

中图法分类号: TP391.41 文献标识码: A 文章编号: 1006-8961(XXXX)XX-0001-17

论文引用格式: Dong Peiyu, Wang Yusen, Bai Chenxi, Zhai Rui, Lu Ke. Latent Conditional Diffusion Purification for Fair Medical Image Diagnosis [J/OL]. Journal of Image and Graphics, XXXX: 1-17. DOI: 10.11834/jig.260326. (董沛宇, 王雨森, 白晨希, 翟锐, 吕科. 面向医学图像公平诊断的潜空间条件扩散去偏[J/OL]. 中国图象图形学报, XXXX: 1-17. DOI: 10.11834/jig.260326.) [DOI: 10.11834/jig.260326]

面向医学图像公平诊断的潜空间条件扩散去偏

董沛宇^{1,2}, 王雨森^{1,2}, 白晨希^{1,2}, 翟锐^{1,2*}, 吕科³

1. 河南大学软件学院, 河南 开封 475004; 2. 河南大学河南省智能数据处理工程研究中心, 河南 开封 475000; 3. 中国科学院大学工程科学学院, 北京 100049

摘要: 目的 医学图像诊断模型易受年龄、性别和种族等人口统计学属性影响, 造成子群间诊断性能差异。现有特征层去偏方法多通过对抗学习或相关性约束抑制人口统计学相关信息, 但当其与病理特征耦合时, 直接抑制可能削弱疾病判别信息, 难以兼顾诊断性能与公平性。为此, 本文提出一种潜空间条件扩散去偏方法, 用于医学图像公平诊断中的特征学习。**方法** 通过双空间正交约束降低潜在病理特征与人口统计学特征的相关性, 对初步解耦的病理特征施加受控高斯扰动, 以人口统计学属性和扩散时间步为条件完成去噪重构, 并采用子群自适应阈值校准预测结果。**结果** 在 CheXpert 的 4 项病理分类任务上, 本文方法的准确率 (accuracy, Acc)、受试者工作特征曲线下面积 (area under the curve, AUC)、交叉子群 AUC 差异和等机会差异分别为 0.796、0.857、0.058 和 0.108。与 7 种对比方法在相应指标上的最佳平均结果相比, 平均准确率和 AUC 分别由 0.781 和 0.841 提高至 0.796 和 0.857, 等机会差异由 0.126 降至 0.108, 交叉子群 AUC 差异为 0.058, 与最佳对比结果持平。在 MIMIC-CXR 跨数据集分布外测试中, 相较于公平性表现最佳的对比方法, 交叉子群 AUC 差异和等机会差异分别由 0.121 和 0.287 降至 0.085 和 0.166。**结论** 所提出的潜空间条件扩散去偏方法可在保持疾病判别能力的同时减小人口统计学子群间的诊断性能差异, 为医学图像公平诊断提供了一种潜空间特征重构的去偏思路。

关键词: 医学图像; 公平诊断; 去偏学习; 条件扩散模型; 人口统计学属性

Latent Conditional Diffusion Purification for Fair Medical Image Diagnosis

Dong Peiyu^{1,2}, Wang Yusen^{1,2}, Bai Chenxi^{1,2}, Zhai Rui^{1,2*}, Lu Ke³

1. School of Software, Henan University, Kaifeng 475004, Henan, China; 2. Henan Provincial Engineering Research Center of Intelligent Data Processing, Henan University, Kaifeng 475000, Henan China; 3. School of Engineering Science, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: Objective Deep learning has shown promise for improving the efficiency and consistency of medical image diagnosis. However, deep learning models may encode demographic attributes, such as age, sex, and race, in ways that contribute to performance disparities across patient subgroups. High overall diagnostic accuracy does not preclude clinically relevant performance gaps across demographic subgroups, which may limit trustworthy clinical deployment. Existing debi-

收稿日期: 2026-06-07; 修回日期: 2026-08-18

* 通信作者: 翟锐 zr@henu.edu.cn

基金项目: 国家自然科学基金重点项目 (62331001); 国家自然科学基金面上项目 (82371980); 河南省科技研发计划联合基金重点项目 (235200810031); 河南省科技攻关项目 (252102211027); 京津冀基础研究合作专项 (H2024202009)

Supported by: National Natural Science Foundation of China (62331001, 82371980); Joint Fund of the Science and Technology R&D Program of Henan Province (235200810031); Henan Science and Technology Research Project (252102211027); Jing-Jin-Ji Incorporation Project of the Natural Science Foundation of Hebei Province (H2024202009)

using methods often use adversarial suppression or discriminative feature removal to reduce demographic information in learned representations. However, when demographic cues are entangled with pathology-related features, suppressing demographic information may also weaken disease-relevant representations. Balancing diagnostic performance and subgroup fairness therefore remains challenging in medical image diagnosis. **Method** To address this challenge, we propose latent conditional diffusion purification (LCDP) for fair medical image diagnosis. LCDP formulates debiasing as the conditional reconstruction of pathology-relevant latent representations rather than the direct removal of demographic information. The framework comprises a dual-branch encoder for representation decomposition, a latent conditional diffusion reconstruction module for feature reconstruction, a disease classifier, and a subgroup-adaptive threshold calibration strategy. First, the dual-branch encoder learns pathology-related and demographic representations separately. Dual-space orthogonality constraints are imposed at the sample-subspace and feature-channel levels to reduce the association between pathology-related and demographic representations before diffusion-based reconstruction. Second, controlled Gaussian perturbation is applied to pathology-related latent representations to disrupt potential demographic shortcut patterns. A conditional denoising network then reconstructs pathology-relevant latent features using demographic attributes and diffusion timesteps as conditioning information. These signals modulate the denoising network through adaptive layer normalization but are not directly provided to the disease classifier, so that demographic information is used to guide feature reconstruction rather than disease prediction. Training jointly optimizes the disease classification loss, dual-space orthogonality loss, demographic-attribute classification loss, and diffusion noise-prediction loss. During inference, the reconstructed representation is used for disease prediction, and subgroup-adaptive thresholds calibrated on the validation set are applied for decision calibration. **Result** Experiments were conducted on four CheXpert chest X-ray classification tasks: cardiomegaly, pleural effusion, pneumonia, and fracture. Sex, age, and race were evaluated as demographic attributes. To assess performance under distributional imbalance, we constructed a stress-test setting by amplifying subgroup differences in positive rates and sample composition. LCDP was compared with empirical risk minimization (ERM) and representative debiasing methods, including fairness via column-row space orthogonality (FCRO), learning not to learn (LNTL), fair adaptive batch normalization (FairAdaBN), adversarial learning (AL), fair re-fusion after disentanglement (FairREAD), and learning adaptive agreement from a biased council (Ada-ABC). Across the four CheXpert tasks, LCDP achieved a mean accuracy of 0.796 and a mean AUC of 0.857, with mean intersectional-subgroup AUC difference and equal opportunity difference values of 0.058 and 0.108, respectively. Compared with the best mean results among the competing methods, LCDP improved the mean accuracy and AUC by 0.015 and 0.016, respectively, reduced the mean equal opportunity difference by 0.018, and achieved a comparable mean subgroup AUC difference. Ablation experiments indicated that dual-space orthogonality constraints, diffusion-based reconstruction, demographic-attribute conditioning, and subgroup-adaptive threshold calibration each contributed to the overall fairness–performance balance. Additional analyses were performed to examine the robustness and subgroup-level performance of the proposed approach. Five-fold cross-validation was used to reduce the influence of data partitioning, and intersectional subgroup analyses were conducted to assess performance for potentially disadvantaged combinations of demographic attributes. Sensitivity analyses of the orthogonality-loss weight and inference timestep showed that LCDP maintained a relatively stable fairness–accuracy balance across settings. In cross-dataset out-of-distribution evaluation on MIMIC-CXR, LCDP achieved mean subgroup AUC difference and equal opportunity difference values of 0.085 and 0.166, respectively, suggesting improved robustness of subgroup performance under the evaluated distribution shift. Grad-CAM and t-SNE analyses were qualitatively consistent with reduced attention to non-lesion regions and weaker subgroup clustering in the learned representations. **Conclusion** This study presents a latent conditional diffusion purification framework for fair medical image diagnosis. By combining dual-space orthogonality constraints, conditional diffusion-based reconstruction, and subgroup-adaptive threshold calibration, LCDP achieved smaller intersectional subgroup performance gaps while maintaining competitive diagnostic performance in the evaluated settings. These findings suggest that conditional latent reconstruction may be a useful debiasing strategy when pathology-related and demographic cues are entangled.

