arXiv:2611.32360v1 [eess.IV] 24 Nov 2026

表征数据集以预测放射影像肺炎检测中的无监督域适应性能

Iftakhar Hossain[a, *], Md. Imran Hasan[a, *], Salahuddin Ahmed[c], Eric D. McCollum[b], Abdullah H. Baqui[d], 以及TaufiqHasan[a, e, †], *IEEE*高级会员

摘要—目的：**基于胸部X线片进行肺炎检测的深度学习模型在跨机构应用时准确率会下降，因此需要采用无监督域适应（UDA）。我们观察到，UDA 性能取决于训练数据集和测试数据集的属性，且性能增益在数据集对之间并不对称。本研究探讨X射线数据集的哪些结构属性最能预测 UDA 性能，以及能否在尝试适应之前对其进行测量。**方法：**我们通过特征的类别可分性来表征五个胸部放射影像数据集。特征取自冻结的 DenseNet-121 的一次前向传播，并以轮廓系数汇总，同时以费舍尔比率作为辅助度量。我们使用域对抗神经网络（DANN）对全部 20 个有向源域—目标域组合进行适应，并评估了未经适应的直接迁移。随后，我们在相同表征上计算了七种已发表的可迁移性估计器，并评估其在预测 UDA 性能方面的显著性。**结果：**我们发现，适应性能取决于两个端点，且二者的作用方向相反。类别可分性更高的目标数据集获得了更高的 AUC 分数（$r = 0.934$，$p = 0.042$），而类别可分性更高的源域则产生较低的 AUC 分数（$r = -0.922$，$p = 0.017$）。可分性的带符号差异与数据集对两个方向之间 AUC 的带符号差异相对应（$r = -0.932$，$p = 0.017$），并在 10 对中的 10 对里识别出较优方向（$p = 0.001$）。在评估目标数据集时，所提出的描述符与表现最佳的已发表估计器相当。**结论：**源域和目标数据集似乎在适应过程中发挥着相反的结构作用。因此，类别可分性较高的数据集更适合作为测试集，但不太适合用于训练适应模型。**意义：**对于基于X射线图像的肺炎检测，所提出的框架可用于判断在低资源环境中，哪个数据集更适合适应目标域，而无需运行完整实验。**

索引词—**非对称迁移、类别可分性、域表征、放射影像肺炎检-测、结构描述符、可迁移性估计、无监督域适应。**

一、引言

肺炎是全球儿童死亡的首要传染病因，2019年导致740,180名五岁以下儿童死亡，占该年龄组所有死亡人数的14% [1]。胸部放射摄影是识别肺炎发作标志——肺泡实变——的既定参考标准，成本低廉、易于推广，也非常适合在现场及流行病学研究中确认病例 [2]。然而，只有经过阅片后，X线片才能用于诊断；可靠的解读需要受过培训的放射科医生或儿科医生按照标准化流程进行，而在肺炎负担集中的低收入地区，这类资源不仅稀缺，分布也极不均衡 [2], [3]。因此，在这些地区现有的便携式放射摄影设备旁，于护理现场部署基于胸部X线片的自动肺炎检测，可发挥至关重要的作用，在没有专家阅片者的情况下进行早期病例识别。

针对该任务的深度学习模型在单一且经过精心整理的数据集上进行训练和测试时，可以达到专家级准确率 [4]。然而，当模型部署到其他机构时，这种准确率并不总能保持。其原因在于域偏移：成像设备、采集方案、患者群体和疾病患病率的差异，使新数据的分布偏离训练数据 [5]。由此导致的性能下降阻碍了安全的跨机构部署，而且在这些模型最能发挥效益的场景中，这种下降最为严重。

在低资源环境下，还存在两个额外限制。本地标注的X射线影像十分稀缺，因为专家阅片所需的人力资源本就短缺，同时用于重新训练的计算预算也十分有限。因此，一种常见策略是

稿件收到日期：XXXX XX, 2026年。

*Iftakhar Hossain 和 Md. Imran Hasan 对本研究贡献相同，并列为共同第一作者。

†通讯作者：Taufiq Hasan。[a]Iftakhar Hossain、Md. Imran Hasan 和 Taufiq Hasan 任职于孟加拉国工程与技术大学（BUET）生物医学工程系及移动健康实验室，地址为孟加拉国达卡 1205（电子邮箱：iftakhar.bme.buet20@gmail.com；imran220.284@gmail.com；taufiq@bme.buet.ac.bd）。[b]Eric D. McCollum 任职于约翰斯·霍普金斯大学医学院 Eudowood 儿科呼吸疾病部全球儿科呼吸科学项目，地址为美国马里兰州巴尔的摩。[c] Salahuddin Ahmed 任职于孟加拉国达卡的 Projahnmo 研究基金会。[d]Abdullah H. Baqui 任职于约翰斯·霍普金斯大学彭博公共卫生学院国际卫生系国际妇幼健康中心，地址为美国马里兰州巴尔的摩。[e]Taufiq Hasan 还兼任约翰斯·霍普金斯大学生物工程创新与设计中心的职务，地址为美国马里兰州巴尔的摩。

将来自不同人群的带标签的公共数据集用作源域，并通过无监督域适应（UDA） [6]–[9]弥合差距。针对每个有向源-目标对分别训练适应模型的成本很高，因此，提前了解哪些配对可能实现良好适应具有重要价值。

适应也具有方向性。从域A适应到域B的效果未必与从域B适应到域A一样好，这种现象称为 非对称迁移 [10]。因此，对迁移进行有用的刻画，不仅应说明适应是否成功，还应解释其成功为何取决于由哪个域提供标签。第二节回顾的现有可迁移性估计器，主要通过目标域来描述迁移，并将源域视为固定的特征提取器。然而，这些估计器没有考虑到源域自身的类别结构可能产生独立的影响。据我们所知，目前尚无已发表的度量能够以一种可用于预判两个域之间首选迁移方向的方式刻画源域。

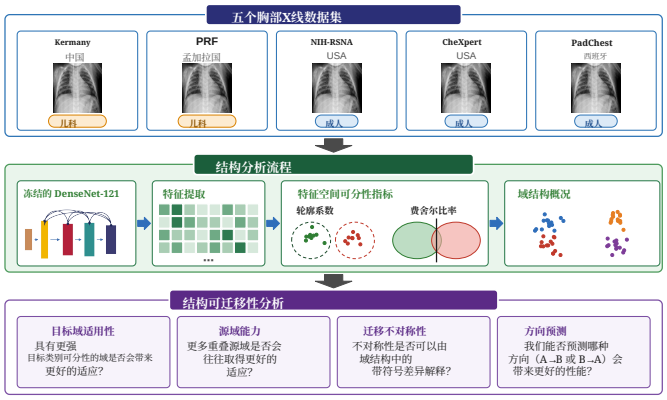

图 1. 结构分析流程概览。五个胸部 X 光数据集经过冻结的 DenseNet-121，以提取倒数第二层特征；随后针对每个数据集计算两个诊断类别的类别可分性，从而得到各域的结构概况，且不依赖于任何特定的源-目标配对。

在本研究中，我们根据胸部X线影像域的特征类别结构对其进行刻画，并在训练任何适应模型之前对这种结构进行度量，再将该刻画与之后观察到的迁移性能联系起来。本研究旨在确定源域和目标域的哪些结构属性与成功适应相关，从而预判两个域之间的迁移是否可能成功，以及可能成功的迁移方向。本研究仅涉及基于胸部X线片的肺炎检测；这些关系是否适用于其他病理类型、任务或模态，尚未经过检验。本研究使用五个数据集，分析具有探索性，因此报告中也将其作为探索性分析。主要贡献如下。

- 我们提出一种可预先计算的胸部X线摄影数据集结构描述符：通过冻结特征提取器进行一次前向处理即可得到一个有界标量，并研究该标量与五个数据集、二十个有向数据集对上的适应性能之间的关系。
- 我们证实，较强的目标类别可分性与更好的适应效果相伴，这与聚类假设一致；此外，我们在相同表示上，将该描述符与七种已发表的可迁移性估计器进行基准比较。
- 我们观察到，源域类别结构具有信息价值，能够相反的方向：重叠更多的源域迁移效果更好。目前没有已发表的估计器以这种方式描述源域。
- 我们表明，迁移不对称性遵循可分性的带符号差异，该差异能在适应过程中识别每一对域中效果更好的方向。仅目标领域评分无法表达这一点：每个域只有一个分值，因此无论迁移方向如何，都会预测相同的结果。
- 我们验证了这些关系在两种机制、两个独立训练的特征提取器和两个可分性度量指标中符号一致，表明这些关系源于领域属性，而非某一种算法、表征或度量指标的产物。

II. 相关工作

A. 可迁移性估计

可迁移性估计无需运行完整的训练过程，即可预测模型的迁移效果。最早的实用度量指标通过预训练模型的一次前向传播对目标数据集进行评分。H分数 [11] 将可迁移性视为类间特征协方差与总特征协方差之比；LEEP [12] 根据源分类器输出分布，计算目标标签的对数期望经验预测值；LogME [13] 则最大化基于冻结特征拟合的线性模型的证据。后续分析表明，H分数的局限性主要来自高维条件下的协方差估计，并提出了一种收缩估计量，使其恢复了很大一部分准确性 [14]。

