arXiv:2303.00500v2 [cs.CV] 8 Aug 2023

利用类别特定反事实实现内在可解释的多标签分类

Susu Sun[1] SUSU.SUN@UNI-TUEBINGEN.DE[1] 卓越集群——科学机器学习，*T*¨宾根大学，德国**Stefano Woerner**[1] STEFANO.WOERNER@UNI-TUEBINGEN.DE
Andreas Maier[2] ANDREAS.MAIER@FAU.DE
[2] 模式识别实验室，埃尔朗根-纽伦堡弗里德里希-亚历山大大学¨¨，德国**Lisa M. Koch**[3], [4] LISA.KOCH@UNI-TUEBINGEN.DE[3] 赫蒂脑健康人工智能研究所，*T*¨宾根大学，德国[4] 眼科研究所，*T*¨宾根大学，德国**Christian F. Baumgartner**[1] CHRISTIAN.BAUMGARTNER@UNI-TUEBINGEN.DE

摘要

在 Medical Image Analysis 等高风险应用领域，可解释性对于机器学习算法至关重要。然而，高性能的黑箱神经网络无法为其预测提供解释，这可能导致不信任，并阻碍高效的人机协作。事后解释技术在实践中被广泛使用，但研究表明，这些技术存在严重的概念性问题。此外，正如本文所示，现有解释技术在多标签场景下表现不佳；在该场景中，单幅图像可能同时出现多种医学发现。我们提出了 Attri-Net[1]，，一种用于多标签分类的固有可解释模型。Attri-Net 是一种强大的分类器，能够提供透明、可信且易于人类理解的解释。该模型首先基于反事实生成类别特定归因图，以识别哪些图像区域对应于特定医学发现。随后，仅使用这些归因图，通过一个简单的逻辑回归分类器进行预测。我们在三个胸部 X 光数据集上，将 Attri-Net 与五种事后解释技术及一种固有可解释分类器进行了比较。结果表明，Attri-Net 能够生成与临床知识一致的高质量多标签解释，并且其分类性能可与最先进的分类模型相媲美。

1. 引言

机器学习（ML）技术在临床上的应用受到深度学习模型黑箱特性的阻碍。模型难以理解，可能导致信任缺失（Dietvorst等人，2015年），也可能使临床医生盲目信任模型（Tschandl等人，2020年；Gaube等人，2021年），并可能引发伦理和法律问题（Grote和Berens，2020年）。因此，透明性被视为在医学等高风险应用领域部署机器学习技术的关键属性之一（Rudin，2019年）。

理解机器学习模型决策机制时最常用的一类技术，是事后 方法。这类方法将启发式方法应用于训练好的模型，试图在预测作出后回溯性地理解其决策机制。

1. Attri-Net 的代码可在以下网址获取：https://github.com/ss-sun/Attri-Net

© 2023 S. Sun、S. Woerner、A. Maier、L.M. Koch 和 C.F. Baumgartner。

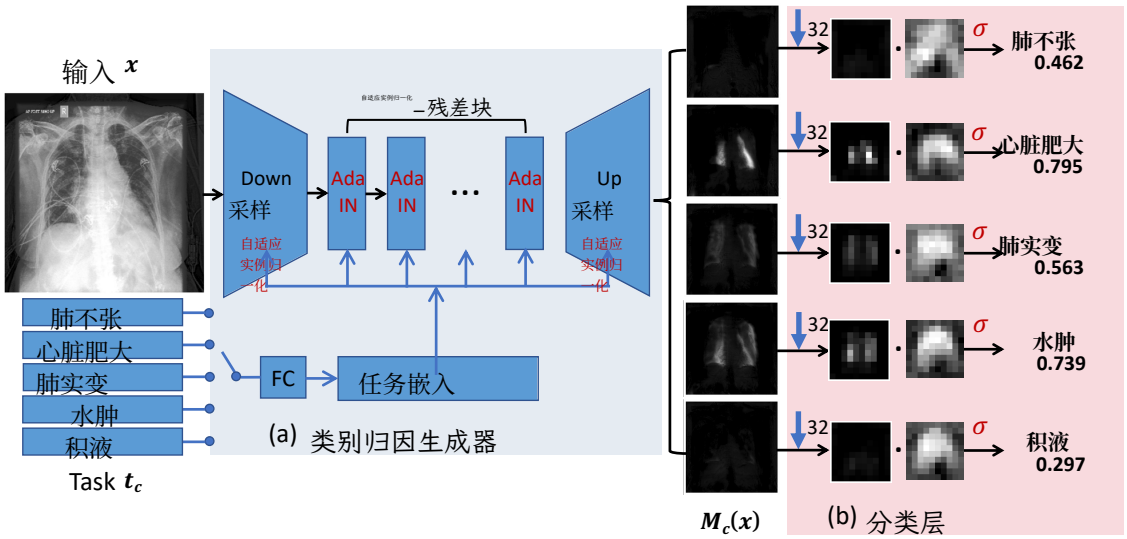


图 1：Attri-Net 框架。给定一张输入图像和一个任务，我们的视觉特征归因生成器 (a) 会生成反事实归因图。逻辑回归分类器 (b) 根据这些归因图，为每个类别生成最终预测。

基于梯度的技术（如引导式反向传播（Springenberg 等人，2014年））通过计算预测值相对于输入像素的梯度，对黑盒模型进行局部函数逼近。Adebayo 等人（2018年）和 Arun 等人（2021年）最近对这类方法忠实于决策机制这一点提出了质疑，他们发现，即使随机化网络权重，解释结果仍然保持不变。基于扰动的方法（如 LIME（Ribeiro 等人，2016）和 SHAP（Lundberg 和 Lee，2017））也会对局部决策函数进行逼近。目前，这些方法无法生成像素级解释，而且计算开销很大。另一类研究包括类激活映射（CAM）（Zhou 等人，2016）和 GradCAM（Selvaraju 等人，2017），旨在构建能够直接推断决策机制的神经网络架构。然而，这些技术受限于其解释的空间分辨率，无法在像素级解释推理机制。BagNet（Brendel 和 Bethge，2019）通过大幅限制网络的全局感受野来解决这一问题，但这可能会对分类性能产生负面影响。在网络不同深度处放置注意力模块，也能为各个特征图提供一定程度的可解释性（Schlemper 等人，2019；Yan 等人，2019）。与我们提出的方法密切相关的一类方法是反事实解释，它们要么试图回答“如果图像属于另一个类别，它会是什么样子？”这一问题（Schutte 等人，2021；Joshi 等人，2018），要么夸大预测类别的特征（Cohen 等人，2021；Singla 等人，2019）。这一类别中的其他方法则从反事实生成器的中间表示中推导分类结果（Bass 等人，2020；Cetin 等人，2022）。我们还注意到，有些技术尝试生成反事实，但并非为了对分类器作出解释（Baumgartner 等人，2018；Nemirovsky 等人，2020）。

尽管事后解释可能看起来合理，但无法保证它们解释了分类器的实际行为；事实上，越来越多的证据表明，它们并不忠实于实际的决策机制（Adebayo 等人，2018年；Han 等人，2022年；White 和 Garcez，2019年）。相比之下，内在可解释方法所采用的预测系统，其决策机制能够直接呈现给用户。根据定义，这些模型忠实于决策机制，因为解释本身就是决策机制。已有研究包括以下方法：最终预测直接基于人类可理解的概念（Alvarez Melis 和 Jaakkola，2018年；Chen 等人，2020年；Koh 等人，2020年）、类别的原型表示（Chen 等人，2019年；Barnett 等人，2021年），或直接将归因对应到图像块（Javed 等人，2022年）。据我们所知，近期提出的卷积动态对齐网络（CoDA-Nets）（Bohle 等人，2021年）是现有唯一能够在像素级别提供内在可解释视觉解释的模型。该方法将网络权重表示为输入图像的函数，从而能够将网络针对特定输入图像所做的决策表述为线性分类器。我们注意到，据我们所知，目前还没有基于反事实解释的内在可解释方法。