Key words: medical image; fair diagnosis; debiasing learning; conditional diffusion model; demographic attribute

© 中国图象图形学报版权所有

论文引用格式: Dong P Y, Wang Y S, Bai C X, Zhai R and Lv K. 2026. Latent Conditional Diffusion Purification for Fair Medical Image Diagnosis(董沛宇, 王雨森, 白晨希, 翟锐, 吕科. 2026. 面向医学图像公平诊断的潜空间条件扩散去偏. 中国图象图形学报)[DOI:10. 11834/jig. 260326]

0 引言

深度学习技术的发展有效推动了医学图像智能诊断的研究与应用,在多种疾病诊断任务中表现出显著的辅助诊断价值(Zhang 等, 2024; 邢素霞 等, 2023)。然而,多项研究表明,无论是传统深度学习模型还是最新的医学图像基础模型,其诊断性能在不同人口统计学子群之间存在系统性差异,导致部分群体面临更高的漏诊或误诊风险(Jin 等, 2024; Larrazabal 等, 2020; Seyyed-Kalantari 等, 2021; Yang 等, 2024, 2025),模型整体诊断精度的提升并未同步缩小群体间的性能差距。造成上述差异的原因包括人口统计学子群样本数量不均、子群间疾病患病率差异以及模型对人口统计学相关非病理特征利用。已有研究表明,医学图像存在人口统计学信息,病理特征中也可能保留此类信息,此外,图像标记、成像体位和图像覆盖范围等非病理因素也可能被模型利用(DeGrave 等, 2021; Gichoya 等, 2022; Glocker 等, 2023; Lotter, 2024)。当这些与人口统计学属性相关但缺乏稳定病理依据的特征与疾病标签形成统计关联时,模型可能将其作为疾病预测捷径,从而扩大不同人口统计学子群之间的性能差异(Brown 等, 2023)。因此,如何有效缓解不同人口统计学子群之间的诊断性能差异,已成为提升医学图像诊断模型公平性与可靠性的重要问题(Ricci Lara 等, 2022)。

为缩小不同人口统计学子群之间的诊断性能差异,现有研究主要从数据、特征和决策三个层面开展改进(Xu 等, 2024; 王玫 等, 2024)。数据层方法通过重采样、样本重加权或生成式增强调整不同子群的训练分布。Ktena 等人(2024)利用条件生成模型生成满足特定疾病标签和人口统计学属性组合的合成图像,以补充训练集中代表性不足的疾病类别和人口统计学子群,并改善模型在分布偏移条件下的子群表现。Skorupko 等人(2024)根据性别、年龄、体质指数等条件生成心脏磁共振图像,用于缓解特定

子群的样本不足。决策层方法不直接改变模型学习到的特征,而是通过阈值调整或预测校准减小不同子群的输出差异。Shui 等人(2023)提出 Cluster-Focal 方法,通过识别并聚类校准较差的样本,再利用组别焦点损失减小子群间的差异。特征层方法主要通过约束特征学习过程,降低病理特征对人口统计学属性或伪相关信息的依赖。Ada-ABC 通过集成多个偏置分类器提取数据中的伪相关信息,利用自适应一致性约束引导模型学习疾病目标特征,以减少对数据偏差的依赖(Luo 等, 2025)。FairAdaBN 依据人口统计学属性自适应调整特征的归一化统计量,并引入子群统计差异约束(Xu 等, 2023)。尽管上述方法在改善模型公平性方面取得了一定进展,但仍存在不同程度的局限:数据层方法主要调整训练样本分布,难以限制模型对图像中人口统计学相关非病理特征的利用,决策层方法主要作用于模型输出端,能够修正不同子群的预测差异,但无法改变模型内部已经形成的非病理特征依赖,特征层方法能够直接干预特征学习过程,但病理特征与人口统计学相关非病理特征在特征空间中高度耦合,通过对抗学习、互信息约束或正交正则直接抑制相关特征时,可能同时削弱疾病分类所需的病理判别信息(Jones 等, 2024; Oakden-Rayner 等, 2020)。

由上述分析可知,医学图像公平诊断中的关键问题并不是简单抑制人口统计学相关信息,而是病理特征与人口统计学相关非病理特征同时存在于特征空间中,增强特征抑制既难以充分去除人口统计学相关信息,又可能削弱疾病分类所需的病理判别信息。因此,本文关注如何在降低模型对人口统计学相关非病理特征依赖的同时,尽可能保留疾病分类所需的病理判别信息。基于这一问题定位,本文将特征去偏设计为显式解耦与条件重构两个阶段。第一阶段,利用双空间正交约束,从样本子空间和特征通道两个层面降低病理特征与人口统计学特征之间的相关性,实现初步解耦,第二阶段,以初步解耦的病理特征为重构对象,通过受控高斯扰动引入扩散过程,并在扩散时间步和人口统计学属性的条件调制下进行去噪重构,进一步减少残留人口统计学相关非病理信息的影响,同时保留疾病分类所需的病理判别信息。现有医学图像去偏方法与本文方法的处理思路对比如图 1 所示。

为实现上述双阶段去偏过程,本文设计了面向
© 中国图象图形学报版权所有

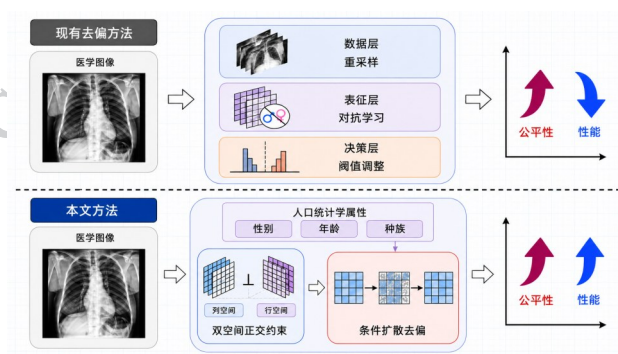


图1 现有医学图像去偏方法与本文方法的对比示意图

Fig. 1 Schematic comparison of existing medical image debiasing methods and the proposed method

医学图像公平诊断的潜空间条件扩散去偏框架 (Latent Conditional Diffusion Purification, LCDP)。医学图像经双分支编码器分别提取病理特征和人口统计学特征,并根据人口统计学特征构建低秩特征子空间,通过列空间正交约束降低病理特征对该子空间的依赖,同时利用行空间正交约束减小两类特征通道之间的线性相关,得到初步解耦的病理特征。在此基础上,对病理特征施加受控高斯扰动,并利用扩散时间步和人口统计学属性对条件去噪网络进行调制,实现病理特征的条件重构。模型通过疾病分类损失、扩散噪声预测损失、双空间正交损失和人口统计学属性分类损失进行联合优化。

