.0875±0.0377
疾病	MLP	0.9337±0.0338	0.9822±0.0138	0.9291±0.0369
	Color Distortion	0.9120±0.0359	0.9200±0.0309	0.8978±0.0375
	CCA	0.4694±0.0218	0.4704±0.0188	0.4840±0.0197
	Adversarial	0.9278±0.0368	0.9613±0.0363	0.9230±0.0400

表 V

不同方法在 TCGA-LUAD 和 TCGA-LUSC 数据集上使用 UNI 模型进行医院来源和疾病分类任务的评估。报告的指标包括准确率、AUC 和 F1 分数的平均值 ± 标准差。

任务	**方法**	**准确率 (± 标准差)**	**AUC（± 标准差）**	**F1(± 标准差)**
医院	MLP	0.7855±0.0438	0.8883±0.0212	0.7845±0.0247
	Adversarial	0.2246±0.1332	0.5568±0.0450	0.1875±0.0377
疾病	MLP	0.9353±0.0192	0.9748±0.0132	0.9288±0.0369
	Adversarial	0.9324±0.0183	0.9646±0.0114	0.9280±0.0208

致谢

我们衷心感谢 TCGA 提供公开数据集以及各种病理学基础模型的开发者们发布他们的预训练权重。他们对数据和模型共享的贡献对于开展这项研究至关重要。所有患者数据均来自在 dbGaP 数据使用认证下公开的癌症基因组图谱（TCGA）计划。这些数据集已完全去标识化，无需额外的机构审查委员会（IRB）批准。

参考文献

[1] A. Radford et al., "Learning Transferable Visual Models from Natural Language Supervision," in Proceedings of the 38th international conference on machine learning, M. Meila and T. Zhang, Eds., in Proceedings of machine learning research, vol. 139. PMLR, July 2021, pp. 8748 – 8763. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[2] M. Y. Lu et al., "A Visual-Language Foundation Model for Computational Pathology," Nat Med, vol. 30, no. 3, pp. 863 – 874, Mar. 2024, doi: 10.1038/s41591-024-02856-4.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA: IEEE, June 2016, pp. 770 – 778. doi: 10.1109/CVPR.2016.90.

[4] H. Xu et al., "A Whole-Slide Foundation Model for Digital Pathology from Real-World Data," Nature, vol. 630, no. 8015, pp. 181 – 188, June 2024, doi: 10.1038/s41586-024-07441-w.

[5] R. J. Chen et al., "Towards a General-Purpose Foundation Model for Computational Pathology," Nat Med, vol. 30, no. 3, pp. 850 – 862, Mar. 2024, doi: 10.1038/s41591-024-02857-3.

[6] T. Ding et al., "Multimodal Whole Slide Foundation Model for Pathology," Nov. 29, 2024, arXiv: arXiv:2411.19666. doi: 10.48550/arXiv.2411.19666.

[7] J. Xiang et al., "A Vision – Language Foundation Model for Precision Oncology," Nature, vol. 638, no. 8051, pp. 769 – 778, Feb. 2025, doi: 10.1038/s41586-024-08378-w.

[8] C. Saillard et al., H-optimus-0. (2024). [Online]. Available: https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0

[9] A. Filiot et al., "Scaling Self-Supervised Learning for Histopathology with Masked Image Modeling," Sept. 14, 2023, medRxiv. doi: 10.1101/2023.07.21.23292757.

[10] A. Filiot, P. Jacob, A. M. Kain, and C. Saillard, "Phikon-v2, A large and public feature extractor for biomarker prediction," Sept. 13, 2024, arXiv: arXiv:2409.09173. doi: 10.48550/arXiv.2409.09173.

[11] E. Vorontsov et al., "A Foundation Model for Clinical-Grade Computational Pathology and Rare Cancers Detection," Nat Med, vol. 30, no. 10, pp. 2924 – 2935, Oct. 2024, doi: 10.1038/s41591-024-03141-0.

[12] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-Efficient and Weakly Supervised Computational Pathology on Whole-Slide Images," Nature Biomedical Engineering, vol. 5, no. 6, pp. 555 – 570, June 2021, doi: 10.1038/s41551-020-00682-w.

[13] Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine, and J. Zou, "A Visual – Language Foundation Model for Pathology Image Analysis Using Medical Twitter," Nat Med, vol. 29, no. 9, pp. 2307 – 2316, Sept. 2023, doi: 10.1038/s41591-023-02504-3.

[14] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," Journal of Machine Learning Research, vol. 9, no. 86, pp. 2579 – 2605, 2008.

[15] E. D. de Jong, E. Marcus, and J. Teuwen, "Current Pathology Foundation Models Are Unrobust to Medical Center Differences," Feb. 01, 2025, arXiv: arXiv:2501.18055. doi: 10.48550/arXiv.2501.18055.

[16] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in arXiv:2010.11929 [cs], ICLR, June 2021. Accessed: Mar. 14, 2022. [Online]. Available: http://arxiv.org/abs/2010.11929

[17] A. Vaidya et al., "Demographic Bias in Misdiagnosis by Computational Pathology Models," Nature Medicine, vol. 30, no. 4, pp. 1174 – 1190, Apr. 2024, doi: 10.1038/s41591-024-02885-z.

[18] W. Lin, S. Liu, R. Zhu, and L. Wang, "Unveiling Institution-Specific Bias in Pathology Foundation Models: Detriments, Causes, and Potential Solutions," Feb. 24, 2025, arXiv: arXiv:2502.16889. doi: 10.48550/arXiv.2502.16889.

[19] M. Macenko et al., "A Method for Normalizing Histology Slides for Quantitative Analysis," in 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, July 2009, pp. 1107 – 1110. doi: 10.1109/ISBI.2009.5193250.

[20] M. Caron et al., "Emerging Properties in Self-Supervised Vision Transformers," May 24, 2021, arXiv: arXiv:2104.14294. doi: 10.48550/arXiv.2104.14294.

[21] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2022.

[22] The Cancer Genome Atlas Breast Invasive Carcinoma Collection (TCGA-BRCA) (Version 3), The Cancer Imaging Archive, 2016. https://doi.org/10.7937/K9/TCIA.2016.AB2NAZRP

[23] The Cancer Genome Atlas Lung Adenocarcinoma Collection (TCGA-LUAD), The Cancer Imaging Archive. https://doi.org/10.7937/TCIA.2018.p33l1qfd

[24] The Cancer Genome Atlas Lung Squamous Cell Carcinoma Collection (TCGA-LUSC), The Cancer Imaging Archive. https://doi.org/10.7937/TCIA.2018.hp6yaq7i