大多数视觉解释技术都是为二分类或多分类问题开发的。然而，许多临床任务属于多标签问题，即多个类别可以同时适用。例如，在本研究关注的胸部 X 光诊断中，一张图像往往包含多个发现。在实验中，我们发现现有的视觉解释技术并不适用于这一重要类型的决策问题。具体而言，这些解释并非针对特定类别，而且往往会为所有类别突出显示相似区域；有时即使图像中并不存在该类别，也会如此。

本文提出了 Attri-Net，这是一种专为多标签场景设计、具有内在可解释性的视觉解释技术。我们的模型将类别特定的反事实归因图预测为中间表示。这些归因图在概念上基于视觉特征归因生成对抗网络（VA-GAN）方法，该方法由 Baumgartner 等人 (2018年) 提出，代表残差图像，其中包含输入图像中某一类别的所有现有证据。随后，Attri-Net 将这些类别特定归因图作为输入特征，输入最终的线性分类层。我们在三个广泛使用的胸部 X 光数据集上评估了 Attri-Net，并证明该方法能够生成高质量、具有内在可解释性的解释，且类别敏感度高，同时分类性能可与最先进的模型相媲美。

2. 方法

本文针对具有 C 个类别的多标签分类场景，其中每个类别 c 具有标签 $y_c \in \{0, 1\}$，且可独立地出现在图像中；也就是说，多个医学发现可以同时存在。接下来，我们首先介绍如何为每个类别 c 生成反事实类别归因图（见图1a）。然后，我们展示如何使用逻辑回归分类器，根据这些归因图得到最终预测（见图1b）。最后，我们说明如何在所提出的 Attri-Net 框架中对这两个组件进行端到端训练。

2.1. 反事实类别归因图生成

我们方法的核心是一个图像到图像网络 $M_c(\mathbf{x})$: $\mathbb{R}^{h\times w} \mapsto \mathbb{R}^{h\times w}$，它为输入图像 $\mathbf{x}$生成残差反事实类别归因图。直观而言， M_c 的输出表示输入图像中的每个像素应如何变化，才能消除类别 c 对图像的影响。与 Baumgartner 等人 (2018年一样，我们学习一种加性映射 M_c，使输出图像看起来像是来自相反的类别，即

$$\hat{\mathbf{x}} = \mathbf{x} + M_c(\mathbf{x}) \,,$$

使得生成的反事实图像 $\hat{}\mathbf{x}$ 与从真实图像分布中采样的图像无法区分 $p(\mathbf{x}|y_c = 0)$，这些真实图像 不 包含类别 c。为确保 M_c行为正确，我们同时训练一个类别特定判别器网络 D_c，使其能够区分真实图像和伪图像，且 $y_c = 0$。具体而言，我们采用 Wasserstein 生成对抗网络损失（Arjovsky 等人，2017年；Baumgartner 等人，2018年）。附录D_c B.1 中给出了 B.1对 D_c 进行优化的细节。给定判别器函数 D_c，我们可以写出以下对抗损失项，以确保 $\hat{}\mathbf{x}$ 是一个不包含类别 c 的真实感反事实图像，并由此确保 M_c 输出真实感残差类别归因图：

$$\mathcal{L}_{\text{adv}}^{(c)} = \mathop{\mathbb{E}}_{\mathbf{x}\sim p(\mathbf{x}|y_c=1)} [-D_c(\mathbf{x} + M_c(\mathbf{x}))] \,. \tag{1}$$

附录中展示了一些生成的反事实图像 $\hat{}\mathbf{x}$ 示例，见A.1。

为避免网络将不属于给定类别的多余像素归因于该类别，我们还通过一个 L_1 正则化项鼓励类别归因图保持稀疏，该项与 Baumgartner 等人 (2018年类似）。为进一步促使生成器在图像中存在该类别时产生比不存在时更小的影响，我们将损失拆分为两个权重不同的项：类别负例使用较大的权重 α_0，而类别正例使用较小的权重 α_1，即：

$$\mathcal{L}_{\text{reg}}^{(c)} = \alpha_0 \mathop{\mathbb{E}}_{\mathbf{x}\sim p(\mathbf{x}|y_c=0)} [\|M_c(\mathbf{x})\|_1] + \alpha_1 \mathop{\mathbb{E}}_{\mathbf{x}\sim p(\mathbf{x}|y_c=1)} [\|M_c(\mathbf{x})\|_1] \,. \tag{2}$$

本文所有实验均使用 $\alpha_0 = 2$， $\alpha_1 = 1$ 。

函数 M_c和 D_c 均以 StarGAN 架构（Choi 等人，2018年）为基础，通过神经网络实现。与我们探索的其他方案（例如原始 VA-GAN 架构）相比，该架构取得了更好的结果。尽管可以设计网络 M，在一次前向传播中通过多个输出通道生成所有标签的类别归因图，但初步实验表明，这种方法在多标签场景中的类别归因效果不足。因此，我们基于近期提出的任务切换网络（Sun 等人，2021年），其中使用自适应实例归一化（AdaIN）层在相关任务之间切换。在我们的工作中，任务对应于生成不同类别的归因图。每个任务由任务向量$\mathbf{t}_c$ 表示，该向量是按（Sun 等人，2021年）中的方法，将独热编码上采样 20 倍得到的。随后，通过一个小型全连接网络将该编码转换为任务嵌入，并输入布置在网络各处的 AdaIN 层（如图1a 所示）。AdaIN 层由此切换网络的行为。为了将这一范式与 StarGAN 架构结合，我们将原始生成器和判别器网络中的所有实例归一化层替换为 AdaIN 层

。该架构的详细说明见附录B.2。现在，掩码生成器和判别器可分别表示为 $M_c(\mathbf{x}) = M(\mathbf{x}, \mathbf{t}_c)$，以及$D_c(\mathbf{x}) = D(\mathbf{x}, \mathbf{t}_c)$。遍历所有类别的$\mathbf{t}_c$ 向量，并对 M 反复进行前向传播，即可得到所有标签的类别归因图。

2.2. 使用逻辑回归分类器进行分类

给定使用 $M(\mathbf{x}、\mathbf{t}_c)$ 得到的类别特定反事实归因图，我们希望预测图像中是否存在类别 c。为此，我们对相应的归因图进行下采样，并将其作为逻辑回归分类器的输入。也就是说，

$$p(y_c|\mathbf{x}) = \sigma\Big(\sum_{i,j} w_{ij}^{(c)} \cdot S_\gamma(M(\mathbf{x}, \mathbf{t}_c))_{ij}\Big), \tag{3}$$

其中，S_γ 是一个二维平均池化算子，以 γ 倍的比例进行下采样；$w_{ij}^{(c)}$ 表示与类别 c 的下采样归因图中各像素相关联的权重；σ是 Sigmoid 函数。在初步实验中，我们发现 $\gamma = 32$ 表现稳健，因此在所有实验中均采用该值。

分类器使用标准的多标签分类损失 $\mathcal{L}^{(c)}_{\text{cls}}$进行训练，该损失是每个类别的二元交叉熵损失之和。需要注意的是，由于我们的框架采用端到端训练，M 也会从该损失中接收梯度，从而促使其生成可通过线性分类的类别归因图。

为进一步促使归因图能够区分每个类别的正样本和负样本，我们采用了Wen 等人（2016提出的中心损失，该损失已被证明能生成区分性更强的特征表示。基于这一思路，我们在此定义类别中心$\mathbf{v}_{y_c=0}$，$\mathbf{v}_{y_c=1} \in \mathbb{R}^{h\times w}$ 这些中心可学习，并会收敛到与每个类别正负实例对应的归因图的原型表示 c。中心损失会使类别归因图更接近各自的类别中心，从而形成聚类程度更高的特征空间，使正负样本更易于线性分离。整体中心损失可写为

$$\mathcal{L}^{(c)}_{\text{ctr}} = \frac{1}{2}\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim p(\mathbf{x}|y_c=0)}\left[\|M(\mathbf{x}, \mathbf{t}_c) - \mathbf{v}_{y_c=0}\|_2^2\right] + \mathop{\mathbb{E}}_{\mathbf{x}\sim p(\mathbf{x}|y_c=1)}\left[\|M(\mathbf{x}, \mathbf{t}_c) - \mathbf{v}_{y_c=1}\|_2^2\right]\right). \tag{4}$$