本文的主要贡献如下:(1)提出面向医学图像公平诊断的潜空间条件扩散去偏框架。该框架在不改变训练样本分布的前提下,直接作用于分类网络提取的潜在病理特征,学习兼顾疾病判别性与群体公平性的特征表示,同时,利用验证集确定的子群自适应阈值校准不同人口统计学交叉子群的决策边界,从特征与决策两个层面改善模型的群体公平性。(2)针对病理特征与人口统计学相关非病理特征高度耦合,以及直接抑制人口统计学相关特征可能造成病

理判别信息损失的问题,本文设计了双空间正交约束与条件扩散重构协同的特征去偏方法。双空间正交约束从样本子空间和特征通道两个层面降低病理特征与人口统计学特征的相关性。在此基础上,利用人口统计学属性和扩散时间步引导去噪重构,对初步解耦后的病理特征进行条件重构,以降低非病理特征对疾病预测的干扰,并保留疾病分类所需的病理判别信息。(3)为了验证方法的有效性,本文在CheXpert数据集的心脏肥大、胸腔积液、肺炎和骨折四项分类任务上开展五折交叉验证,并通过偏差压力测试、四项任务的组件消融、弱势交叉子群分析以及MIMIC-CXR跨数据集测试,从疾病诊断性能、人口统计学子群公平性和跨数据集稳定性等方面对所提方法进行评估。实验结果表明,LCDP能够在保持较高疾病诊断性能的同时,缩小不同人口统计学子群之间的性能差异。

1 方法

本文提出的潜空间条件扩散去偏框架如图2所示,整体由双空间正交约束模块、潜空间条件扩散去偏模块和分类器以及子群自适应阈值校准策略组成。模型首先通过双分支编码器分别提取病理特征和人口统计学特征,并利用双空间正交约束从样本子空间和特征通道两个层面降低二者的相关性,随后,对初步解耦的病理特征施加受控高斯扰动,并将人口统计学属性编码和扩散时间步编码作为条件信号调制去噪重构过程。重构后的病理特征用于疾病预测,以在保留疾病判别信息的同时减小不同人口统计学子群之间的性能差异。本文涉及的主要符号及其含义如表1所示。

表1 本文涉及的主要符号及含义

Table 1 Main symbols and their definitions

符号	含义	符号	含义
X	医学图像	y	疾病标签
A	人口统计学属性向量	Z_T	病理特征矩阵
Z_A	人口统计学特征矩阵	S_A	低秩人口统计学特征子空间
C_{SA}, C_t	人口统计学属性条件和时间步条件	C	全局条件控制信号
Z_i	加噪病理特征	Z_p	重构后的病理特征

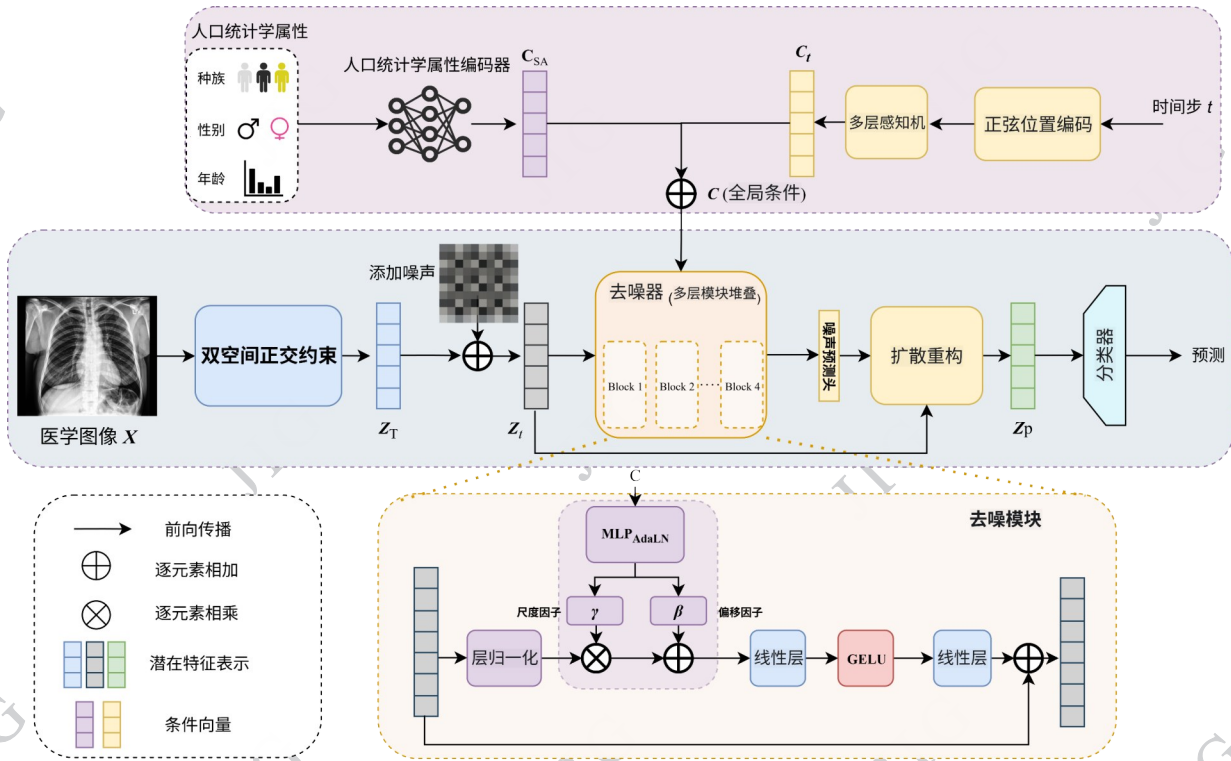


图 2 潜空间条件扩散去偏框架总体结构

Fig. 2 Overall framework of the proposed latent conditional diffusion purification method

1.1 双空间正交约束

为实现病理特征与人口统计学特征的初步解耦,本文设计双空间正交约束模块,其结构如图 3 所示。该模块包含两个参数相互独立的 DenseNet 编码分支。同一输入医学图像分别经过病理特征编码

器和人口统计学特征编码器,得到病理特征 Z_T 和人口统计学特征 Z_A 。为使人口统计学特征编码分支能够有效表达人口统计学属性信息,在其末端设置辅助属性预测头,并采用人口统计学属性分类损失 L_{fA} 进行监督。