自 2021 年以来提出的四种度量从不同方向对这一思路加以改进。NLEEP [15] 将 LEEP 中的源分类器替换为基于目标特征拟合的高斯混合模型，从而完全无需源分类头。Tran- sRate [16] 通过编码率替代量估计特征与标签之间的互信息。PACTran [17]推导微调风险的 PAC-贝叶斯界，并给出高斯和狄利克雷变体。SFDA [18]将特征嵌入正则化费舍尔空间，并加入一个扰动简单样本的自挑战步骤。GBC [19] 则从目标类别的几何结构入手处理同一问题：将每个类别建模为预训练网络特征空间中的高斯分布，并以成对的巴塔恰里亚系数衡量类条件分布之间的重叠程度；重叠越低，预期可迁移性越高。其依据是，该系数能够界定相应问题的贝叶斯错误率。

尽管这些度量的构造方式各不相同，但它们有一个共同前提：在给定表征下，目标类别的可分性决定了模型迁移到该域的效果。适应理论中也独立出现了同一原理：聚类假设认为，当目标特征形成判别性簇，使决策边界能够穿过低密度区域时，适应更容易成功 [20], [21]。因此，目标侧可分性作为迁移预测指标已得到确立；

我们在此将其视为检验所提出描述符的参照，而非一项发现；为此，我们在当前表征上计算了其中七个估计量（IV-B节）。

这些度量指标都无法表征源域。在每种情况下，源域仅作为冻结特征提取器参与其中，而得分仅取决于目标域。没有一种方法将源域自身的类别结构视为独立因素，因此也无法得出可用于预测两个域之间更佳迁移方向的量。另一个局限涉及评估，而非方法构想。这些度量指标主要是在自然图像基准上开发并验证的；对于医学影像领域，其类别边界依据诊断而非物体，且数据集规模较小、采集方式也更加多样，这些指标的表现尚未得到系统性研究。

B. 胸部*X*光片域适应

胸部X线片域适应一直是一个活跃的研究领域。早期研究采用监督式适应，即在有标签源数据集之外，还提供少量有标签目标数据。Feng等提出了一种用于肺炎诊断的深度监督式域适应网络，该网络在对齐源域和目标域特征分布的同时，在源域专用分类器和目标域专用分类器之间共享特征提取层 [22]。沿用类似策略，研究人员将监督式对抗适应应用于来自不同国家的胸部X线摄影数据集之间的跨人群偏移，同样依赖有标签的目标样本 [23]。这一方法基于 Ganin等提出的域对抗神经网络（DANN）框架 [24]，该框架训练一个共享特征提取器，使其既能有效区分标签预测任务中的类别，又不受域来源影响。

第二类方法利用生成对抗网络在不同域之间转换图像。CX-DaGAN采用图像到图像转换网络，并通过相似性驱动的源域选择步骤进行引导，从而合成类似目标域的源图像，用于在小型目标数据集上进行肺炎诊断 [25]。C2ADA在此基础上引入了循环一致性对抗模块，在没有配对图像的情况下对齐源域和目标特征 [26]。最近，完全无监督的方法进一步摆脱了对任何带标签或配对目标数据的需求。TUNA-Net仅使用无标签目标图像，将成人胸部X线片转换为类似儿科X光片的图像，用于肺炎检测，并挽回了因域偏移而损失的大部分性能 [27]。

在这一发展过程中，源域要么被视为固定不变，要么被视为被动的特征提取器；源数据的选择通常不会参考源域本身任何经度量的属性。CX-DaGAN 是一个部分例外，它会根据与目标域的相似性来选择源图像，但并未对整个源域进行表征。医学图像分析领域的大多数域适应研究都是如此。本研究旨在弥补这一空白，对源域和目标域进行平等的描述，并将这两种描述与观察到的迁移性能联系起来。

III. 方法

A. 数据集

本文采用了来自不同机构、地理区域和患者群体的五个胸部放射影像数据集。所有数据集均被设定为二分类，包括正常 和肺炎；对于包含其他疾病标签的数据集，仅保留相关类别。表I总结了各类别的分布情况，后续所有表格中的数据集均按照IV-B节所报告的可分性进行排序。

Kermany 是一个公开的儿科数据集，采集自一家儿科医疗中心的常规诊疗过程中。发布前，专家医生对每张X光片的质量和诊断结果进行评估 [28]。其中包含1,583张正常影像和4,273张肺炎影像。

NIH-RSNA 基于NIH ChestX-ray14数据集构建 [29]，其中的肺炎标注由放射科医生为RSNA肺炎检测挑战赛进行了补充 [30]。研究使用了一个平衡子集，其中包含3,350张正常影像和3,350张肺炎影像。

PadChest 是一个来自西班牙医疗机构的大型标注数据集 [31]。从中提取了一个平衡子集，包含3,350张正常影像和3,350张肺炎影像。

CheXpert是一个基于斯坦福大学常规临床影像构建的大型数据集 [32]。由于原始发布版本包含多种胸部病变及不确定性标签，因此构建了一个平衡的二分类子集，其中包含3,350张正常影像和3,350张肺炎影像。

PRF 是一个私有儿科数据集，采集自孟加拉国农村锡尔赫特地区 Projahnmo 研究基金会研究区域内的门诊诊所，属于一项肺炎球菌结合疫苗有效性研究 [3]。对疑似肺炎的3至35个月龄儿童进行仰卧位胸部前后位X线摄影，使用便携式模拟设备拍摄影像，并通过计算机X线摄影暗盒读取器将图像数字化；放射科医生和儿科医生组成的专家组按照流行病学研究中胸部X线片判读的标准化方法对影像进行判读 [2]。质量控制过程中剔除了一部分影像，最终保留2,777张正常影像和1,999张肺炎影像。

数据集划分与标签处理。 每个数据集仅划分一次，分为训练部分和互不重叠的留出评估集；无论该数据集承担何种用途，均采用这一划分。Kermany 数据集的训练部分包含5,232张X线片（1,349张正常、3,883张肺炎）；NIH-RSNA、PadChest和CheXpert数据集各包含6,000张（每类3,000张）；PRF数据集包含4,107张（2,388张正常、1,719张肺炎）。Kermany采用随数据集发布的官方划分。第III-C节中的所有结构描述符均仅根据训练部分计算，因此，在训练任何适应模型之前，只使用当时可用的数据来描述一个域。

当一个数据集作为目标域时，其标签会被丢弃。第III-E节所述的适应过程会在没有标签的情况下查看目标域训练部分中的图像；目标域标签

表I按表 III 的可分性排序的数据集汇总

数据集	正常	肺炎	训练	总数
Kermany（儿科）	1,583	4,273	5,856	5,232
NIH-RSNA（成人）	3,350	3,350	6,700	6,000
PadChest（成人）	3,350	3,350	6,700	6,000
CheXpert（成人）	3,350	3,350	6,700	6,000
PRF（儿科）	2,777	1,999	4,776	4,107

仅使用一次，即在不重叠的留出部分上计算所报告的曲线下面积。任何数据集留出部分中的标签都不会参与任何一种机制下的训练。

B. 特征表示

受像素空间统计量将与类别相关的结构与扫描仪、曝光和定位等采集干扰因素纠缠在一起这一观察的启发，结构描述符并非在原始像素上计算，而是在固定的已学习特征表示上计算。设 D 为五个数据集之一，由 N幅灰度射线照片 $\{I_i\}_{i=1}^{N}$ 及二元标签 $y_i \in \{0,1\}$组成。主提取器是在 NIH ChestX-ray14 集合上针对胸部射线照片分类训练的 DenseNet-121，该数据集即 CheXNet [4] 最初接受训练所用的数据集；权重为随 TorchXRayVision 库 [33] 分发的 densenet121-res224-nih。该网络定义了一个映射

$$\phi : \mathcal{I} \to \mathbb{R}^{1024}, \tag{1}$$

从图像空间 $\mathcal{I}$ 到倒数第二层在最终整流和全局平均池化之后取得的1024维激活。每幅图像在推理模式下仅通过网络一次，且不进行梯度更新，得到$\mathbf{x}_i = \phi(I_i)$。每个数据集只需一次前向传播，无需训练，这比针对每个有向数据集对训练一个适应模型要便宜得多。

需要注意的是，特征提取器的权重是在 NIH ChestX-ray14 上训练的，而 NIH ChestX-ray14 是五个受评估数据集之一，也是源域。相应地，第 IV-E 节使用独立训练的 DenseNet-121（densenet121-res224-chex，在 CheXpert 上训练）重复了完整分析，并报告了比较结果。

随后，使用基于该数据集两个类别共同拟合得到的统计量，对每个数据集的 1024 个维度分别进行标准化，使所得向量具有零均值和单位方差。整体流程如图 1 所示。

C. 结构描述符

数据集在该特征空间中的类别结构由两个标量概括。令$\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N} \subset \mathbb{R}^d$, d =1024，表示数据集D的标准化特征向量集合，并将其划分为$\mathbf{X}_0$ 和$\mathbf{X}_1$ 两个子集，分别对应正常类和肺炎类，大小为 n_0 和 n_1。全文中，粗体小写字母表示向量，普通斜体字母表示标量。