如Wen 等人（2016）所述，每个小批量都会单独执行一次梯度更新，以更新类别中心图像；该更新与网络参数的更新交错进行。最终的类别中心图像以及逻辑回归权重 $w_{ij}^{(c)}$ 可用于从全局层面进一步解释模型的行为。附录A.2展示了二者的示例。不过，全局可解释性的探索留待未来工作。

2.3. 训练

我们的 Attri-Net 框架可以通过四个损失项进行端到端训练，以满足我们的核心要求：首先，归因图应保留足够的类别相关信息，从而获得令人满意的分类结果。其次，

表1：以 ROC 曲线下面积（曲线下面积）衡量的分类性能。

模型	CheXpert	ChestX-ray8	VindrCXR
ResNet50(Azizi 等人，2021年)	0.7687	-	-
SimCLR(Azizi 等人，2021年)	0.7702	-	-
LSE (Ye 等人，2020年)	-	0.7554	-
ChestNet（Ye 等人，2020年）	-	0.7896	-
ResNet50	0.7727	0.7445	0.8986
CoDA-Nets	0.7659	0.7727	0.9322
ours	0.7405	0.7762	0.9405

表2：类别敏感度分数比较。

模型	CheXpert	ChestX-ray8	VindrCXR
ResNet + GB	0.3183	0.3028	0.1727
ResNet + GCam	0.1434	0.1570	0.1931
ResNet + LIME	0.2347	0.2609	0.2422
ResNet + SHAP	0.4745	0.4122	0.3714
ResNet + Gifsplan。	0.2748	0.5817	0.4396
CoDA-Nets	0.3576	0.4138	0.4464
ours	**0.4880**	**0.6160**	**0.5509**

归因图应具有良好的可解释性。类别归因生成器 M 在权重参数 φ 下的整体训练目标为

$$\min_{\varphi} \sum_{c} \lambda_{\text{cls}} \mathcal{L}_{\text{cls}}^{(c)} + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}}^{(c)} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}^{(c)} + \lambda_{\text{ctr}} \mathcal{L}_{\text{ctr}}^{(c)}, \tag{5}$$

其中我们使用超参数 λ_* 来平衡各项损失。我们的实验选用了 $\lambda_{\text{cls}} = 100$，$\lambda_{\text{adv}} = 1$，$\lambda_{\text{reg}} = 100$，$\lambda_{\text{ctr}} = 0.01$。关于不同损失影响的消融研究见附录A.3。训练过程中，我们反复遍历不同类别 c，并针对每个类别采样两个小批次，一个包含当前类别的正样本，另一个包含负样本。利用这两个批次，我们先对所有损失执行梯度更新，然后再处理下一个类别。我们使用 ADAM 优化器（Kingma 和 Ba，2014），以学习率10^{-4} 和批次大小 4 来优化模型。此外，遵循Wen 等人 (2016)，我们使用随机梯度下降来更新中心损失参数。在 Nvidia V100 GPU 上，训练在 72 小时内收敛。训练结束后，我们针对验证集中的每个类别选择使 Youden指数（敏感度 + 特异度 - 1）最大的决策阈值。我们也对基线方法执行这一步骤。

3. 实验与结果

数据。我们在三个广泛使用的胸部 X 光数据集 CheXpert（Irvin 等人，2019年）、ChestX-ray8（Wang 等人，2017年）和 VinDrCXR（Nguyen 等人，2020年）上评估了我们提出的 Attri-Net。遵循（Irvin 等人，2019年）和（Azizi 等人，2021年）的方法，在 CheXpert 和 ChestX-ray8 数据集中，我们使用了“Atelectasis”“Cardiomegaly”“Consolidation”这几个类别，

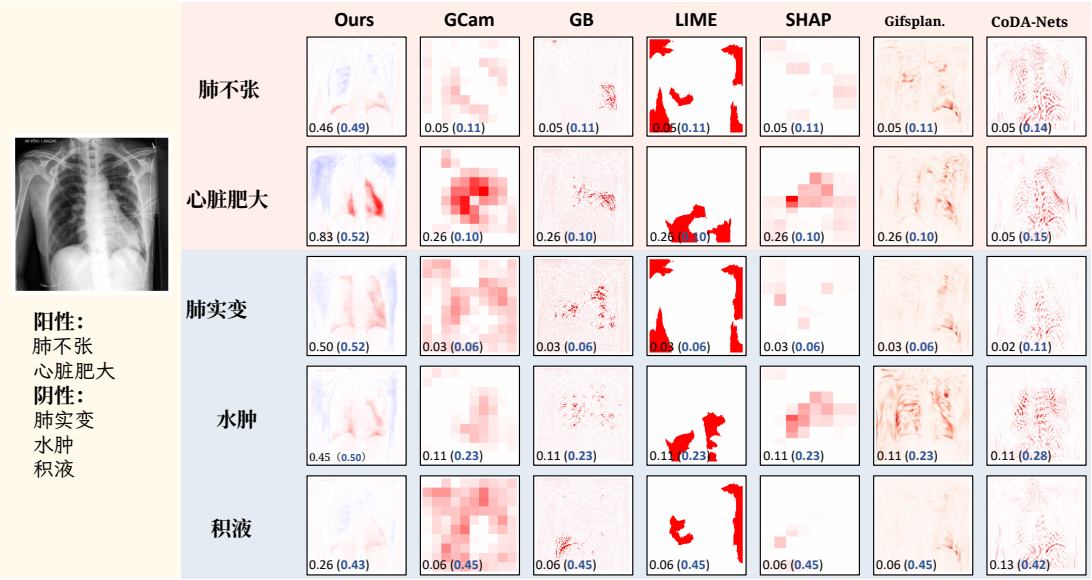


图2：CheXpert 数据集中一张示例图像的解释可视化对比。每张归因图的左下角标示了预测类别的概率，括号中为相应的决策阈值。

"Edema"和"Pleural Effusion"。对于 VinDr-CXR 数据集，我们选取了样本数量最多的五种病变，分别为"Aortic enlargement""Cardiomegaly""Pulmonary fibrosis""Pleural thickening"和"Pleural effusion"。我们将所有数据集划分为训练集（80%）、测试集（10%）和验证集（10%）。由于 CheXpert 的测试集未公开，且官方验证集规模较小，我们采用了（Azizi 等人，2021年）中使用的方法，将训练集划分为训练集、验证集和测试集。**分类性能**。为了评估分类性能，我们将模型与最先进的内在可解释模型 CoDA-Nets（Bohle 等人，2021年）以及标准的黑盒 ResNet50 模型进行了比较。我们还报告了 Azizi 等人（2021年）和Ye 等人（2020年）分别在 CheXpert 和 ChestX-ray8 上取得的结果。总体而言，Attri-Net 的表现与最先进方法相当（见表 1），其 ROC 曲线下面积在 CheXpert 上略低，在 ChestX-ray8 上与其他方法相近，在 VindrCXR 上则略好。

可解释性。Khakzar 等人（2021年）认为，如果图像的不同区域分别负责预测不同类别，那么相应的解释也应该有所不同。他们将这一属性称为"类别敏感度"。在多标签分类中，包含某个类别的图像，其解释中的归因应高于不含该类别的图像。我们参照Bohle 等人（2021年）的方法测量类别敏感度，并创建了一系列 2×2 解释网格，每个网格仅包含某一类别的一个正例（示例网格见附录A.4 ）。随后，我们用正例中的归因总和除以网格中所有归因的总和，以此表示类别敏感度。在理想情况下，只有疾病阳性图包含归因，而疾病阴性归因图为空，此时敏感度为 1。我们计算了每个类别在 200 个网格上的平均敏感度 c。