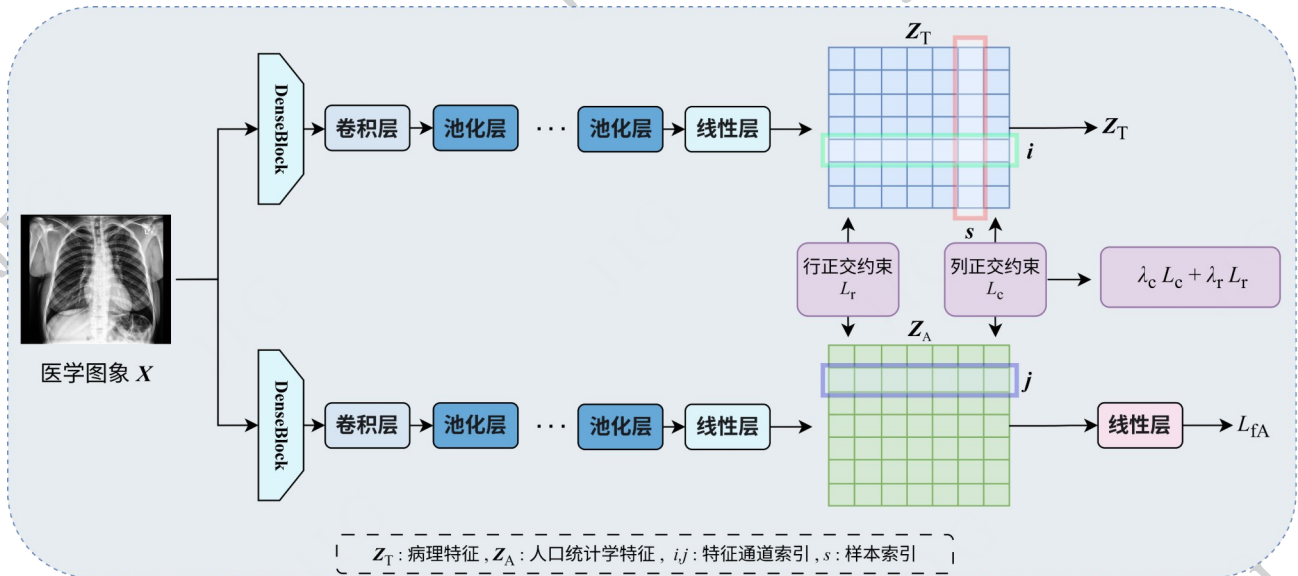


图 3 双空间正交约束模块结构示意图

Fig. 3 Schematic architecture of the dual-space orthogonality constraint module

对于由 n 个样本组成的训练批次,两个编码分支输出的病理特征和人口统计学特征分别组成特征矩阵 $\mathbf{Z}_T$ 和 $\mathbf{Z}_A$,二者的维度均为 $d_z \times n$,其中每一列对应一个样本,每一行对应一个特征通道, d_z 表示两个编码分支共同的特征维度。由于两个编码分支的参数相互独立并不能保证其输出特征保持低相关,本文进一步引入列空间正交约束和行空间正交约束,分别从低秩人口统计学特征子空间和特征通道相关性两个层面降低两类特征之间的相关性。

1.1.1 列空间正交约束

列空间正交约束用于从低秩人口统计学特征子空间层面降低病理特征对人口统计学相关方向的依赖。首先,对人口统计学特征矩阵 $\mathbf{Z}_A$ 进行奇异值分解 $\mathbf{Z}_A = \mathbf{U}_A \mathbf{\Sigma}_A \mathbf{V}_A^T$ 式中, $\mathbf{U}_A$ 和 $\mathbf{V}_A$ 分别为左奇异向量矩阵和右奇异向量矩阵, $\mathbf{\Sigma}_A$ 为奇异值对角矩阵,上标 T 表示矩阵转置。根据奇异值由大到小的排列顺序,从 $\mathbf{U}_A$ 中选取前 k 个左奇异向量,构成低秩人口统计学特征子空间 $\mathbf{S}_A$ 。其中, k 表示低秩人口统计学特征子空间维度,用于控制该子空间所包含的人口统计学相关特征方向。记病理特征矩阵 $\mathbf{Z}_T$ 的第 s 列为 $\mathbf{z}_T^{(s)}$,即第 s 个样本的病理特征。通过最小化各样本病理特征在 $\mathbf{S}_A$ 上的归一化投影,列空间正交损失定义为:

$$L_c = \sum_{s=1}^n \frac{\|\mathbf{S}_A^T \mathbf{z}_T^{(s)}\|_2^2}{\|\mathbf{z}_T^{(s)}\|_2^2} \quad (1)$$

式中, L_c 为列空间正交损失, n 为当前批次中的样本数, $\mathbf{S}_A^T \mathbf{z}_T^{(s)}$ 表示第 s 个样本的病理特征在低秩人口统计学特征子空间上的投影,分母用于减小不同样本特征幅值对投影结果的影响。最小化 L_c 能够降低病理特征在人口统计学特征方向上的投影。由于该约束仅作用于由前 k 个左奇异向量构成的低秩子空间,其余特征空间仍可用于保留疾病预测所需的信息。

1.1.2 行空间正交约束

行空间正交约束关注病理特征与人口统计学特征在各特征通道上的相关性。对于特征矩阵 $\mathbf{Z}_T$ 和 $\mathbf{Z}_A$,每一行表示一个特征通道在当前小批量样本上的响应。将 $\mathbf{Z}_T$ 的第 i 行和 $\mathbf{Z}_A$ 的第 j 行分别记为 $\mathbf{z}_{T,i}$ 和 $\mathbf{z}_{A,j}$ 。对通道响应进行中心化后,通过最小化两类特征通道之间的批内协方差,定义行空间正交损失:

$$L_r = \frac{1}{d_z^2} \sum_{i=1}^{d_z} \sum_{j=1}^{d_z} \left[\frac{1}{n-1} (\mathbf{z}_{T,i} - \boldsymbol{\mu}_{T,i})^T (\mathbf{z}_{A,j} - \boldsymbol{\mu}_{A,j}) \right]^2 \quad (2)$$

式中, L_r 为行空间正交损失, d_z 为特征通道数, i 和 j 分别为病理特征通道和人口统计学特征通道的序号, $\mathbf{z}_{T,i}$ 和 $\mathbf{z}_{A,j}$ 分别为第 i 个病理特征通道和第 j 个人口统计学特征通道在当前小批量上的响应向量, $\boldsymbol{\mu}_{T,i}$ 和 $\boldsymbol{\mu}_{A,j}$ 分别表示对应通道响应向量在当前批次样本上的均值,方括号中的项表示两个特征通道之间的批内协方差。列空间正交约束和行空间正交约束分别面向不同层次的相关结构:前者通过低秩人口统计学特征子空间投影,降低病理特征对人口统计学属性主导方向的依赖,后者通过批内通道协方差,降低病理特征通道与人口统计学特征通道之间的成对线性相关。两项约束相互补充,双空间正交损失定义为:

$$L_o = \lambda_c L_c + \lambda_r L_r \quad (3)$$

式中, L_o 为双空间正交损失, λ_c 和 L_c 分别为列空间正交损失的权重系数和损失项, λ_r 和 L_r 分别为行空间正交损失的权重系数和损失项。经过双空间正交约束后,得到初步解耦的病理特征矩阵 $\mathbf{Z}_T$,并将其作为潜空间条件扩散去偏模块的输入。

1.2 潜空间条件扩散去偏

双空间正交约束通过限制病理特征在低秩人口统计学特征子空间中的投影以及两类特征通道之间的批内协方差,主要降低病理特征与人口统计学特征之间的显式线性相关,但对潜在特征中更复杂的非线性关联约束有限。针对上述约束范围,本文设计潜空间条件扩散去偏模块,将病理特征的进一步去偏建模为受控扰动下的条件扩散重构过程。该模块首先对初步解耦的病理特征施加与扩散时间步相关的高斯扰动,随后将人口统计学属性和扩散时间步编码为条件信号,通过自适应层归一化调制去噪网络,并根据预测噪声完成病理特征重构。在疾病分类任务的监督下,条件去噪网络学习从受扰动特征中恢复病理判别信息,以降低与人口统计学属性相关的非病理特征对疾病预测的影响。

1.2.1 条件控制信号构建

设每个样本的人口统计学属性向量为 $\mathbf{A}$ 。人口统计学属性编码器将其映射为 256 维人口统计学属性条件向量 $\mathbf{C}_{SA}$ 。该编码器以给定的人口统计学属性向量为输入,独立于 1.1 节中从医学图像提取人