轮廓系数。 如果一个数据集中的点与同类别其他点的距离小于与相反类别点的距离，则该数据集的类别分离良好。对于类别为$\mathbf{x}_i$ 的点 k，设 $a(i)$ 为其到类别 k 中其他点的平均欧几里得距离， $b(i)$ 表示其到相反类别各点的平均欧几里得距离。$\mathbf{x}_i$ 的轮廓系数为

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{2}$$

数据集级别的轮廓系数是其平均值，

$$s(D) = \frac{1}{N}\sum_{i=1}^{N} s(i), \tag{3}$$

值越接近 1 表示分离效果越强 [34]。为便于计算，使用固定随机种子，从训练部分中抽取最多 3,000 个点构成随机子样本，以估算 $s(D)$，因为底层的成对距离计算为 $\mathcal{O}(N^2)$。

费舍尔比率。 作为基于完全不同原理构建的辅助度量，费舍尔比率通过方差比而非距离比来表示分离程度 [35]。本文采用多变量、按维度求和的形式。设$\mathbf{c}_0$,$\mathbf{c}_1 \in \mathbb{R}^d$ 为类别质心，则类间项为它们之间的平方欧几里得距离，

$$B(D) = \|\mathbf{c}_0 - \mathbf{c}_1\|_2^2, \tag{4}$$

类内项则为两个类别各维度无偏方差之和，

$$W(D) = \sum_{j=1}^{d} \mathrm{Var}(\mathbf{X}_0)_j + \sum_{j=1}^{d} \mathrm{Var}(\mathbf{X}_1)_j. \tag{5}$$

因此，费舍尔比率为

$$F(D) = \frac{B(D)}{W(D) + \epsilon}, \tag{6}$$

其中 $\epsilon = 10^{-12}$ 用于确保数值稳定性，数值越大表示可分辨性越强。这是一种计算效率较高的经典类间/类内方差比多变量推广形式，采用各维度方差之和，而非完整的 1024×1024 协方差矩阵。它不应与线性判别分析（LDA）中的标量费舍尔判别函数混淆。

全文始终采用轮廓系数作为主要描述符，并同时报告费舍尔比率。轮廓系数的取值范围为 $[-1, 1]$ ，且具有无尺度特性：它既不取决于特征幅值，也不取决于维度或样本数量；在比较规模和采集特征各不相同的五个数据集时，这一点十分重要。费舍尔比率没有上界，度量的是不同的几何量，因此二者的一致性表明，相关发现并非任一构造方式造成的伪象。如 IV-B节所述，二者对五个数据集的排序完全一致，我们在第三-F节中利用了这一事实。

D. 参考可迁移性估计量

为确认目标侧关系是域本身的属性，我们在相同的冻结表示上计算了七种已发表估计器，并将结果与所提出的描述符一并报告：H分数 [11] 及其收缩变体[14],LogME [13], NLEEP [15],TransRate [16], PACTran 的高斯形式 [17]和 SFDA [18], , 以及 GBC[19]。这些方法均以特征和

目标标签为唯一依据，因此都可以与所提出的描述符直接比较。LEEP [12]和 PACTran 的狄利克雷变体 [17] 不予纳入，因为二者都需要 $\theta(\mathbf{x}) = P(\text{源标签} \mid \mathbf{x})$，即定义在源任务标签集上的分类分布；此处唯一可用的分类头是预训练的多标签病理分类器，其针对各病理类别的 Sigmoid 输出并不构成这样的分布。

GBC 按照其作者推荐的配置计算，并在此完整描述，因为它与所提出的描述符在理念上最为接近。未标准化的向量 $\tilde{\mathbf{x}}_i = \boldsymbol{\phi}(I_i)$被投影到前 $d' = 64$个主成分上，这些主成分保留了五个数据集中 97.6% 至 98.4% 的方差；每个类别 $k \in \{0,1\}$ 均建模为球形高斯分布 $\mathcal{N}(\boldsymbol{\mu}_k,\ \sigma_k^2\mathbf{I}_{d'})$，其中 $\boldsymbol{\mu}_k$ 为类别均值，σ_k^2 为各分量无偏方差的均值。给定 $m = (\sigma_0^2 + \sigma_1^2)/2$，类条件分布之间的巴塔恰里亚距离为

$$D_B(0,1) = \frac{\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|_2^2}{8\,m} + \frac{d'}{2}\ln\frac{m}{\sigma_0\sigma_1}, \tag{7}$$

数据集 D 的得分通过对所有有序类别对求和得到，

$$\mathrm{GBC}(D) = \sum_{k \neq l} D_B(k,l) = 2\,D_B(0,1), \tag{8}$$

值越大表示可分性越强。报告的得分采用巴塔恰里亚距离形式，而非 [19]中使用的指数化系数 BC $= \exp(-D_B)$；二者单调相关，给出的排序也完全相同，但在这五个数据集中，该系数的取值范围跨越约八个数量级，因此线性比较缺乏意义。参考估计量从未与所提出的描述符结合使用，因此二者之间的一致性是一种独立验证，而非循环验证。

E. 迁移情形

对五个数据集中的每一对有序组合，均在两种机制下进行评估。采用 DANN 进行适应是本研究的主要机制；将未经适应的直接迁移作为次要机制进行评估，以便检验有无适应过程时，任何结构与性能之间的关系是否保持一致。

采用 DANN 进行适应。本文采用 Ganin等人 [24] 提出的域对抗神经网络原始形式，不使用辅助损失或额外的架构设计，以便将观察到的关系归因于域本身，而非更复杂方法的机制。架构如图 2 所示。训练了三个组件

联合训练：共享特征提取器 G_f、标签分类器 G_y 和域判别器 G_d，其参数分别为 θ_f、θ_y 和 θ_d。训练特征提取器，使其生成的表征能够区分任务，同时不包含域来源信息。记标签预测损失为 $\mathcal{L}_y$，域分类损失为 $\mathcal{L}_d$；给定大小分别为 n_s 和 n_t的有标签源批次 $\mathcal{D}_s$和无标签目标批次 $\mathcal{D}_t$，联合

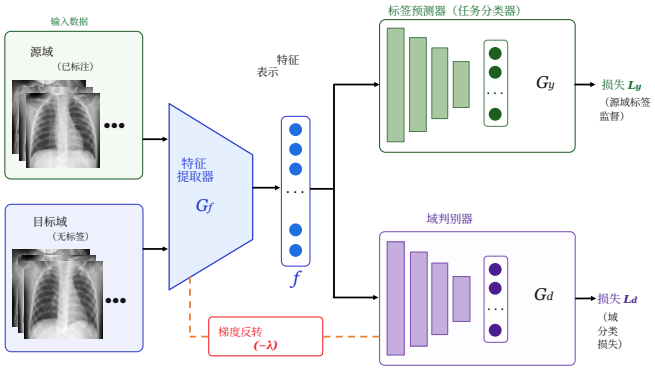

图 2. 域对抗神经网络架构。共享特征提取器 $\boldsymbol{G_f}$ 将带标签的源域数据和无标签目标批次映射到公共特征空间 $\mathbf{f}$。标签预测器 $\boldsymbol{G_y}$ 仅使用源数据进行训练（损失 $\boldsymbol{\mathcal{L}_y}$）；域判别器 $\boldsymbol{G_d}$ 则使用两个域的数据进行训练（损失 $\boldsymbol{\mathcal{L}_d}$）。位于 $\boldsymbol{G_f}$ 与 $\boldsymbol{G_d}$之间的梯度反转层在前向传播时原样传递激活值，并在反向传播时将梯度取反，从而推动 $\boldsymbol{G_f}$ 学习域不变特征。

目标函数为

$$\mathcal{L}(\theta_f,\theta_y,\theta_d) = \frac{1}{n_s}\sum_{i\in\mathcal{D}_s}\mathcal{L}_y^{(i)} - \lambda\left(\frac{1}{n_s+n_t}\sum_{i\in\mathcal{D}_s\cup\mathcal{D}_t}\mathcal{L}_d^{(i)}\right), \tag{9}$$

其中，$\lambda > 0$ 用于平衡这两项。由于目标标签被保留不用（第III-A节），标签分类器仅使用源数据进行训练；域判别器则接收两个域的数据。直接优化 θ_f 以最大化 $\mathcal{L}_d$ 需要另设一个方向相反的目标，因此 DANN 改为在 G_f 与 G_d之间插入一个梯度反转层：前向传播时执行恒等映射，反向传播时将进入 G_f 的梯度乘以 $-\lambda$。这样，网络只需经过一次前向传播、一次反向传播并使用一个优化器即可完成训练，同时促使 G_f 远离那些能够用于域判别的特征。遵循 Ganin等人 [24]的方法，λ 在训练过程中从 0 逐渐增加至 1，具体如下：

$$\lambda(p) = \frac{2}{1+\exp(-10p)} - 1, \tag{10}$$

其中，$p \in [0,1]$ 表示已完成的训练迭代总数比例，因此在标签分类器仍处于条件不佳状态时，域项会受到抑制。完整流程见表 II。

直接迁移。在单个源数据集上训练分类器，并将其未经修改地应用于每个目标域，最小化

$$\mathcal{L}_{\text{direct}}(\theta_f,\theta_y) = \frac{1}{n_s}\sum_{i\in\mathcal{D}_s}\mathcal{L}_y^{(i)}, \tag{11}$$