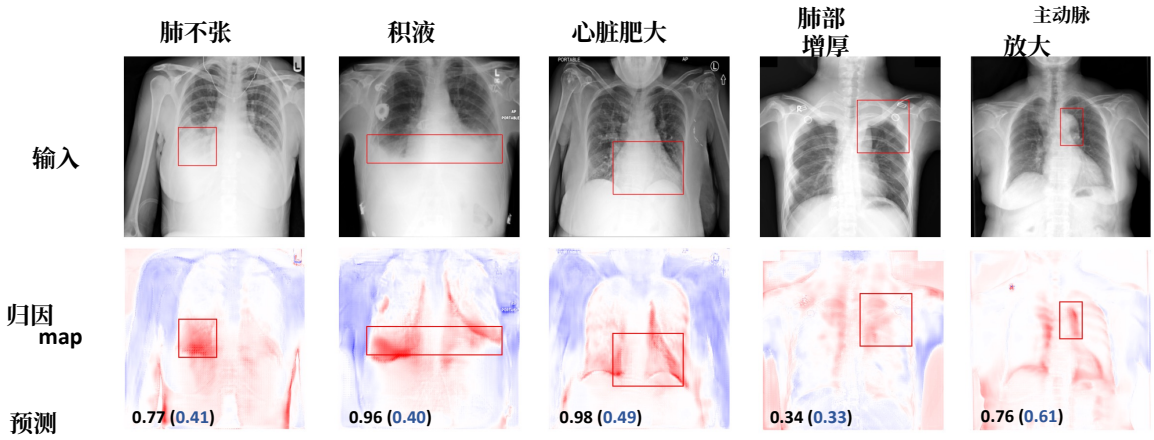


图3：ChestX-ray8 和 VindrCXR 中带有 GT 边界框的样本的归因图。括号中给出了决策阈值。

与可解释性内在具备的基线模型 CoDA-Nets 相比，我们的方法在所有数据集上的类别敏感度都显著且稳定地更高（见表2）。对于黑盒 ResNet，我们比较了五种事后解释技术，即引导式反向传播（Springenberg 等人，2014）、GradCAM（Selvaraju 等人，2017年）、LIME（Ribeiro 等人，2016）、SHAP（Lundberg 和 Lee，2017年）以及近期提出的 Gifsplanation（Cohen 等人，2021年）。不同事后方法的表现差异很大；SHAP 和 Gifsplanation 的表现与 CoDA-Nets 相当，但明显逊于我们的 Attri-Net。

对示例解释进行定性检查进一步支持了这些结果。我们提出的 Attri-Net 生成了类别归因图，清晰地突出显示支持相应类别的底层解剖结构部分（参见图2 ，其中展示了 CheXpert 数据集中的一个代表性示例）。此外，不同类别的归因彼此明显不同，每种归因都聚焦于不同的解剖区域。ChestX-ray8 和 VinDr-CXR 数据集中的示例见附录 A.5。相比之下，具有内在可解释性的基线方法 CoDA-Nets 为所有类别生成了视觉上相似的归因（图2最右列）。我们还观察到，基线技术大多无法有效识别解剖结构中哪些部分促成了某项预测。在各基线方法中，尽管引导式反向传播在定性上提供了最有用的解释，但其归因非常嘈杂，这也是该技术的典型特征。我们进一步在每个类别的示例图像上检验了 Attri-Net 的解释，这些图像均有可用的病理边界框（图3）。总体而言，Attri-Net 突出了与相应病理相关的区域；当最终预测具有较高置信度时，其归因图尤其敏感（即肺不张、积液和心脏肥大的示例）。我们还观察到，边界框之外的区域中也存在一些相对较强的归因。由于我们的类别归因图基于旨在逼真地移除病理全部影响的反事实，我们推测，这些归因图可能揭示了与类别相关、但不属于临床分级方案的其他相关影响。

4. 讨论与结论

我们提出了 Attri-Net，这是一种新颖的具有内在可解释性的多标签分类器，并证明其能够生成高质量的解释：在类别敏感度方面显著优于所有基线方法，同时保持与当前最先进黑盒模型相当的分类性能。黑盒模型的解释高度依赖事后技术，即使针对同一图像，不同技术生成的解释也存在根本差异。这削弱了人们对其在高风险应用中提供必要透明性的能力的信任，也表明有必要采用像我们的方法这样具有内在可解释性的模型；在这类模型中，预测结果直接且线性地由视觉上可解释的类别归因图形成。

本文的定性和定量评估表明，我们的方法能够提供有用的解释，但未来仍有许多重要的研究方向。我们认为，迈向临床影响的关键一步，是在实际的人机协作场景中评估可解释模型，并通过临床相关终点评估其效用。

致谢

本研究由德国研究联合会（DFG，德国研究基金会）通过“德国卓越战略”资助，项目编号为 EXC 2064/1，项目编号 390727645。作者感谢卡尔蔡司基金会对“医疗保健安全机器学习系统的认证与基础”项目的支持，以及赫蒂基金会的支持。作者感谢国际马克斯·普朗克智能系统研究生院（IMPRS-IS）为 Susu Sun 和 Stefano Woerner 提供支持。

参考文献

Julius Adebayo、Justin Gilmer、Michael Muelly、Ian Goodfellow、Moritz Hardt 和 Been Kim。《显著性图的健全性检查》。神经信息处理系统进展，第 31 卷，2018年。

David Alvarez Melis 和 Tommi Jaakkola。通过自解释神经网络实现稳健的可解释性。《神经信息处理系统进展》，第31卷，2018年。

Martin Arjovsky、Soumith Chintala 和 L´eon Bottou。Wasserstein生成对抗网络。载于《国际机器学习会议论文集》，第214–223页。PMLR，2017年。

Nishanth Arun、Nathan Gaw、Praveer Singh、Ken Chang、Mehak Aggarwal、Bryan Chen、Katharina Hoebel、Sharut Gupta、Jay Patel、Mishka Gidwani 等。评估显著性图在医学影像异常定位中的可信度。《放射学：人工智能》，3(6)，2021年。

Shekoofeh Azizi、Basil Mustafa、Fiona Ryan、Zachary Beaver、Jan Freyberg、Jonathan Deaton、Aaron Loh、Alan Karthikesalingam、Simon Kornblith、Ting Chen等。大型自监督模型推进医学图像分类。载于《IEEE/CVF国际计算机视觉会议论文集》，第3478–3488页，2021年。

Alina Jade Barnett、Fides Regina Schwartz、Chaofan Tao、Chaofan Chen、Yinhao Ren、Joseph Y Lo和Cynthia Rudin。使用基于案例的推理和深度学习进行可解释的乳腺X线图像分类。arXiv预印本 arXiv:2107.05605，2021年。Cher Bass、Mariana da Silva、Carole Sudre、Petru-Daniel Tudosiu、Stephen Smith和Emma Robinson。Icam：通过解耦表示和特征归因映射实现可解释分类。《神经信息处理系统进展》，33：7697–7709，2020年。Christian F. Baumgartner、Lisa M. Koch、Kerem Can Tezcan、Jia Xi Ang和Ender Konukoglu。使用Wasserstein生成对抗网络进行视觉特征归因。载于《IEEE计算机视觉与模式识别会议论文集》（CVPR），2018年6月。Moritz Bohle、Mario Fritz和Bernt Schiele。用于可解释分类的卷积动态对齐网络。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第10029–10038页，2021年。Wieland Brendel和Matthias Bethge。使用局部特征包模型近似卷积神经网络在ImageNet上取得了出人意料的良好效果。arXiv预印本 arXiv:1904.00760，2019年。Irem Cetin、Maialen Stephens、Oscar Camara和Miguel A Gonz´alez Ballester。Attri-vae：利用变分自编码器对医学图像进行基于属性的可解释表示。《计算机化医学成像与图形学》，第102158页，2022年。Chaofan Chen、Oscar Li、Daniel Tao、Alina Barnett、Cynthia Rudin和Jonathan K Su。这看起来像那样：用于可解释图像识别的深度学习。《神经信息处理系统进展》，32，2019年。Zhi Chen、Yijie Bei和Cynthia Rudin。用于可解释图像识别的概念白化。《Nature Machine Intelligence》，2(12)：772–782，2020年。Yunjey Choi、Minje Choi、Munyoung Kim、Jung-Woo Ha、Sunghun Kim和Jaegul Choo。StarGAN：用于多域图像到图像转换的统一生成对抗网络。载于《IEEE计算机视觉与模式识别会议论文集》，第8789–8797页，2018年。Joseph Paul Cohen、Rupert Brooks、Sovann En、Evan Zucker、Anuj Pareek、Matthew P Lungren和Akshay Chaudhari。通过潜变量偏移实现Gifsplanation：一种用于胸部X线片反事实生成的简单自编码器方法。载于《深度学习医学成像》，第74–104页。PMLR，2021年。Berkeley J Dietvorst、Joseph P Simmons和Cade Massey。算法厌恶：人们在看到算法出错后会错误地回避算法。《实验心理学杂志：综合》，144(1)：114，2015年。Susanne Gaube、Harini Suresh、Martina Raue、Alexander Merritt、Seth J Berkowitz、Eva Lermer、Joseph F Coughlin、John V Guttag、Errol Colak和Marzyeh Ghassemi。听从人工智能的指示：临床决策辅助工具部署中的易感性。《NPJ数字医学》，4(1)：1–8，2021年。