口统计学特征 Z_A 的编码分支。对于扩散时间步 t , 首先生成 128 维正弦位置编码, 随后依次经过线性映射、SiLU 激活函数和第二个线性映射, 得到 256 维时间条件向量 C_t 。两类条件向量逐元素相加, 得到 256 维全局条件信号 $C = C_{SA} + C_t$ 。式中, C 为全局条件向量, C_{SA} 为人口统计学属性条件向量, C_t 为扩散时间步条件向量。 C 用于生成自适应层归一化中的尺度因子和偏移因子, 从而调制去噪网络的中间特征, 疾病分类器则以重构后的病理特征为输入。

1.2.2 潜空间加噪与条件噪声预测

本文借鉴去噪扩散概率模型中的随机时间步加噪与噪声预测机制(Ho 等, 2020), 在潜在特征空间中对初步解耦后的病理特征 Z_T 施加高斯扰动。训练时, 从预设的扩散时间步集合中均匀采样时间步 t , 并采样与 Z_T 维度相同的高斯噪声 ϵ , 得到加噪病理特征:

$$Z_t = \sqrt{\bar{\alpha}_t} Z_T + \sqrt{1 - \bar{\alpha}_t} \epsilon \quad (4)$$

式中, Z_t 为时间步 t 对应的加噪病理特征, $\bar{\alpha}_t$ 为时间步 t 对应的累计信号保留系数, $\bar{\alpha}_t$ 决定病理特征与高斯噪声的混合比例, 其取值随扩散时间步变化, 从而形成不同强度的特征扰动。噪声 ϵ 服从零均值、单位协方差的多元正态分布。

为预测加噪过程中引入的噪声, 本文构建由全局条件信号 C 调制的条件去噪网络 ϵ_θ 。对于去噪器中的中间特征 H , 采用自适应层归一化将条件信号转换为尺度因子和偏移因子:

$$\text{AdaLN}(H, C) = \gamma(C) \odot \text{LN}(H) + \beta(C) \quad (5)$$

式中, AdaLN 表示自适应层归一化函数, $\gamma(C)$ 和 $\beta(C)$ 分别表示由全局条件信号生成的尺度因子和偏移因子, $\odot$ 表示逐元素相乘。 $\text{LN}(\cdot)$ 表示层归一化。

通过上述方式, 不同人口统计学属性和扩散时间步对应的条件信息能够作用于去噪器的中间特征, 并参与噪声预测。条件去噪网络由 4 个残差多层感知机块串联构成。每个残差块首先通过 AdaLN 完成归一化与条件调制, 随后依次经过维度为 $1024 \rightarrow 4096$ 的线性映射、GELU 激活函数和维度为 $4096 \rightarrow 1024$ 的线性映射, 并采用残差连接保持输入与输出维度一致, Dropout 率设置为 0.1。去噪器最终输出与 Z_t 维度相同的预测噪声 $\epsilon_\theta(Z_t, C)$, 用于后续病理特征重构。

1.2.3 病理特征重构

条件去噪网络根据加噪病理特征 Z_t 和全局条件信号 C , 预测时间步 t 下所注入的高斯噪声。根据上述前向加噪关系, 可以利用预测噪声直接重构病理特征:

$$Z_p = \frac{Z_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(Z_t, C)}{\sqrt{\bar{\alpha}_t}} \quad (6)$$

式中, $\epsilon_\theta(Z_t, C)$ 表示条件去噪网络预测的噪声, Z_p 表示重构后的病理特征。该模块以病理特征 Z_T 加噪得到的 Z_t 为输入, 计算起点仍保留部分原始病理信息, 因此可根据当前时间步的噪声预测结果恢复 Z_p , 形成一次加噪、一次噪声预测和一次特征重构的计算路径。

训练时, 模型在每次前向传播中分别采样扩散时间步和高斯噪声, 并完成相应的条件扩散重构。推理时采用固定时间步 $t=25$, 同时为每次前向传播独立采样高斯噪声, 使扰动强度保持一致并保留随机扰动。整个推理过程仅调用一次条件去噪网络, 其计算次数不随扩散总步数 T 增加。

1.3 疾病预测与决策校准

经过潜空间条件扩散去偏后, 模型利用重构后的病理特征 Z_p 进行疾病预测, 并在推理阶段采用子群自适应阈值校准二值分类结果。训练阶段联合优化疾病分类损失、扩散噪声预测损失、双空间正交损失和人口统计学属性分类损失, 使特征解耦、条件重构与疾病预测在统一目标下进行。模型训练完成后, 根据验证集的预测结果为不同人口统计学子群确定校准阈值, 并在推理阶段将疾病预测概率转换为最终的二值预测结果。

1.3.1 总体优化目标

训练阶段, 将 Z_p 输入疾病分类器, 得到疾病预测概率。针对本文的疾病二分类任务, 采用二元交叉熵损失 L_c 约束预测概率与真实疾病标签之间的一致性, 以保持模型的疾病判别能力。为训练条件去噪网络预测潜空间加噪过程中注入的高斯噪声, 参照去噪扩散概率模型的噪声预测方式, 采用均方误差度量前向加噪过程中采样的高斯噪声 ϵ 与条件去噪网络预测噪声 $\epsilon_\theta(Z_t, C)$ 之间的差异, 并将其记为扩散噪声预测损失 L_e 。

为联合优化疾病分类、扩散噪声预测、病理特征解耦和人口统计学特征学习, 本文构建总体优化

目标:

$$L = L_y + \lambda_e L_e + \lambda_o L_o + \lambda_{fa} L_{fa} \quad (7)$$

式中, L 为总体损失, L_y 为疾病分类损失, λ_e 和 L_e 分别表示扩散噪声预测损失的权重系数和损失项, λ_o 和 L_o 分别表示双空间正交损失的权重系数和损失项, λ_{fa} 和 L_{fa} 分别表示人口统计学属性分类损失的权重系数和损失项。

1.3.2 子群自适应阈值校准

考虑到不同人口统计学交叉子群的预测概率分布存在差异, 单一固定阈值难以充分适应各子群的决策边界。为减小这种差异对二值分类结果的影响, 本文引入子群自适应阈值校准策略。针对每项疾病分类任务, 本文在验证集上分别计算各人口统计学交叉子群在不同判别阈值下的真阳性率和真阴性率, 并将二者差值绝对值最小时对应的阈值确定为该子群的校准阈值 θ_g , 其中, g 表示人口统计学交叉子群。

推理阶段, 根据样本所属人口统计学交叉子群选取相应的校准阈值; 当疾病预测概率不低于该阈值时, 将样本判定为阳性, 否则判定为阴性。所有校准阈值均由验证集确定, 并在测试阶段保持固定。该策略仅作用于决策阶段, 不改变模型参数及其输出的疾病预测概率。

2 实验与分析

2.1 实验数据集与偏差压力测试

本文实验基于公开胸部 X 射线数据集 CheXpert (Irvin 等, 2019) 开展。该数据集包含 65240 名患者的 224316 张胸部 X 射线图像, 并提供多种胸部疾病标签及患者人口统计学信息。本文选取心脏肥大、胸腔积液、肺炎和骨折 4 种病理作为二分类任务, 以评价模型在不同疾病类型和数据分布条件下的诊断性能与群体公平性。

为考察模型在不同人口统计学子群中的性能差异, 本文选取性别、年龄和种族作为人口统计学属性。其中, 性别划分为男性和女性, 年龄以 60 岁为界划分为低龄组 (<60 岁) 和高龄组 (≥ 60 岁), 种族划分为白人和非白人。将上述 3 种二元属性组合形成 8 个人口统计学交叉子群, 并据此计算群体公平性指标。