这相当于仅在源域上进行经验风险最小化，训练过程中不涉及目标域。该机制代表直接部署源域训练模型，并作为评估未经适应时域偏移影响的参考点。

在两种机制下，通过计算预测的肺炎概率与真实标签之间ROC曲线下面积，在目标数据集的留出部分上评估所得分类器，从而得到该配对所报告的有向AUC。对于每个适应后的配对，所报告的数值

表II适应阶段的训练方案

组件	设置
特征提取器 G_f	ResNet-18，基于ImageNet预训练，移除最后一层（512 维）
标签预测器 G_y	$512 \to 256$，批归一化、ReLU，随机失活 0.5，$256 \to 2$
域判别器 G_d	相同的头部，$256 \to 2$
梯度反转	$\lambda(p)$ 每（10）
输入尺寸	224×224，ImageNet 归一化
批次大小	每个域 32
优化器	Adam，权重衰减 10^{-5}
学习率	10^{-4}
学习率调度	平台期降低，因子 0.5，耐心值 3
梯度裁剪	最大范数 5.0
训练轮数	30，提前停止耐心值 7
类别平衡（源域）	加权随机采样
随机种子	42
报告的曲线下面积	目标评估中的最佳值
	训练期间的部分

是在训练过程中目标评估部分上达到的最高曲线下面积。该数值代表适应在该配对上所能达到的最佳性能，并消除了停止规则这一干扰变量。因此，当问题是领域结构是否决定适应能够达到的效果时，它是合适的衡量指标。但它并不是对部署模型在无法获取目标标签的情况下所能获得性能的估计。

F. 迁移潜力

为用一个量表示源域与目标域之间的结构关系，有向数据集对（$S \to T$）的迁移潜力定义为两个域之间轮廓系数的差值，

$$\mathrm{TP}(S \to T) = s(T) - s(S). \tag{12}$$

正值表示目标域比源域更易分离，负值则表示相反情况。

该指数仅基于轮廓系数，而不是同时基于两种描述符构建。如 IV-B节所报告的，费舍尔比率和轮廓系数对五个数据集的排序完全一致（$\rho = 1.00$），并且几乎完全共线（$r = 0.999$）。因此，在该数据集集合中，费舍尔比率没有提供轮廓系数之外的任何信息，将二者合成为一个指标不会增加预测信息，反而需要进行任意加权。在两个可互换的度量中，我们选择有界且无尺度的那个。由于（12）是原始轮廓系数之差，因此它仅取决于所涉及的两个域，而与所比较的数据集集合无关。

G. 统计分析

结构描述符与迁移性能相关，迁移性能以曲线下面积进行度量，并采用皮尔逊相关进行分析。由于数据集数量较少，除每个皮尔逊系数外，还报告斯皮尔曼秩相关，因为该方法不假设任何线性形式，并且对任一单独数据集不那么敏感。

每个问题均在其适当的分析单位上进行评估。在五个数据集中检验目标域和源域角色使用四个配对数据集上的平均曲线下面积。在十个无序对上，通过将带符号的结构差异与不同方向之间曲线下面积的带符号差异进行相关分析，检验方向效应。若将全部二十个有序对视为相互独立，则每个数据集会被重复计算四次，并低估 p值；采用这一选择可避免这种情况。

由于仅有 5 个观测值，渐近 p 值并不可靠，因此采用置换检验评估显著性。数据集层面的分析针对两个系数，使用所有 5! 种 $= 120$ 分配构成的精确置换分布。对于配对层面分析，由于各配对共享数据集，因此在固定曲线下面积值的情况下，将五个结构值在各数据集间置换。p 值为 120 种分配中达到或超过观测到的相关性绝对值的比例，因此双侧检验的下限为 $1/120 \approx 0.008$。

未对多重性进行校正，因此这些结果应视为探索性结果，而非验证性结果。这是规模上的限制，而非方法上的限制。对于五个域，若针对这些关系采用邦费罗尼阈值，则需要 $p < 0.002$，低于 0.008 的置换下限，因此在当前样本量下无法获得校正后的显著性。因此，该分析旨在生成假设，重点关注效应的方向和一致性，而非统计显著性。

IV. 结果

A. 两种方案下的迁移性能

在二十个有序的源域-目标域对中，采用 DANN 的适应方法取得了 0.866 的平均曲线下面积，而直接迁移的平均曲线下面积为 0.830，相较于直接迁移，适应方法的相对提升为 4.3%。图3展示了两种机制下完整的有向AUC值，以及每个数据集作为源域和目标域时的平均曲线下面积。图4总结了各目标域和各源域-目标域对的数值。

从图3可以看出，在适应机制下，平均输入 AUC的范围从 PRF 的 0.762 到 NIH-RSNA 的 0.968，跨度为 0.207；而平均传出AUC的范围从 Kermany 的 0.816 到 CheXpert 的 0.893，跨度为 0.077。因此，目标域角色造成了更大比例的差异，而源域角色造成的差异较小，但如下文所示，这种差异是一致的。两个极端在两种角色之间也发生了反转：Kermany 是表现最佳的目标域之一，却是表现最差的源域；而 PRF 是表现最差的目标域，却是表现最佳的源域之一。

B. 域的结构表征

表 III 报告了每个数据集在 1024 维表征上的两个所提出的描述符和七个参考估计量。相应的二维 UMAP [36] 嵌入按类别着色，见图5。

从表 III 可以清楚看出，九种度量均以相同顺序排列数据集：Kermany、NIH-RSNA、PadChest、CheXpert、PRF。两个所提出描述符之间的吻合几乎是完全的：Silhouette Score 与 Fisher Ratio 之间的 Spearman $\rho = 1.00$ 和 Pearson $r = 0.999$，以及 Silhouette Score 与GBC之间的 $\rho = 1.00$ 和 $r = 0.975$。因此，该表征中的可分性并非

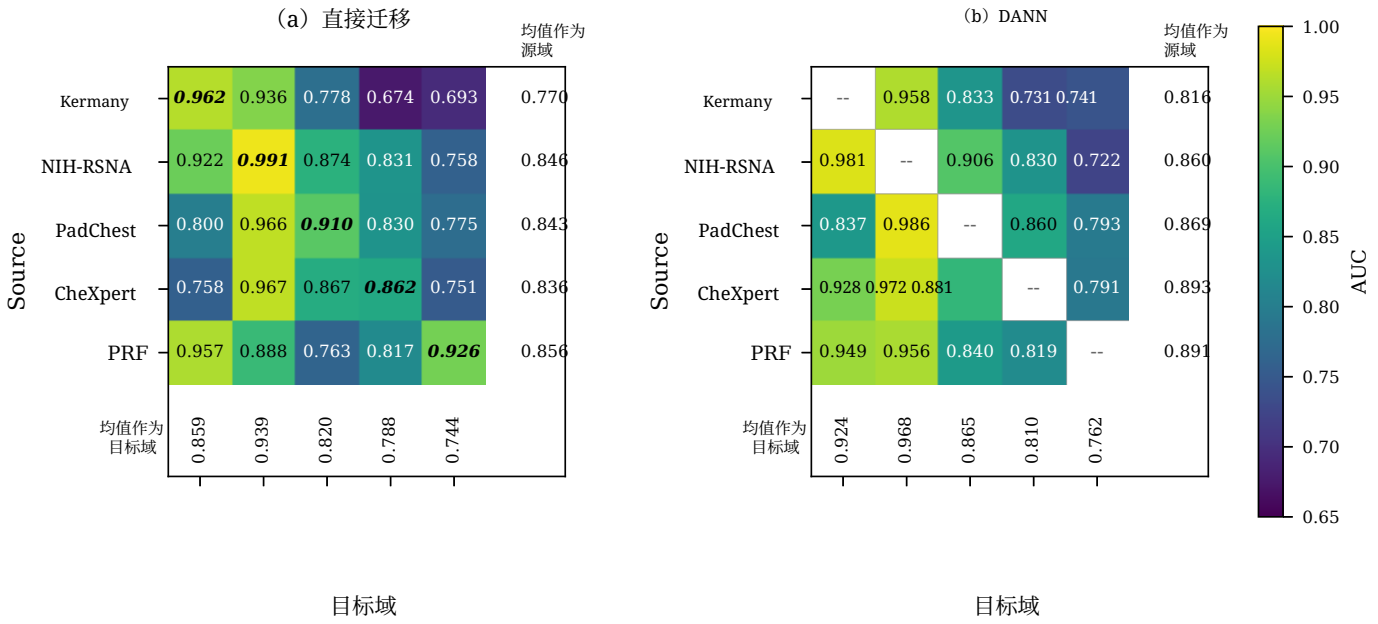

图3。所有源域-目标域对在 (a) 直接迁移（源域训练模型未经适应，直接在目标域上评估）和 (b) DANN 下的有向AUC。行表示源域，列表示目标域，二者均按可分性排序；非对角单元格显示二十个跨域对。在 (a) 中，对角线（粗斜体）报告域内曲线下面积，即模型在同一数据集上训练和评估所得的结果，作为参考点，而非迁移结果。在 (b) 中，对角线未定义，因为 DANN 要求域彼此不同。每个面板右侧的色带显示各数据集作为源域时的平均AUC，底部色带显示其作为目标域时的平均AUC，均仅根据跨域对计算。两个面板共用同一颜色刻度。