Thomas Grote 和 Philipp Berens。《医疗保健中算法决策的伦理问题》。《医学伦理学杂志》，46(3):205–211，2020年。Ishaan Gulrajani、Faruk Ahmed、Martin Arjovsky、Vincent Dumoulin 和 Aaron C Courville。《改进 Wasserstein 生成对抗网络的训练》。《神经信息处理系统进展》，30，2017年。Tessa Han、Suraj Srinivas 和 Himabindu Lakkaraju。《我应该选择哪种解释？从函数逼近的角度刻画事后解释》arXiv 预印本 arXiv:2206.01254，2022年。Jeremy Irvin、Pranav Rajpurkar、Michael Ko、Yifan Yu、Silviana Ciurea-Ilcus、Chris Chute、Henrik Marklund、Behzad Haghgoo、Robyn Ball、Katie Shpanskaya 等。《CheXpert：一个包含不确定性标签并可进行专家对比的大型胸部X线影像数据集》。载于《人工智能促进协会年会论文集》，第33卷，第590–597页，2019年。Syed Ashar Javed、Dinkar Juyal、Harshith Padigela、Amaro Taylor-Weiner、Limin Yu 和 Aaditya Prakash。《加性多示例学习：病理学的内在可解释性》。arXiv 预印本 arXiv:2206.01794，2022年。Shalmali Joshi、Oluwasanmi Koyejo、Been Kim 和 Joydeep Ghosh。《XGEMs：生成示例以解释黑盒模型》。arXiv 预印本 arXiv:1806.08867，2018年。Ashkan Khakzar、Yang Zhang、Wejdene Mansour、Yuezhi Cai、Yawei Li、Yucheng Zhang、Seong Tae Kim 和 Nassir Navab。《通过识别信息性输入特征解释 COVID-19 和胸部病理模型预测》。载于《医学图像计算与计算机辅助干预国际会议》，第391–401页。Springer，2021年。Diederik P Kingma 和 Jimmy Ba。《Adam：一种随机优化方法》。arXiv 预印本 arXiv:1412.6980，2014年。Pang Wei Koh、Thao Nguyen、Yew Siang Tang、Stephen Mussmann、Emma Pierson、Been Kim 和 Percy Liang。《概念瓶颈模型》。载于《国际机器学习会议》，第5338–5348页。PMLR，2020年。Scott M Lundberg 和 Su-In Lee。《一种解释模型预测的统一方法》。《神经信息处理系统进展》，30，2017年。Daniel Nemirovsky、Nicolas Thiebaut、Ye Xu 和 Abhishek Gupta。《CounterGAN：利用残差生成对抗网络生成逼真的反事实》。arXiv 预印本 arXiv:2009.05199，2020年。Ha Q Nguyen、Khanh Lam、Linh T Le、Hieu H Pham、Dat Q Tran、Dung B Nguyen、Dung D Le、Chi M Pham、Hang TT Tong、Diep H Dinh 等。《VinDr-CXR：一个带有放射科医师标注的开放胸部X线片数据集》。arXiv 预印本 arXiv:2012.15029，2020年。

Marco Tulio Ribeiro、Sameer Singh 和 Carlos Guestrin。《为什么我应该信任你？解释任意分类器的预测》。载于《第22届ACM SIGKDD知识发现与数据挖掘国际会议论文集》，第1135–1144页，2016。

Cynthia Rudin。停止为高风险决策解释黑箱机器学习模型，改用可解释模型。Nature Machine Intelligence，1(5)：206–215，2019年。

Jo Schlemper、Ozan Oktay、Michiel Schaap、Mattias Heinrich、Bernhard Kainz、Ben Glocker 和 Daniel Rueckert。注意力门控网络：学习利用医学图像中的显著区域。Medical Image Analysis，53：197–207，2019年。

Kathryn Schutte、Olivier Moindrot、Paul H´erent、Jean-Baptiste Schiratti 和 Simon J´egou。使用 StyleGAN 实现医学图像上深度学习模型的视觉可解释性。arXiv 预印本 arXiv:2101.07563，2021年。

Ramprasaath R Selvaraju、Michael Cogswell、Abhishek Das、Ramakrishna Vedantam、Devi Parikh 和 Dhruv Batra。梯度加权类激活映射：通过基于梯度的定位实现深度网络的视觉解释。载于 IEEE国际计算机视觉会议论文集，第618–626页，2017年。

Sumedha Singla、Brian Pollack、Junxiang Chen 和 Kayhan Batmanghelich。通过渐进式夸张进行解释。arXiv 预印本 arXiv:1911.00483，2019年。

Jost Tobias Springenberg、Alexey Dosovitskiy、Thomas Brox 和 Martin Riedmiller。力求简洁：全卷积网络。arXiv 预印本 arXiv:1412.6806，2014年。

Guolei Sun、Thomas Probst、Danda Pani Paudel、Nikola Popovi´c、Menelaos Kanakis、Jagruti Patel、Dengxin Dai 和 Luc Van Gool。用于多任务学习的任务切换网络。载于 IEEE/CVF 计算机视觉国际会议论文集，第8291–8300页，2021年。

Philipp Tschandl、Christoph Rinner、Zoe Apalla、Giuseppe Argenziano、Noel Codella、Allan Halpern、Monika Janda、Aimilios Lallas、Caterina Longo、Josep Malvehy 等。人机协作进行皮肤癌识别。Nature Medicine，26(8):1229–1234，2020年。

Xiaosong Wang、Yifan Peng、Le Lu、Zhiyong Lu、Mohammadhadi Bagheri 和 Ronald M Summers。ChestX-ray8：医院规模的胸部X光数据库，以及针对常见胸部疾病弱监督分类与定位的基准测试。载于 IEEE计算机视觉与模式识别会议论文集，第 2097–2106 页，2017年。

Yandong Wen、Kaipeng Zhang、Zhifeng Li 和 Yu Qiao。深度人脸识别的判别性特征学习方法。载于 欧洲计算机视觉会议论文集，第 499–515 页。Springer，2016。

Adam White 和 Artur d’Avila Garcez。适用于任意分类器的可测量反事实局部解释。arXiv 预印本 arXiv:1908.03020，2019年。

Yiqi Yan、Jeremy Kawahara 和 Ghassan Hamarneh。通过视觉注意力进行黑色素瘤识别。载于《医学影像信息处理：第26届国际会议，IPMI 2019》，中国香港，2019年6月2日至7日，论文集第26卷，第793–804页。Springer，2019年。

Wenwu Ye、Jin Yao、Hui Xue 和 Yi Li。采用概率CAM池化的弱监督病灶定位。arXiv 预印本 arXiv:2005.14480，2020年。

Bolei Zhou、Aditya Khosla、Agata Lapedriza、Aude Oliva 和 Antonio Torralba。《学习用于判别式定位的深度特征》。载于 IEEE计算机视觉与模式识别会议论文集，第 2921–2929 页，2016年。

附录 A. 补充评估

A.1. 反事实生成示例

通过将特定类别的视觉归因图添加到输入图像（即 $\hat{\mathbf{x}} = \mathbf{x} + M_c(\mathbf{x})$）获得的反事实图像示例如图4所示。