为在可控的数据分布不均衡条件下评价不同方

法的公平性, 本文针对每项病理任务构建偏差压力测试数据。首先分别预设备单一人口统计学属性子群的目标样本构成比例和疾病阳性率, 并据此确定相应子群的阳性样本数和阴性样本数。随后按照上述目标数量进行无放回欠采样, 仅从样本数量充足的类别中删除病例, 不重复使用同一病例。四项病理任务中各单一人口统计学属性子群的样本量及疾病阳性率见表 2。

表 2 CheXpert 不同病理任务的人口统计学子群样本量及阳性率

Table 2 Sample sizes and positive rates of demographic subgroups across CheXpert pathology tasks

属性	子群	统计量	心脏肥大	胸腔积液	肺炎	骨折
性别	男性	样本量	88,458	92,981	55,909	62,046
		阳性率	14.86	48.15	6.33	9.54
	女性	样本量	29,041	48,009	14,938	14,876
		阳性率	30.12	64.85	16.53	20.83
种族	白人	样本量	75,866	82,192	51,625	54,407
		阳性率	13.46	42.64	5.29	8.32
	非白人	样本量	41,633	58,798	19,222	22,515
		阳性率	28.05	69.50	17.03	19.94
年龄	≥ 60 岁	样本量	70,184	87,191	56,056	55,766
		阳性率	24.31	60.12	6.26	8.87
	<60 岁	样本量	47,315	53,799	14,791	21,156
		阳性率	10.21	43.66	16.88	19.24

注: 阳性率表示相应人口统计学子群中疾病阳性样本所占比例, 单位为 %。

在交叉验证划分前, 按照上述采样规则固定偏差压力测试样本, 并将其中 10% 作为独立测试集, 其余 90% 用于五折交叉验证, 每折选取 1 份作为验证集、其余 4 份作为训练集。训练集、验证集和测试集均来自同一偏差压力测试分布, 而非 CheXpert 原始自然分布。所有对比方法使用相同的数据划分和样本索引。

2.2 实验设置

本文实验在 Ubuntu 服务器环境上完成, 硬件平台为 NVIDIA GeForce RTX 4090 GPU (24 GB), 模型基于 PyTorch 实现。所有图像统一缩放至 224×224 , 不采用随机翻转、随机裁剪等额外数据增强。

© 中国图象图形学报版权所有

病理特征编码分支和人口统计学特征编码分支均采用 DenseNet-121, 并进行端到端训练。模型使用 Adam 优化器, 初始学习率设为 $2e-4$, 批次大小为 128, 最大训练轮数为 12, 训练过程采用 16 位混合精度, 随机种子设为 0。当验证集损失连续 3 轮未下降时提前终止训练, 并保存验证集损失最低时的模型参数。

LCDP 的潜在特征维数为 1024, 人口统计学属性条件向量和时间条件向量的维数均为 256, 其中扩散时间步首先采用 128 维正弦位置编码。条件去噪网络包含 4 个 AdaLN 残差多层感知机块, 每个残差块的线性映射维度为 $1024 \rightarrow 4096 \rightarrow 1024$, Dropout 率为 0.1。扩散总步数 T 设为 50, 并采用余弦噪声调度。训练阶段从 0~49 中均匀采样时间步 t , 推理阶段固定取 $t=25$ 。

列空间正交损失和行空间正交损失的权重 λ_c 和 λ_r 分别设为 80 和 500, 低秩人口统计学特征子空间的维数 k 设为 3。总体优化目标中, 双空间正交损失权重 λ_o 设为 0.10, 扩散噪声预测损失权重 λ_ϵ 按照余弦策略由 1.0 衰减至 0.05, 人口统计学属性分类损失权重 λ_{ta} 设为 0.05。

2.3 性能评价指标

为综合评价模型的疾病诊断性能与群体公平性, 本文采用准确率 (accuracy, Acc) 和受试者工作特征曲线下面积 (area under the curve, AUC) 衡量整体诊断性能, 并采用等机会差异 (equal opportunity difference, ΔEO) (Hardt 等, 2016) 和子群 AUC 差异 (intersectional subgroup AUC difference, ΔAUC) 衡量不同人口统计学子群之间的性能差异。将性别、年龄和种族 3 种二元属性组合形成 8 个人口统计学交叉子群。 ΔEO 为各交叉子群真阳性率的最大差值, ΔAUC 为各交叉子群 AUC 的最大差值。Acc 和 AUC 越高, ΔEO 和 ΔAUC 越低, 表明模型的综合性能越好。

Acc 和 ΔEO 根据子群自适应阈值校准后的二值预测结果计算, AUC 和 ΔAUC 根据连续疾病预测概率计算。各子群阈值均在相应折的验证集上确定, 并应用于固定独立测试集。各项指标以五折模型在该测试集上所得结果的均值 $\pm$ 标准差报告。

2.4 实验结果与分析

为验证 LCDP 在医学图像公平诊断任务中的有效性, 本文将其与 ERM、FCRO、LNTL、FairAdaBN、

AL、FairREAD 和 Ada-ABC 共 7 种方法进行比较, 为保证比较公平, 所有对比方法均采用与本文方法一致的子群自适应阈值校准策略。各方法在 CheXpert 偏差压力测试基准上的五折交叉验证结果如表 3 所示。

从 4 项病理任务的平均结果来看, LCDP 的 Acc、AUC、 ΔAUC 和 ΔEO 分别为 0.796、0.857、0.058 和 0.108。与 7 种对比方法在相应指标上的最佳平均结果相比, Acc 和 AUC 分别由 0.781 和 0.841 提高至 0.796 和 0.857, 分别提高 0.015 和 0.016, ΔEO 由 0.126 降至 0.108, 降低 0.018, ΔAUC 为 0.058, 与最佳对比结果持平。与未进行去偏处理的 ERM 相比, LCDP 的 Acc 和 AUC 分别提高 0.056 和 0.040, ΔAUC 和 ΔEO 分别降低 0.036 和 0.063, 表明该方法能够兼顾疾病诊断性能与群体公平性。

从具体病理任务来看, LCDP 在心脏肥大任务上的 Acc、 ΔAUC 和 ΔEO 均取得最优结果, 在胸腔积液任务上, LCDP 的 ΔAUC 取得最优结果, Acc 为 0.804, 与最优结果 0.805 接近, ΔEO 为 0.073, 与最优结果 0.071 相差 0.002。在肺炎和骨折任务上, LCDP 均取得最高的 Acc 和 AUC, 其中骨折任务的 ΔEO 由 ERM 的 0.220 降至 0.105。尽管 LCDP 在部分任务的单项公平性指标上未取得最优结果, 但其在 4 项病理任务的平均结果上表现最佳, 说明该方法在不同病理任务中具有较为稳定的公平性—诊断性能综合表现。

2.5 消融实验与参数敏感性分析

2.5.1 核心组件消融实验

为考察 LCDP 各组成模块的作用, 本文以 ERM 为基础模型, 依次引入双空间正交约束 (Orth)、潜空间扩散重构模块 (Diff) 和人口统计学属性条件 (AttrCond)。其中, Diff 包含对潜在病理特征的受控加噪、扩散时间步编码和去噪重构, AttrCond 通过 AdaLN 将人口统计学属性信息引入去噪网络。4 项病理任务的消融实验结果如表 4 所示。

加入双空间正交约束后, 模型的平均 Acc 由 0.740 提高至 0.777, ΔAUC 由 0.094 降至 0.082, 但 AUC 由 0.817 降至 0.788, ΔEO 由 0.171 增至 0.193。这说明双空间正交约束能够降低交叉子群间的 AUC 差异, 但单独使用时可能削弱疾病判别信息, 且难以稳定改善子群间的真阳性率差异。