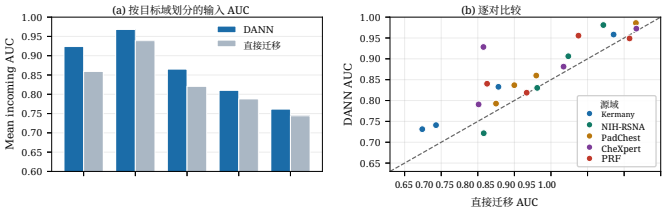

图4. 不同机制下的迁移性能。(a) 两种机制下每个目标数据集的平均输入AUC。(b) 每对源域与目标域的 AUC；位于恒等线上方的点表示适应模型在相应数据对上的取值更高。点的颜色表示源数据集。

某一种构造的产物：基于距离的度量、基于方差的度量和高斯重叠度量完全一致。正如第 三-F节所述，这种冗余为仅基于 Silhouette Score 构建式（12）的迁移潜力提供了依据。尽管如此，下面的分析始终同时报告这两个描述符，并且任何结论都不取决于采用哪一个。

C. 结构-性能关系

我们从三个角度、以三个分析单位考察结构与迁移之间的关系：仅考虑目标域的作用、仅考虑源域的作用，以及二者之间的带符号差异。

目标数据集。表IV将各数据集的可分性与其平均输入AUC，以及在相同表征上计算的七种参考估计量联系起来。从该表可以看出，类别区分度更高的目标域在适应后获得的 AUC 也更高（$r = 0.934$、$p = 0.042$ 对应轮廓系数；$r = 0.932$、$p = 0.042$ 对应费舍尔比率）。所有参考估计量都呈现相同的正相关关系，而本文提出的描述符也位居表现最强者之列：NLEEP 略高，为 $r = 0.966$；收缩H分数与之相同，为 $r = 0.934$；其余值介于 0。840 和 0。908 之间。十种度量中有九种（包括两个提出的描述符）得到 $\rho = 0.900$，与 AUC 排序相比，仅有 Kermany 和 NIH-RSNA 这两个相邻项交换了位置；未收缩的 H分数则得到 $\rho = 0.800$。因此，所有考察的度量都再现了目标域侧的关系，这正是预期结果，也是我们将其视为参考基准而非本文贡献的原因。

源数据集。源域的关系恰恰相反，而且尚无已发表的估计器对此进行研究。表V列出了各数据集的可分性与其平均传出AUC之间的关系。从表中可以看到，类别重叠程度更高的源域与更强的迁移相关（$r = -0.922$, $p = 0.017$ 轮廓系数；$r = -0.919$, $p = 0.017$费舍尔比率）。在适应机制下，这是本研究中唯一一种在两种描述符上均达到 0.05 显著性水平的关系。在直接迁移下，这种关系较弱，且未达到该显著性水平（$r = -0.706$, $p = 0.133$），但其方向仍呈现一致趋势。

两个端点。由于这两个角色分别与相反意义上的可分性相关，因此，作为有利目标域的数据集往往是较弱的源域。这种对立构成了迁移不对称性的结构基础。表V报告了两项配对层面分析。两个方向之间带符号差异的可分性与AUC的带符号差异密切对应，在两种机制下以及两个描述符中均呈现相同趋势（适应条件下为$r = -0.932$、$p = 0.017$）：迁移更有利于从重叠程度更高的

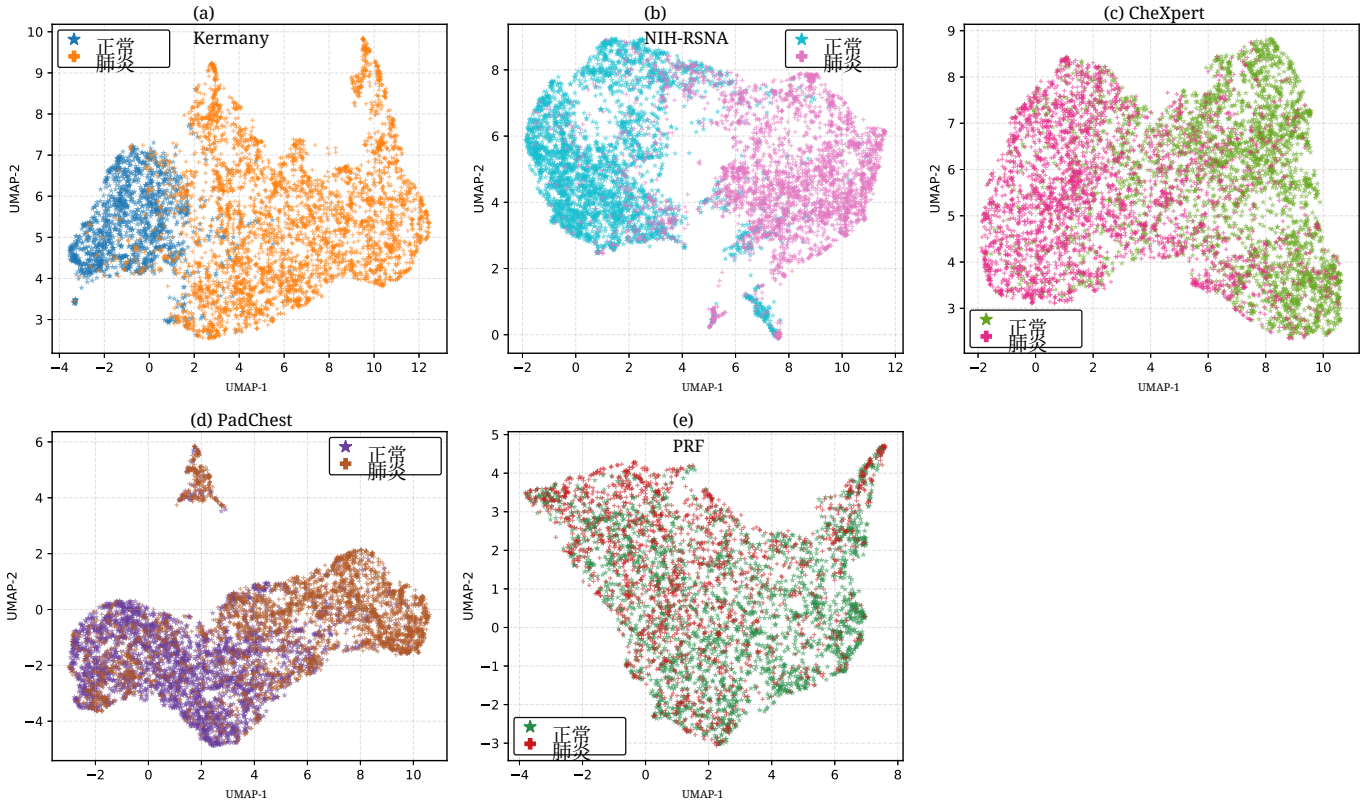

图5. 每个数据集的 DenseNet-121 特征表示的二维 UMAP 嵌入，按类别标签着色。轮廓系数较高的数据集，其类别分离得更为明显。

表 III

主提取器下各数据集的结构描述符与参考可迁移性估计量

每一列中，数值越高表示类别可分性越强。数据集按轮廓系数排序；所有九种度量得出的排序均相同。

	提出的描述符		参考估计量						
数据集	轮廓系数	费舍尔比率	GBC	LogME	H分数 (s)	TransRate	NLEEP	PACTran-G	SFDA
Kermany	0.1535	0.3486	17.9147	+0.2347	0.8637	60.6346	−0.3230	−4.9686	−0.2554
NIH-RSNA	0.1265	0.2891	11.3861	−0.0693	0.7848	43.2202	−0.3432	−4.3534	−0.3702
PadChest	0.0671	0.1412	4.8135	−0.3791	0.5890	30.5928	−0.5162	−9.1413	−0.4502
CheXpert	0.0326	0.0731	3.4402	−0.5465	0.4330	20.1721	−0.6136	−13.0551	−0.5489
PRF	0.0086	0.0206	0.4197	−0.6024	0.4024	14.8996	−0.6682	−10.5807	−0.5858

表IV目标结构与参考估计量相对于平均输入 AUC 的关系（$n = 5$；精确置换检验 p）

	DANN				直接迁移			
度量	r	p	ρ	p_ρ	r	p	ρ	p_ρ
轮廓系数	+0.934	0.0417	+0.900	0.0833	+0.850	0.0583	+0.900	0.0833
费舍尔比率	+0.932	0.0417	+0.900	0.0833	+0.852	0.0583	+0.900	0.0833
GBC [19]	+0.840	0.0667	+0.900	0.0833	+0.735	0.1500	+0.900	0.0833
LogME [13]	+0.853	0.0583	+0.900	0.0833	+0.747	0.1417	+0.900	0.0833
H分数 [11]	+0.908	0.0583	+0.800	0.1333	+0.821	0.1083	+0.800	0.1333
收缩H分数 [14]	+0.934	0.0417	+0.900	0.0833	+0.850	0.0583	+0.900	0.0833
TransRate [16]	+0.854	0.0750	+0.900	0.0833	+0.738	0.1500	+0.900	0.0833
NLEEP [15]	+0.966	0.0250	+0.900	0.0833	+0.901	0.0333	+0.900	0.0833
PACTran-G [17]	+0.889	0.0833	+0.900	0.0833	+0.846	0.0750	+0.900	0.0833
SFDA [18]	+0.874	0.0583	+0.900	0.0833	+0.758	0.1250	+0.900	0.0833