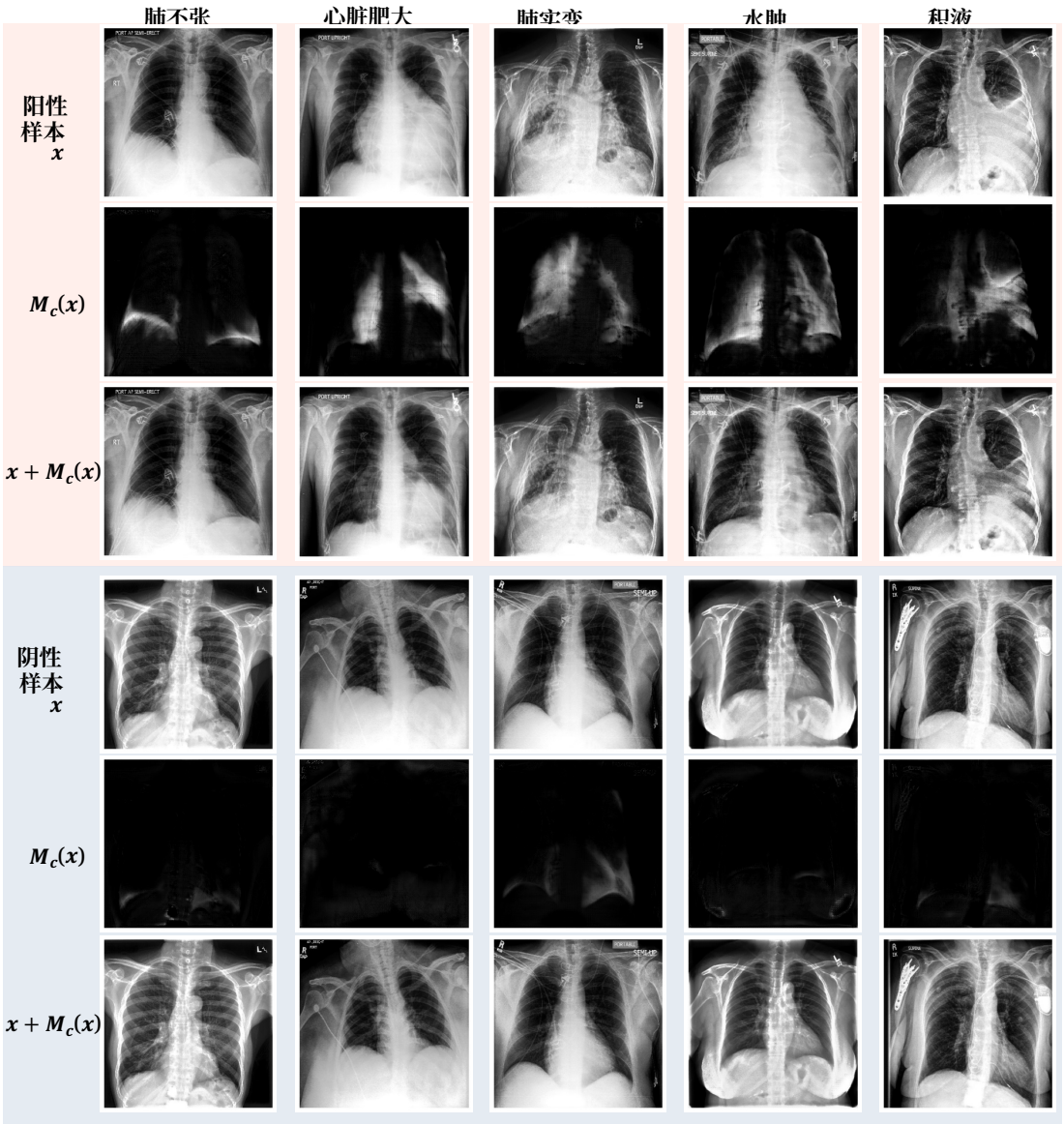


图 4：反事实图像生成。顶部几组行中的示例展示了包含不同类别证据的输入图像 c。对于这些图像，通过添加 $M_c(\mathbf{x})$，移除了类别 c 的证据。底部几组行中的图像不包含类别 c的证据。添加 $M_c(\mathbf{x})$的输出后，这些图像基本保持不变。

A.2. 全局可解释性

人们通常会区分局部解释和全局解释：局部解释说明特定输入图像的预测结果，而全局解释则说明机器学习算法整体的决策机制（即针对所有输入图像的决策机制）。尽管本文主要关注局部可解释性，但我们也可以通过解读第2.2节介绍的正类和负类中心，以及逻辑回归分类器的类别特定权重，来深入了解分类器的决策机制。类别中心捕捉了相应类别的一些原型特征，而分类器权重则可以告诉我们，分类器针对每个类别会关注图像中的哪些区域。图5 展示了在 ChestX-ray8 数据集上训练的五种疾病的类别中心及相应分类器的权重。

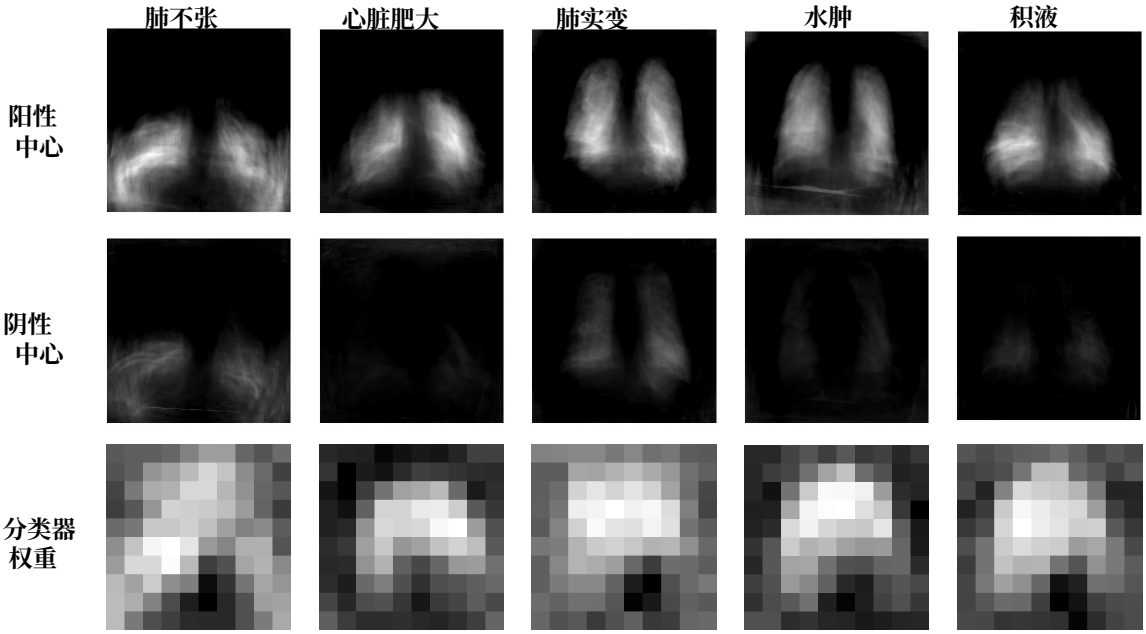


图 5：疾病的归因中心及相应的分类器权重为 Attri-Net 提供了全局解释。

A.3. 损失项的消融研究

关于用于训练 Attri-Net 的各项损失影响的消融研究见表3。ChestX-ray8 数据集中一张图像在所有组合下的归因示例见图6。

Ta表 3：四种损失项的消融研究。在 VinDr-CXR 数据集上进行评估 .