表 3 不同方法在 CheXpert 偏差压力测试基准上的五折交叉验证结果

Table 3 Five-fold cross-validation results of different methods on the CheXpert bias stress test benchmark									
病理	方法	Accuracy		AUC		Δ AUC ($\downarrow$)		Δ EO ($\downarrow$)	
心脏肥大	ERM	0.798 $\pm$ 0.017		0.846 $\pm$ 0.011		0.068 $\pm$ 0.008		0.176 $\pm$ 0.014	
	FCRO(Deng 等,2023)	0.792 $\pm$ 0.013		0.835 $\pm$ 0.012		0.082 $\pm$ 0.010		0.133 $\pm$ 0.011	
	LNTL(Kim 等,2019)	0.720 $\pm$ 0.018		0.831 $\pm$ 0.014		0.104 $\pm$ 0.013		0.152 $\pm$ 0.016	
	FairAdaBN(Xu 等,2023)	0.786 $\pm$ 0.011		0.876 $\pm$ 0.009		0.075 $\pm$ 0.008		0.140 $\pm$ 0.012	
	AL(Madras 等,2018)	0.803 $\pm$ 0.010		0.870 $\pm$ 0.010		0.073 $\pm$ 0.007		0.143 $\pm$ 0.013	
	FairREAD(Gao 等,2026)	0.781 $\pm$ 0.014		0.877	0.008	0.080 $\pm$ 0.011		0.127 $\pm$ 0.010	
	Ada-ABC(Luo 等,2025)	0.748 $\pm$ 0.016		0.799 $\pm$ 0.015		0.083 $\pm$ 0.012		0.161 $\pm$ 0.017	
	本文方法	0.805	0.009	0.869 $\pm$ 0.011		0.046	0.006	0.119	0.009
胸腔积液	ERM	0.796 $\pm$ 0.011		0.886 $\pm$ 0.008		0.047 $\pm$ 0.007		0.091 $\pm$ 0.010	
	FCRO(Deng 等,2023)	0.793 $\pm$ 0.012		0.879 $\pm$ 0.009		0.042 $\pm$ 0.006		0.081 $\pm$ 0.009	
	LNTL(Kim 等,2019)	0.771 $\pm$ 0.015		0.862 $\pm$ 0.012		0.051 $\pm$ 0.009		0.088 $\pm$ 0.011	
	FairAdaBN(Xu 等,2023)	0.801 $\pm$ 0.010		0.895 $\pm$ 0.007		0.057 $\pm$ 0.010		0.087 $\pm$ 0.010	
	AL(Madras 等,2018)	0.796 $\pm$ 0.013		0.887 $\pm$ 0.008		0.043 $\pm$ 0.008		0.092 $\pm$ 0.012	
	FairREAD(Gao 等,2026)	0.805	0.009	0.897	0.006	0.045 $\pm$ 0.007		0.077 $\pm$ 0.008	
	Ada-ABC(Luo 等,2025)	0.787 $\pm$ 0.014		0.873 $\pm$ 0.011		0.044 $\pm$ 0.009		0.071	0.009
	本文方法	0.804 $\pm$ 0.010		0.879 $\pm$ 0.010		0.041	0.006	0.073 $\pm$ 0.008	
肺炎	ERM	0.691 $\pm$ 0.018		0.778 $\pm$ 0.014		0.096 $\pm$ 0.013		0.197 $\pm$ 0.018	
	FCRO(Deng 等,2023)	0.780 $\pm$ 0.015		0.715 $\pm$ 0.016		0.118 $\pm$ 0.015		0.193 $\pm$ 0.017	
	LNTL(Kim 等,2019)	0.652 $\pm$ 0.020		0.740 $\pm$ 0.015		0.107 $\pm$ 0.014		0.179 $\pm$ 0.016	
	FairAdaBN(Xu 等,2023)	0.629 $\pm$ 0.019		0.783 $\pm$ 0.013		0.100 $\pm$ 0.012		0.172 $\pm$ 0.015	
	AL(Madras 等,2018)	0.770 $\pm$ 0.013		0.764 $\pm$ 0.012		0.092 $\pm$ 0.011		0.218 $\pm$ 0.020	
	FairREAD(Gao 等,2026)	0.699 $\pm$ 0.017		0.812 $\pm$ 0.011		0.095 $\pm$ 0.013		0.146 $\pm$ 0.014	
	Ada-ABC(Luo 等,2025)	0.757 $\pm$ 0.014		0.822 $\pm$ 0.010		0.070	0.009	0.128	0.012
	本文方法	0.801	0.012	0.851	0.009	0.090 $\pm$ 0.011		0.136 $\pm$ 0.013	
骨折	ERM	0.674 $\pm$ 0.020		0.757 $\pm$ 0.015		0.164 $\pm$ 0.017		0.220 $\pm$ 0.019	
	FCRO(Deng 等,2023)	0.759 $\pm$ 0.016		0.717 $\pm$ 0.017		0.129 $\pm$ 0.014		0.252 $\pm$ 0.022	
	LNTL(Kim 等,2019)	0.597 $\pm$ 0.022		0.721 $\pm$ 0.016		0.118 $\pm$ 0.015		0.197 $\pm$ 0.018	
	FairAdaBN(Xu 等,2023)	0.640 $\pm$ 0.019		0.749 $\pm$ 0.014		0.141 $\pm$ 0.016		0.172 $\pm$ 0.015	
	AL(Madras 等,2018)	0.728 $\pm$ 0.015		0.756 $\pm$ 0.013		0.146 $\pm$ 0.015		0.252 $\pm$ 0.021	
	FairREAD(Gao 等,2026)	0.713 $\pm$ 0.017		0.779 $\pm$ 0.010		0.098 $\pm$ 0.012		0.157 $\pm$ 0.014	
	Ada-ABC(Luo 等,2025)	0.748 $\pm$ 0.014		0.777 $\pm$ 0.011		0.035	0.005	0.145 $\pm$ 0.013	
	本文方法	0.772	0.011	0.829	0.008	0.053 $\pm$ 0.007		0.105	0.009

注:加粗字体表示各列最优结果;↓表示数值越小性能越优。

进一步加入潜空间扩散重构后,平均 Acc 和 AUC 分别提高至 0.789 和 0.846, Δ AUC 和 Δ EO 分别降至 0.067 和 0.131,表明该模块能够恢复正交

解耦过程中受到影响的病理判别信息,并进一步缩小子群间的性能差异。在此基础上加入人口统计学属性条件后,平均 Acc 和 AUC 进一步达到 0.792

表 4 LCDP各组成模块及子群自适应阈值的消融实验结果
Table 4 Ablation results of LCDP components and subgroup-adaptive threshold calibration