域迁移到可分性更高的域。相比之下，差异的绝对大小在任何一种机制下都不包含信号（均为 $|r| \le 0.33$，均为 $p \ge 0.28$）。结构距离较远的域并不会比结构距离较近的域产生更差的迁移效果。真正相关的是两个域中哪一个更

表V源域结构与平均传出AUC（$n = 5$；精确置换检验 p）。没有参考估计量能够描述源域角色。

	DANN				直接迁移			
度量 r	p		ρ	p_ρ	r	p	ρ	p_ρ
轮廓系数	−0.922	0.0167	−0.900	0.0833	−0.706	0.1333	−0.600	0.3500
费舍尔比率	−0.919	0.0167	−0.900	0.0833	−0.706	0.1250	−0.600	0.3500

可分离，且并非差值的大小。

最后，表VII将式（12）的迁移潜力与全部二十个有序对的有向AUC联系起来，图6展示了这种关系。作为一个兼顾两个端点、易于解释的单一指标，它在两种机制下均与迁移性能呈正相关；在适应条件下，它能解释约一半的方差（$R^2 = 0.521$）。

这些关系并非由任何单一数据集驱动。依次省略每个数据集后，每个相关性仍保持其符号不变，

表VI配对层面分析（$n = 10$无序对；数据集标签置换 p）。带符号差异为 $s(A) - s(B)$，相对于 **曲线下面积$(A \to B)$ − 曲线下面积$(B \to A)$**。

描述符分析		DANN		直接迁移	
		r	p	r	p
轮廓系数	带符号差异与带符号 Δ曲线下面积	−0.932	0.0167	−0.859	0.0250
	绝对差距与配对平均曲线下面积	−0.301	0.3083	+0.321	0.2833
Fisher	带符号差异与带符号 Δ曲线下面积	−0.932	0.0167	−0.860	0.0250
	绝对差距与平均成对 AUC	−0.269	0.3667	+0.294	0.3167

表VIII迁移潜力与有向 AUC（$n = 20$有向配对；数据集标签置换 p）

描述符	机制	r	p	ρ	p_ρ	R^2
轮廓系数	DANN	+0.722	0.0167	+0.693	0.0250	0.521
	直接迁移	+0.568	0.0167	+0.505	0.0583	0.323
Fisher	DANN	+0.720	0.0167	+0.693	0.0250	0.519
	直接迁移	+0.569	0.0167	+0.505	0.0583	0.324

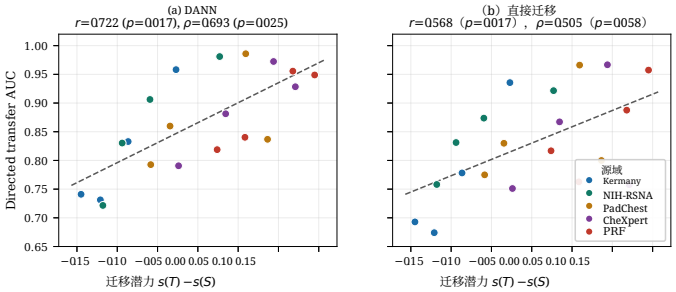

图6. 在 DANN 和 (b) 直接迁移下，所有二十个有序对的迁移潜力与直接迁移曲线下面积。虚线表示线性趋势，点的颜色表示源数据集。迁移潜力为正，意味着目标域比源域更可分。

在适应过程中，目标域、源域和成对层面的关联在最坏情况下仍分别保持至少 0.89、0.91 和 0.78 的幅度。

D. 预测迁移方向

上述关系可以表述为一条决策规则。对于十个无序对中的每一对，预测迁移效果较好的方向是迁移潜力为正的方向，也就是朝向可分性更强的域的方向。表 VIII 报告了该规则的正确率。

从表 VIII 可以看出，在适应条件下，该规则在全部十对中都能识别出效果更好的方向（相对于随机猜测的精确二项检验，$p = 0.001$）。在直接迁移条件下，该规则在十对中有八对判断正确（$p = 0.055$），例外是 Kermany 与 NIH-RSNA，以及 PRF 与 PadChest。采用费舍尔比率后，得到的两项计数完全相同；GBC 和参考估计量也是如此，因为它们得出的排序相同。数据集移除后，该规则仍然成立。依次省去每个数据集后，在五个子集中，该规则在剩余的六对中均全部判断正确。

我们有意不对微小差异设置弃权区间。对于十组配对数据，任何宽到足以排除两个直接迁移例外的阈值，也会舍弃十组中的六组，而选择该阈值的唯一依据只是它

表 VIII方向预测。十个无序配对中，迁移潜力的符号能够识别两种迁移方向中较优方向的比例。

描述符	机制	正确	二项式 p
轮廓系数	DANN	10/10	0.0010
	直接迁移	8/10	0.0547
Fisher	DANN	10/10	0.0010
	直接迁移	8/10	0.0547

表 IX特征提取器的鲁棒性：皮尔逊 r 在两个独立训练的 DenseNet-121 网络下

关系	机制	轮廓系数 NIH	轮廓系数 CheXpert	费舍尔比率 NIH	费舍尔比率 CheXpert
可分性 → 输入 AUC	DANN	+0.934	+0.963	+0.932	+0.823
	直接	+0.850	+0.927	+0.852	+0.726
可分性 → 输出 AUC	DANN	−0.922	−0.761	−0.919	−0.909
	直接	−0.706	−0.572	−0.706	−0.842
带符号差异 → Δ曲线下面积	DANN	−0.932	−0.908	−0.932	−0.847
	直接	−0.859	−0.889	−0.860	−0.809
迁移潜力 →曲线下面积	DANN	+0.722	+0.704	+0.720	+0.657
	直接	+0.568	+0.575	+0.569	+0.539

能够消除观察到的错误。因此，我们报告未经过条件筛选的计数。

E. 特征提取器的稳健性

主提取器是在 NIH ChestX-ray14 上训练的，该数据集是五个受评估数据集之一的源域，因此我们使用在 CheXpert（另一个受评估数据集的源域）上训练的 DenseNet-121 重复了整个分析。两个网络对这五个数据集的排序几乎完全一致（轮廓系数的斯皮尔曼 $\rho = 0.90$ 相关系数为。00，GBC 为 1）。从表 IX 可以看出，在两种描述符下，八种关系的符号均保持不变，总计十六项。两个特征提取器都没有将其自身的训练域评为具有虚高的可分性。使用 NIH 训练权重时，NIH-RSNA 的可分性排名第二；使用 CheXpert 训练权重时则排名第一，而 CheXpert 在两种情况下均排名第四。需要注意的是，每个特征提取器都与一个受评估数据集重合，因此，这项比较表明研究结果并不依赖于单一的特征提取器与数据集重合情况；但并未消除这种重合。

F. 类别平衡

Kermany 和 PRF 是仅有的类别不平衡数据集，且它们分别处于可分性尺度的两端。因此，使用主要的 NIH 加权提取器，在这两个数据集的类别平衡子样本上重新计算了所提出的描述符；这些子样本通过剔除多数类中多余的图像获得（两类分别保留 1349 和 1719 张图像）。其余三个数据集原本就完全平衡，每类均有 3000 张图像，因此平衡处理不会改变它们，数值沿用原值。对十次随机抽样取平均后，Kermany 的轮廓系数从 0.154 变为 0.137 ± 0.002，PRF 的轮廓系数从 0.009 变为 0.011 ± 0.001；费舍尔比率也呈现相同趋势（$0.349 \to 0.320 \pm 0.005$ 和 $0.021 \to 0.021 \pm 0.001$）。这两种变化都比其所处的可分性差距小一个数量级。平衡后的 Kermany 可分性仍是平衡后 PRF 的十二倍以上，

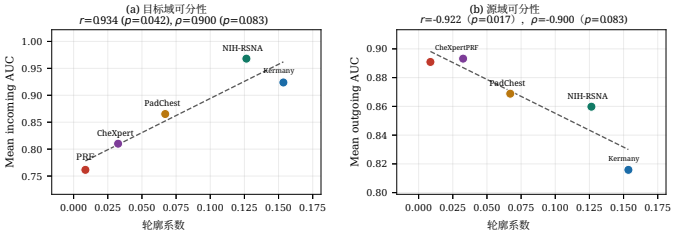

图7. 这两个端点在适应过程中发挥着相反的结构作用。(a) 类别可分性较高的目标域具有较高的平均输入 AUC。(b) 类别可分性较高的源域产生较低的平均传出AUC，因此，重叠程度较高的源域往往具有更强的迁移能力。