模型	损失项	分类 AUC	类别敏感度
Attri-Net$_{\text{cls}}$	$\mathcal{L}_{\text{cls}}$	0.9339	0.2516
Attri-Net$_{\text{cls_adv}}$	$\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{adv}}$	**0.9444**	0.1602
Attri-Net$_{\text{cls 对抗正则化}}$	$\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{reg}}$	0.9397	0.5259
Attri-Net$_{\text{all}}$	$\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{ctr}}$	0.9405	**0.5509**

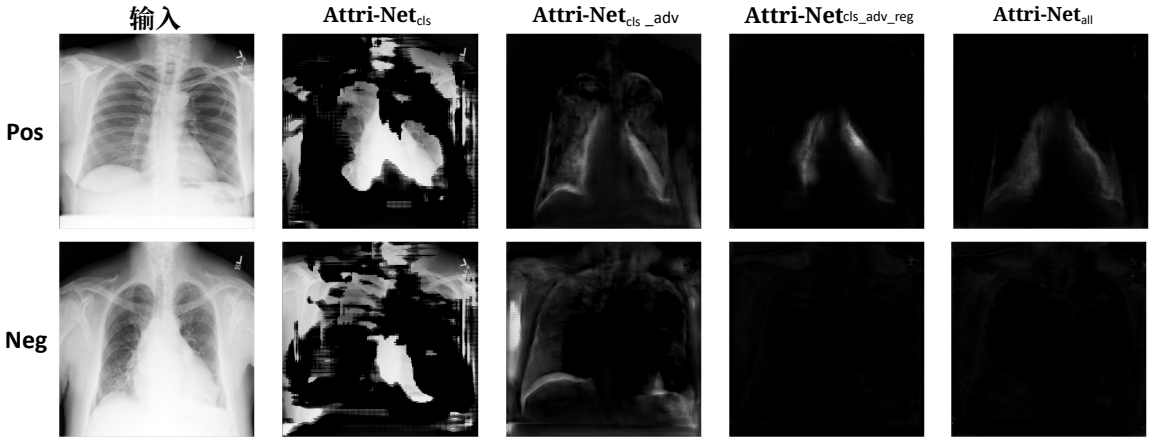


图 6：在 ChestX-ray8 上，针对不同损失子集，心脏肥大疾病的正负样本所得到的归因图示例。

A.4. 类别敏感性图像网格

第 3 节中的类别敏感性评估 基于 Bohle 等人 (2021年) 提出的类别敏感性网格。图7、图8和图 9 分别展示了 CheXpert 数据集上所有研究类别的此类网格示例，所用方法依次为 Attri-Net、CoDA-Nets 和 Gifsplanation。