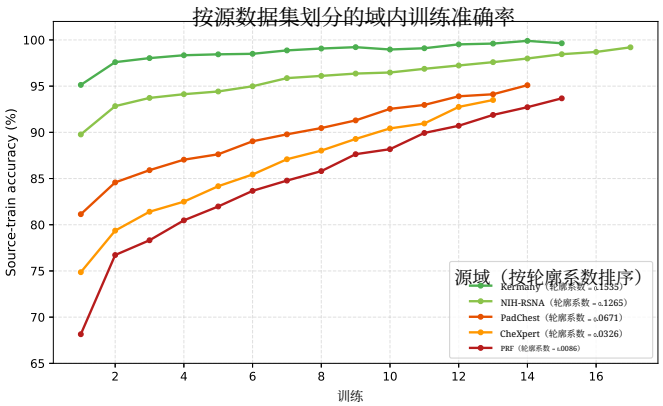

图8. 每个用作源域的数据集的域内训练准确率，按轮廓系数排序。可分性最高的源域起始准确率接近上限，并在几个训练轮次内达到饱和；类别重叠更多的源域起始准确率则低得多，并逐步提升。

且完整的五数据集排序保持不变。按照原始类别比例抽取的大小匹配对照证实，出现的少量变化归因于类别平衡，而非样本量减少；对于 Kermany 轮廓系数，平衡因素的影响是样本量因素的三倍（-0.013 对 -0.004）。

对所得描述符向量重复每项分析后，所有结论均保持不变。在 DANN 下，目标结构可预测输入 AUC，数值为 $r = +0.959$ ($p = 0.025$)；源域结构可预测传出 AUC，数值为 $r = -0.892$ ($p = 0.033$)；迁移潜力可预测有向AUC，数值为 $r = +0.730$($p = 0.017$)；并且所有十个数据集对的性能不对称性方向均能被正确识别（$p = 0.001$）；数据集对层面的绝对间隔仍是唯一无法提供信息的量。在全部十次抽样中，每项相关性均保持原有方向和显著性，而且每次抽样中十个数据集对的方向规则都判断正确。因此，类别不平衡既不会导致所提出描述符度量的结构差异，也不会导致基于这些度量建立的关系。

V. 讨论

在这五个胸部X线影像域之间进行肺炎检测适应，其性能取决于两个端点的类别结构，而这两种依赖关系的符号相反。我们将依次讨论两者，随后探讨二者结合后呈现的方向性关系。

A. 目标域可分性作为参考点

类别分离度更高的目标域更容易适应（图7(a)），这与聚类假设一致：类别本就清晰可分的目标域具有更明确的几何结构，迁移来的决策边界可以据此对齐。我们并不声称这一发现具有新颖性。在相同表征上计算的七种已发表估计量无一例外地重现了这一关系，其中两种的表现与所提描述符相当或更优（表IV）。这一比较表明，目标域一侧已是解决充分的参考点，并由此凸显出尚未解决的两个方面，即源域和方向性。

B. 源域重叠与源域强度

源域关系则较难预料，且尚无已发表的估计器对此进行了描述。类别重叠程度更高的数据集会成为更强的源域（图7(b)），这与以下直觉相悖：一个易于清晰区分的源域问题更简单，因此迁移效果应该最好。

源模型的训练行为与一种基于其所学习表征的解释相一致。图8展示了每个用作源域的数据集的域内训练准确率，并按可分性排序。从图中可以看到，可分性最高的两个源域的起始准确率约为90%至95%，并在几个训练轮次内达到饱和；而类别重叠更多的源域起始准确率则低得多，并逐步提升。类别分离紧密的源域使模型能够快速拟合其任务，这表明模型可以依靠一小组无需具备泛化能力的数据集特定特征来完成任务。类别重叠的源域则无法采用这种解决方式。要将其类别区分开来，就需要更广泛、分布更分散的一组特征，而这些特征似乎更易于迁移。因此，使数据集单独来看更难的类别重叠，也可能正是让它成为更有效源域的属性。

这一现象被报告为一种稳定的观察结果，而非已确立的机制。在两种机制、两种描述符和两种特征提取器下，方向性都成立，训练动态也与所提出的解释一致。不过需要指出的是，这些动态只是与该解释相符，并不能证明它，因为拟合源域的速度也取决于数据集大小和类别平衡。要确认这一点，还需要更多域。

C. 两个端点与迁移的不对称性

源域和目标域在结构上扮演着相反的角色：类别分离良好的数据集是有利的目标域，却是较弱的源域；而类别重叠严重的数据集则恰好相反，Kermany 和 PRF 的反转现象便说明了这一点。迁移不对称性最清楚地表明，两个端点都会产生影响。从 A 迁移到 B 通常不同于从 B 迁移到 A，而首选方向取决于可分性的带符号差异，即从重叠程度较高的域迁移到可分性较高的域。

仅考虑目标域的解释无法表达这一点。因为它只对目标域进行评分，忽略了究竟由哪个域提供

标签。因此，可预测的迁移不对称性直接证明了源域自身的结构包含独立信息。本研究不同于第二节A中的估计量：那些估计量仅将源域视为冻结的特征提取器，因而得出的只是目标域本身的属性。我们通过在同一表征上计算其中七种估计量，对此进行了实证验证。

D. 方向而非幅度

两个域之间的绝对结构差距，在任一机制或任一描述符下，均未显示出与性能存在可检测的关系：结构上相距较远的域，其迁移效果并不比接近的域更差。迁移取决于两个域中哪一个具有更高的可分性，而不是取决于它们差异的程度。因此，相关量是带符号差异，而非距离。迁移潜力恰好捕捉了这一点，并构成第四节D中方向规则的基础。

应当注意所提出指数的局限性。迁移潜力对性能的预测并不比仅使用目标域可分性更准确；目标项占据了大部分权重，而源域项则是一个较小但一致的组成部分。它的价值在于可解释性：将两个端点都发挥作用且符号相反这一事实表达为一个量。

E. 不同机制、表征与度量之间的一致性

本文报告的每一种关系，在直接迁移和适应机制下均呈现相同的符号；在两种独立训练的特征提取器下，以及基于不同原理构建的两种可分性度量下，也同样如此。如果这些关系只在适应机制下出现，那么它们很可能是对抗训练造成的伪象。由于在没有对抗训练的情况下它们同样出现，因此更恰当地说，这些关系是域本身的属性，而适应机制反映了这些属性，而非产生了它们。在直接迁移下，这些关系较弱，尤其是源域侧效应未达到显著性；因此，该机制验证了研究结果的方向，但并未独立确立这些结果。

F. 实践意义

对这五个数据集进行表征只需分别完成一次处理；而通过训练评估相同的二十个有向配对则需要进行二十次适应运行，并且随着候选域数量的增加，这一差距将呈二次扩大。

这一结果的实际价值取决于可获得哪些标签。源侧结果只需要源域标签，因此，拥有多个候选公开数据集的团队可以在接触目标域之前先对其进行排名，将其作为源域，并优先选择重叠程度更高的数据集。目标域结果和方向性结果需要一个带标签的目标域子集，而无监督适应并不假设存在这样的子集；它们适用于已经存在小型验证集的情况——这在临床部署前通常如此——而不适用于完全无标签的情形。该子集究竟可以小到什么程度，本文尚未确定。在这两种情况下，描述符都不能替代实际运行适应：它只能解释迁移性能变化的一部分；与其用于自动进行选择，不如用来排列候选项并设定预期。

G. 局限性

该分析基于五个数据集；如第三节-G所述，在多重性校正后，这一规模下的任何结果都无法达到显著性。因此，这些发现属于假设生成性发现，尤其是源侧关系还需要在更多域中进行验证。研究范围也较为狭窄：任务是基于胸部X线片检测肺炎，本文没有任何证据表明相同的关系适用于其他病理、多分类问题或其他模态。研究只考察了一种适应算法，因此，关于这些关系并非特定于某种方法的结论，是基于DANN与无适应基线共同得出的。

有三点测量方面的注意事项与结果解读相关。报告中的适应后的AUC是在训练过程中目标评估部分所达到的最佳值，衡量的是适应所能达到的效果，而非部署模型可能获得的效果。两个特征提取器均在五个评估数据集中的一个上进行训练，因此二者都不能完全独立于该数据集集合；不过，它们的一致性表明，结论并不依赖于单一的数据重叠。参考估计量仅用于目标侧检查，而 LEEP 和 PACTran 的狄利克雷变体完全无法纳入（第三节-D）。将表征范围扩展到单一学习特征空间中的类别可分性之外，仍是一个有待研究的方向。

VI. 结论

通过冻结的 DenseNet-121 进行一次处理获得特征，并依据这些特征的类别可分性对五个胸部X线影像域进行了表征；随后，在无监督域适应条件下，针对全部二十个有向源域-目标域对，将该表征与放射影像肺炎检测性能联系起来。性能取决于方向相反的两个端点：可分性更高的目标域获得了更高的输入 AUC（$r = 0.934$，$p = 0.042$），而可分性更高的源域产生了更低的输出 AUC（$r = -0.922$，$p = 0.017$）。两个端点之间的带符号差异在 10 对中的 10 对中识别出了更优的迁移方向（$p = 0.001$），而其幅度未显示出任何信号。在直接迁移、使用第二个特征提取器以及使用第二种可分性度量的情况下，每种关系的符号均保持不变。

在相同表征上计算的七种已发表可迁移性估计器均再现了目标侧的关系；本文将其视为参照点，而非本文的贡献。这些估计器并未描述源域，而两个端点之间的对立正是能够预测迁移方向的原因。由于仅涉及五个域，这一发现仍属探索性结果；若能在更大规模的研究中、跨越更多适应方法并拓展至放射影像肺炎检测之外予以验证，就能确立域表征作为一项实用步骤的价值，帮助在适应运行出错代价最高的场景中部署这些模型的团队。

数据可用性

Kermany、CheXpert、NIH ChestX-ray14、RSNA和PadChest数据集可从其原始分发方公开获取。PRF数据集为私人收集的数据，不公开提供。代码、预先计算的

结构描述符和完整的统计流程可在 https://github.com/mHealthBuet/CXR-Domain-Structure-UDA获取。

利益冲突

作者声明不存在利益冲突。

参考文献

[1] 世界卫生组织，《儿童肺炎》，世界卫生组织情况说明书，2022年11月。 [在线]。可访问：https://www.who.int/news-room/fact-sheets/detail/pneumonia [2] E. D. McCollum等人，“孟加拉国肺炎球菌疫苗有效性研究中的胸部X线片阅片小组表现”，《BMJ开放呼吸研究》，第6卷，第1期，2019年4月，文章编号e000393，doi：10.1136/bmjresp-2018-000393。[3] A. H. Baqui等人，“孟加拉国肺炎球菌结合疫苗影响评估”，《盖茨开放研究》，第2卷，第21页，2018年4月，doi：10.12688/gatesopenres.12805.1。[4] P. Rajpurkar等人，“CheX Net：利用深度学习实现放射科医生水平的胸部X光肺炎检测”，arXiv:1711.05225，2017年。[5] J. R. Zech、M. A. Badgeley、M. Liu、A. B. Costa、J. J. Titano和E. K. Oermann，“用于检测胸部X线片中肺炎的深度学习模型的泛化性能存在差异：一项横断面研究”，《PLoS医学》，第15卷，第11期，2018年11月，文章编号e1002683。[6] H. Guan和M. Liu，“医学图像分析的域适应：综述”，《IEEE生物医学工程汇刊》，第69卷，第3期，第1173–1185页，2022年3月，doi：10.1109/TBME.2021.3117407。[7] E. Ahn、A. Kumar、M. Fulham、D. Feng和J. Kim，“利用零偏置卷积自编码器和基于上下文的特征增强进行医学图像分类的无监督域适应”，《IEEE医学成像汇刊》，第39卷，第7期，第2385–2394页，2020年7月，doi：10.1109/TMI.2020.2971258。[8] P. Thiam等人，“跨域胸部X光图像中心脏肥大检测的无监督域适应”，《前沿人工智能》，第6卷，2023年，文章编号1056422，doi：10.3389/frai.2023.1056422。[9] G. Wilson和D. J. Cook，“无监督深度域适应综述”，《ACM智能系统与技术汇刊》，第11卷，第5期，2020年10月，文章编号51，doi：10.1145/3400066。[10] B. Kulis、K. Saenko和T. Darrell，“所见并非所得：利用非对称核变换进行域适应”，载于《IEEE计算机视觉与模式识别会议论文集》（CVPR），美国科罗拉多州科罗拉多斯普林斯，2011年，第1785–1792页。[11] Y. Bao等人，“一种面向任务迁移学习中可迁移性的信息论方法”，载于《IEEE国际图像处理会议论文集》（ICIP），2019年，第2309–2313页。[12] C. V. Nguyen、T. Hassner、M. Seeger和C. Archambeau，“LEEP：一种评估学习到的表征可迁移性的新度量”，载于《国际机器学习会议论文集》（ICML），《机器学习研究论文集》，第119卷，2020年，第7294–7305页。[13] K. You、Y. Liu、J. Wang和M. Long，“LogME：用于迁移学习的预训练模型实用评估”，载于《国际机器学习会议论文集》（ICML），《机器学习研究论文集》，第139卷，2021年，第12133–12143页。[14] S. Ibrahim、N. Ponomareva和R. Mazumder，“更新并不总是更好：重新思考可迁移性指标及其特性、稳定性和性能”，载于《欧洲机器学习与知识发现数据库会议论文集》（ECML PKDD），2022年，第693–709页。[15] Y. Li等人，“神经网络检查点排序”，载于《IEEE/CVF计算机视觉与模式识别会议论文集》（CVPR），2021年，第2663–2673页。[16] L.-K. Huang、J. Huang、Y. Rong、Q. Yang和Y. Wei，“出人意料地简单的可迁移性估计”，载于《国际机器学习会议论文集》（ICML），《机器学习研究论文集》，第162卷，2022年，第9201–9225页。[17] N. Ding、X. Chen、T. Levinboim、S. Changpinyo和R. Soricut，“PACTran：用于估计预训练模型向分类任务可迁移性的PAC-贝叶斯指标”，载于《欧洲计算机视觉会议论文集》（ECCV），2022年，doi：10.1007/978-3-031-19830-4 15。[18] W. Shao等人，“并非所有模型都相同：在自我挑战的费舍尔空间中预测模型可迁移性”，载于《欧洲计算机视觉会议论文集》（ECCV），2022年，doi：10.1007/978-3-031-19830-4。[19] M. P´andy、A. Agostinelli、J. Uijlings、V. Ferrari和T. Mensink，“利用Bhattacharyya类可分性进行可迁移性估计”，载于《IEEE/CVF计算机视觉与模式识别会议论文集》（CVPR），2022年，第9172–9182页。[20] R. Shu、H. H. Bui、H. Narui和S. Ermon，“一种面向无监督域适应的DIRT-T方法”，载于《国际学习表征会议论文集》（ICLR），2018年。[21] S. Ben-David、J. Blitzer、K. Crammer、A. Kulesza、F. Pereira和J. W. Vaughan，“跨域学习理论”，《机器学习》，第79卷，第1–2期，第151–175页，2010年5月。[22] Y. Feng等人，“基于深度有监督域适应的胸部X光图像肺炎诊断”，《IEEE生物医学与健康信息学杂志》，第26卷，第3期，第1080–1090页，2022年3月，doi：10.1109/JBHI.2021.3100119。[23] A. Musa、R. Prasad和M. Hernandez，“通过有监督对抗域适应解决胸部X光分类中的跨人群域偏移”，《科学报告》，第15卷，第1期，2025年4月，文章编号11383，doi：10.1038/s41598-025-95390-3。[24] Y. Ganin等人，“神经网络的域对抗训练”，《机器学习研究杂志》，第17卷，第59期，第1–35页，2016年。[25] K. Sanchez、C. Hinojosa、H. Arguello、D. Kouam´e、O. Meyrignac和A. Basarab，“CX-DaGAN：面向小型胸部X光数据集肺炎诊断的域适应”，《IEEE医学成像汇刊》，第41卷，第11期，第3278–3288页，2022年11月，doi：10.1109/TMI.2022.3182168。[26] Y. Zhou、X. Li和Y. Wu，“用于肺炎诊断的循环一致对抗胸部X光域适应”，《神经计算》，第610卷，2024年12月，文章编号128604。[27] Y. Tang、Y. Tang、V. Sandfort、J. Xiao和R. M. Summers，“TUNA-Net：用于跨域胸部X光疾病识别的面向任务的无监督对抗网络”，载于《医学图像计算与计算机辅助干预国际会议论文集》（MICCAI），《计算机科学讲义》，第11769卷，2019年，第431–440页。[28] D. S. Kermany等人，“通过基于图像的深度学习识别医学诊断和可治疗疾病”，《细胞》，第172卷，第5期，第1122–1131.e9页，2018年2月。[29] X. Wang等人，“ChestX-ray8：医院规模的胸部X光数据库，以及常见胸部疾病弱监督分类与定位基准”，载于《IEEE计算机视觉与模式识别会议论文集》（CVPR），美国夏威夷州檀香山，2017年，第2097–2106页。[30] G. Shih等人，“利用专家对可能肺炎的标注扩充美国国立卫生研究院胸部X线摄影数据集”，《放射学人工智能》，第1卷，第1期，2019年1月，文章编号e180041，doi：10.1148/ryai.2019180041。[31] A. Bustos、A. Pertusa、J.-M. Salinas和M. de la Iglesia-Vay´a，“PadChest：包含多标签标注报告的大型胸部X光图像数据集”，《医学图像分析》，第66卷，2020年12月，文章编号101797，doi：10.1016/j.media.2020.101797。[32] J. Irvin等人，“CheXpert：包含不确定性标签和专家比较的大型胸部X线摄影数据集”，载于《AAAI人工智能会议论文集》，美国夏威夷州檀香山，2019年，第590–597页。[33] J. P. Cohen等人，“TorchXRayVision：胸部X光数据集和模型库”，载于《医学成像深度学习会议论文集》（MIDL），《机器学习研究论文集》，第172卷，2022年，第231–249页。[34] P. J. Rousseeuw，“轮廓系数：用于聚类分析解释与验证的图形辅助工具”，《计算与应用数学杂志》，第20卷，第53–65页，1987年11月。[35] T. K. Ho和M. Basu，“监督分类问题的复杂度度量”，《IEEE模式分析与机器智能汇刊》，第24卷，第3期，第289–300页，2002年3月。[36] L. McInnes、J. Healy和J. Melville，“UMAP：用于降维的均匀流形近似与投影”，arXiv:1802 03426，2018年。