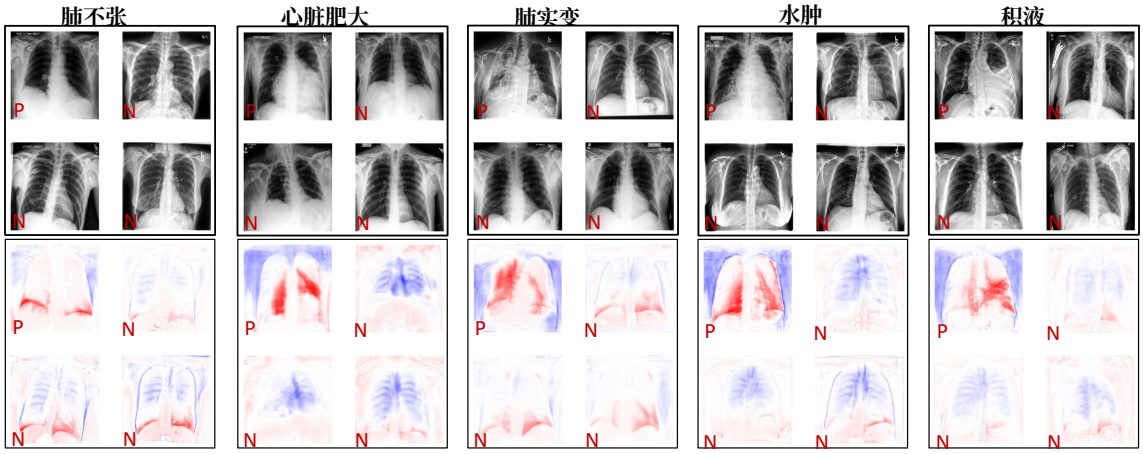


图 7：使用 Attri-Net 获得的类别敏感性图像网格。第一行显示图像网格，第二行显示相应的归因图。P 和 N 分别表示类别正例和类别负例。

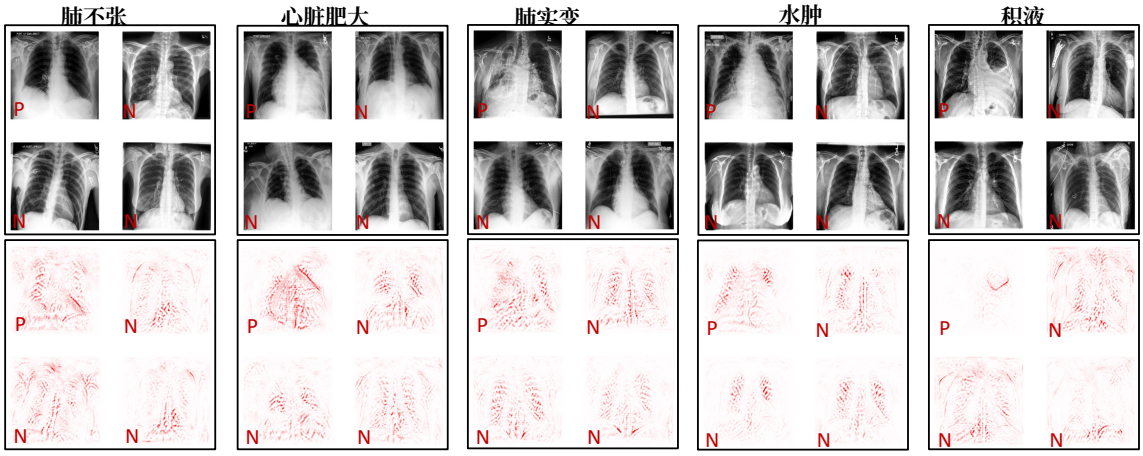


图 8：使用 CoDA-Nets 获得的类别敏感性图像网格。第一行显示图像网格，第二行显示相应的归因图。P 和 N 分别表示类别正例和类别负例。

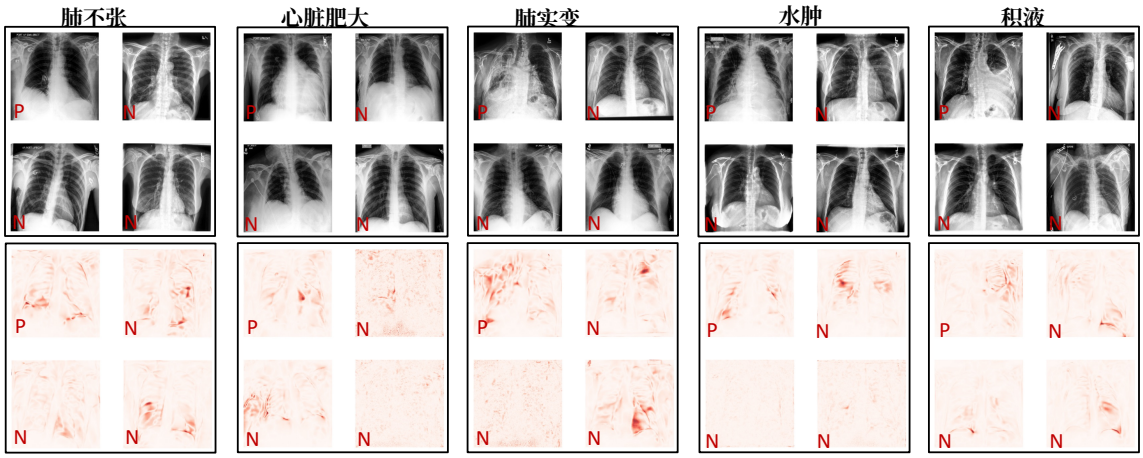


图9：使用 Gifsplanation 获得的类别敏感性图像网格。第一行显示图像网格，第二行显示相应的归因图。P 和 N 分别表示类别阳性和类别负例。

A.5. ChestX-ray8 和 VinDr-CXR 的示例解释

图10 和图 11 分别展示了使用所有对比方法从 ChestX-ray8 和 VinDr-CXR 数据集中得到的视觉归因的其他示例。

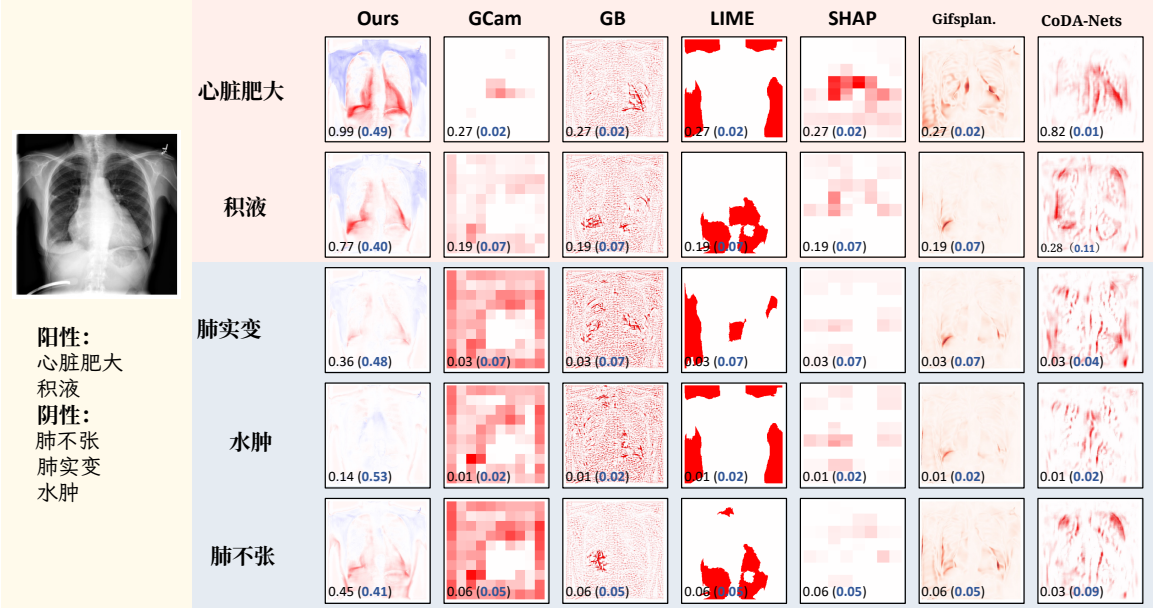


图10：ChestX-ray8 数据集中的一张示例图像的解释 .

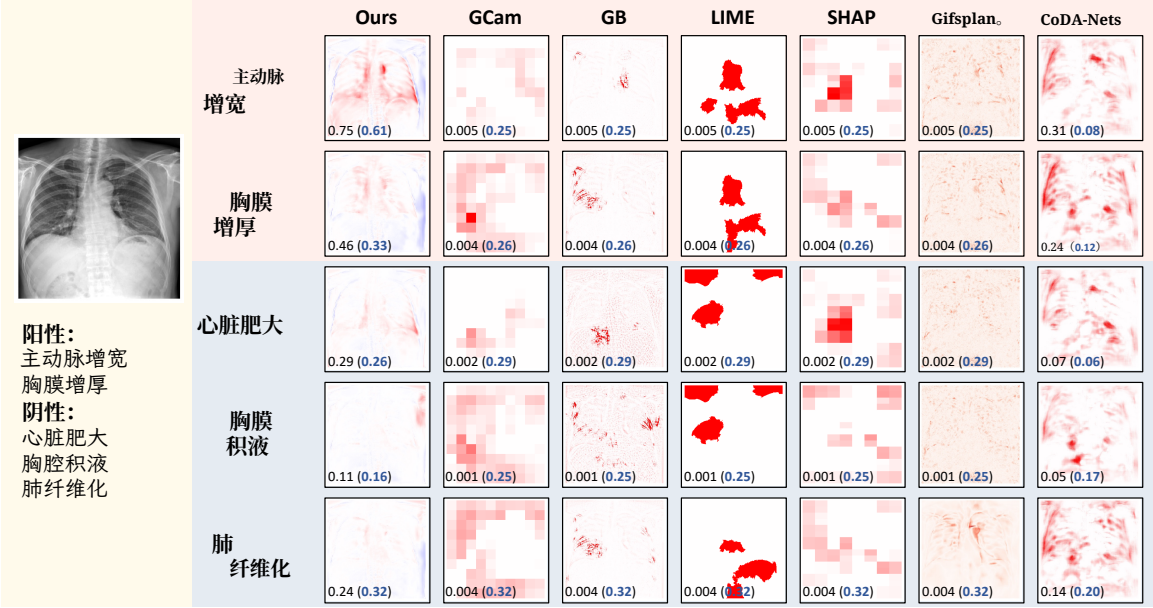


图 11：VindrCXR 数据集中一个示例图像的解释。

附录 B. 补充训练细节

B.1. 判别器训练

Attri-Net 框架需要并行训练一个判别器函数 D 和类别归因生成器 M。判别器的权重参数 θ 通过独立的梯度更新步骤，使用 Wasserstein 生成对抗网络（Arjovsky 等人，2017年）目标函数计算。完整的判别器优化目标如下：

$$\min_{\theta} \sum_{c} \mathop{\mathbb{E}}_{\mathbf{x}\sim p(\mathbf{x}|y_c=0)} [D_c(\mathbf{x}|\theta)] + \mathop{\mathbb{E}}_{\mathbf{x}\sim p(\mathbf{x}|y_c=1)} [D_c(\mathbf{x}+M_c(\mathbf{x})|\theta)] \,,$$

其中省略了梯度惩罚损失，该损失可确保判别器满足 Wasserstein 生成对抗网络目标函数（Gulrajaniet al.，2017年）规定的 1-Lipschitz 约束。

B.2. 网络架构

Attri-Net 框架中的归因图生成器和判别器的网络架构分别如表4 和表5所示。L 表示输入/输出特征的长度，N 表示输出通道数，K 表示卷积核大小。

表 4：Attri-Net 类别归因生成网络架构。

层	输入 → 输出	层信息
任务嵌入层	任务代码 $\mathbf{t}_c \rightarrow$ 任务嵌入 $\mathbf{t}'_c$	$8 \times$ 全连接层(L100,L100)
下采样	（输入图像 $\mathbf{x}$、$\mathbf{t}'_c$） $\rightarrow$ **输出**$_{\text{down}}$	自适应卷积：卷积(N64, K7x7)、自适应实例归一化、修正线性单元 自适应卷积：卷积(N128, K4x4)、自适应实例归一化、修正线性单元 自适应卷积：卷积(N256, K4x4)、自适应实例归一化、修正线性单元
瓶颈层	（**输出**$_{\text{down}}$, **时间**$'_c$） $\rightarrow$ **输出**$_{\text{bn}}$	自适应残差块：卷积(N256, K3x3)、自适应实例归一化、修正线性单元 自适应残差块：卷积(N256, K3x3)、自适应实例归一化、修正线性单元 自适应残差块：卷积(N256, K3x3,)、自适应实例归一化、修正线性单元 自适应残差块：卷积(N256, K3x3)、自适应实例归一化、修正线性单元 自适应残差块：卷积(N256, K3x3), 自适应实例归一化, 修正线性单元 自适应残差块：卷积(N256, K3x3), 自适应实例归一化, 修正线性单元
上采样	（**输出**$_{\text{bn}}$,$\mathbf{t}'_c$） $\rightarrow$ **输出**$_{\text{u p}}$	自适应反卷积(N128, K4x4), 自适应实例归一化, 修正线性单元 自适应反卷积(N64, K4x4), 自适应实例归一化, 修正线性单元 卷积(N1, K7x7)
输出层	(**x**, **输出**$^{\text{up}}$) $\rightarrow M_c$(**x**)	M_c(**x**) $=$ 双曲正切(**x** $+$**输出**$^{\text{up}}$) $-$**x**

表5：Attri-Net 判别器网络架构。

层	输入 → 输出	层信息
任务嵌入层	任务代码 $\mathbf{t}_c \rightarrow$ 任务嵌入 $\mathbf{t}'_c$	$8 \times$全连接层(L100,L100)
输入层		自适应卷积：卷积(N64, K4x4)、自适应实例归一化、修正线性单元
隐藏层	（$\mathbf{x}/\hat{}\mathbf{x}$,$\mathbf{t}'_c$） $\rightarrow$ **输出**$_{\text{hid}}$	自适应卷积：卷积(N128, K4x4)、自适应实例归一化、修正线性单元 自适应卷积：卷积(N256, K4x4)、自适应实例归一化、修正线性单元 自适应卷积：卷积(N512, K4x4)、自适应实例归一化、修正线性单元 自适应卷积：卷积(N1024, K4x4)，自适应实例归一化，修正线性单元 自适应卷积：卷积(N2048, K4x4)，自适应实例归一化，修正线性单元
输出层	**输出**$_{\text{hid}} \rightarrow \mathcal{L}^{(c)}_{\text{adv}}$	卷积(N1, K3x3)