病理	模型配置	Accuracy		AUC		$\Delta AUC(\downarrow)$		$\Delta EO(\downarrow)$	
心脏肥大	ERM	0.798 $\pm$ 0.017		0.846 $\pm$ 0.011		0.068 $\pm$ 0.008		0.176 $\pm$ 0.014	
	+ Orth	0.801 $\pm$ 0.015		0.829 $\pm$ 0.013		0.061 $\pm$ 0.009		0.191 $\pm$ 0.017	
	+ Orth + Diff	0.803 $\pm$ 0.012		0.861 $\pm$ 0.010		0.052 $\pm$ 0.007		0.137 $\pm$ 0.012	
	+ Orth + Diff + AttrCond	0.803 $\pm$ 0.010		0.869 0.011		0.046 0.006		0.132 $\pm$ 0.011	
	+ Orth + Diff + AttrCond + Threshold	0.805 0.009		0.869 0.011		0.046 0.006		0.119 0.009	
胸腔积液	ERM	0.796 $\pm$ 0.011		0.886 $\pm$ 0.008		0.047 $\pm$ 0.007		0.091 $\pm$ 0.010	
	+ Orth	0.799 $\pm$ 0.012		0.868 $\pm$ 0.011		0.044 $\pm$ 0.007		0.104 $\pm$ 0.012	
	+ Orth + Diff	0.802 $\pm$ 0.011		0.875 $\pm$ 0.010		0.043 $\pm$ 0.006		0.086 $\pm$ 0.010	
	+ Orth + Diff + AttrCond	0.801 $\pm$ 0.010		0.879 0.010		0.041 0.006		0.081 $\pm$ 0.009	
	+ Orth + Diff + AttrCond + Threshold	0.804 0.010		0.879 0.010		0.041 0.006		0.073 0.008	
肺炎	ERM	0.691 $\pm$ 0.018		0.778 $\pm$ 0.014		0.096 $\pm$ 0.013		0.197 $\pm$ 0.018	
	+ Orth	0.748 $\pm$ 0.017		0.737 $\pm$ 0.016		0.094 $\pm$ 0.014		0.225 $\pm$ 0.021	
	+ Orth + Diff	0.787 $\pm$ 0.014		0.834 $\pm$ 0.011		0.092 $\pm$ 0.012		0.154 $\pm$ 0.015	
	+ Orth + Diff + AttrCond	0.795 $\pm$ 0.013		0.851 0.009		0.090 0.011		0.149 $\pm$ 0.014	
	+ Orth + Diff + AttrCond + Threshold	0.801 0.012		0.851 0.009		0.090 0.011		0.136 0.013	
骨折	ERM	0.674 $\pm$ 0.020		0.757 $\pm$ 0.015		0.164 $\pm$ 0.017		0.220 $\pm$ 0.019	
	+ Orth	0.759 $\pm$ 0.016		0.717 $\pm$ 0.017		0.129 $\pm$ 0.014		0.252 $\pm$ 0.022	
	+ Orth + Diff	0.763 $\pm$ 0.014		0.814 $\pm$ 0.011		0.081 $\pm$ 0.010		0.147 $\pm$ 0.015	
	+ Orth + Diff + AttrCond	0.768 $\pm$ 0.012		0.829 0.008		0.053 0.007		0.123 $\pm$ 0.011	
	+ Orth + Diff + AttrCond + Threshold	0.772 0.011		0.829 0.008		0.053 0.007		0.105 0.009	

注:加粗字体表示各列最优结果;↓表示数值越小性能越优;“+”表示在前一配置基础上增加相应模块;Orth、Diff、AttrCond 和 Threshold 分别表示双空间正交约束、潜空间扩散重构、人口统计学属性条件和子群自适应阈值。

和 0.857, ΔAUC 和 ΔEO 分别降至 0.058 和 0.121。上述结果表明,双空间正交约束、潜空间扩散重构和人口统计学属性条件分别在特征解耦、病理信息恢复和条件引导重构中发挥互补作用。

2.5.2 子群阈值校准策略分析

为考察子群自适应阈值校准(Threshold)的作用,本文固定 LCDP 模型及其连续预测概率,仅改变测试阶段的阈值策略。各折分别采用统一阈值 0.5 和由相应验证集确定的子群自适应阈值,并应用于独立测试集。由于阈值调整不改变连续预测概率及其排序,本实验仅比较与二值决策相关的 Acc 和 ΔEO 。

由表 4 可见,采用固定阈值 0.5 时,4 项病理任务的平均 Acc 和 ΔEO 分别为 0.792 和 0.121,采用子群自适应阈值后,平均 Acc 提高至 0.796, ΔEO 降

至 0.108,分别改善 0.004 和 0.013。结果表明,子群自适应阈值能够适应不同人口统计学交叉子群的预测概率分布,在保持整体分类性能的同时缩小子群间的真阳性率差异。

2.5.3 正交权重敏感性分析

正交损失权重 λ_o 用于调节双空间正交约束的强度。本文在肺炎任务上设置不同的 λ_o 进行敏感性实验,结果如图 4 所示。 λ_o 较小时,正交约束对病理特征与人口统计学特征相关性的抑制作用有限;适当增大 λ_o 有助于降低两类特征之间的相关性。当 $\lambda_o=0.10$ 时,模型的 Acc 和 AUC 达到最高, ΔAUC 和 ΔEO 也保持在相对较低水平,表明此时诊断性能与群体公平性之间取得了较好的平衡。继续增大 λ_o 后,诊断性能有所下降,公平性指标出现波动,说明过强的正交约束可能在抑制人口统计学相关信息

的同时削弱部分疾病判别信息。综合 4 项指标的变化,本文将 λ_0 设为 0.10。

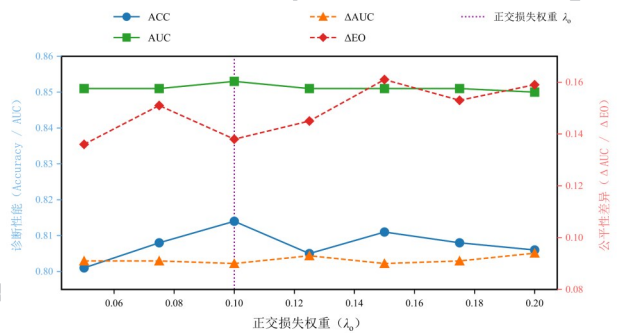


图 4 正交损失权重 λ_0 对 LCDP 性能与公平性的影响
Fig. 4 Effect of the orthogonality-loss weight λ_0 on the diagnostic performance and fairness of LCDP

2.5.4 推理时间步敏感性分析

推理时间步 t 决定推理阶段施加于潜在病理特征的噪声强度。为考察该参数对诊断性能与群体公平性的影响,本文在心脏肥大任务上设置 $t=5$ 、15、25、35 和 45 进行敏感性实验,结果如表 5 所示。

当 t 由 5 增大至 25 时,模型的 Acc 和 AUC 整体保持稳定, ΔAUC 由 0.051 降至 0.046, ΔEO 由 0.227 降至 0.119。其中, $t=25$ 时的 Acc 和 AUC 分别为 0.805 和 0.869, 均为最高值, ΔAUC 也取得最低值。当 t 进一步增大至 45 时, ΔEO 降至 0.108, 但 Acc 和 AUC 分别下降至 0.774 和 0.831, ΔAUC 升至 0.075, 说明过强的噪声扰动可能造成较大的疾病判别信息损失。综合诊断性能与群体公平性, 本文将 $t=25$ 作为默认设置。

表 5 推理时间步 t 对 LCDP 在心脏肥大任务上性能的影响
Table 5 Effect of the inference timestep t on the performance of LCDP for the cardiomegaly task

时间步 t	Acc	AUC	$\Delta AUC (\downarrow)$	$\Delta EO (\downarrow)$
5	0.802 ± 0.001	0.867 ± 0.013	0.051 ± 0.007	0.227 ± 0.017
15	0.804 ± 0.003	0.863 ± 0.031	0.047 ± 0.005	0.177 ± 0.013
25	0.805 0.009	0.869 0.011	0.046 0.006	0.119 ± 0.009
35	0.801 ± 0.002	0.866 ± 0.015	0.054 ± 0.006	0.139 ± 0.016
45	0.774 ± 0.004	0.831 ± 0.018	0.075 ± 0.002	0.108 0.023

注:加粗字体表示各列最优结果;↓表示数值越小性能越优。

2.6 特征响应可视化分析

为直观比较不同方法在疾病判别过程中的关注区域差异,本文采用梯度加权类激活映射 (gradient-weighted class activation mapping, Grad-CAM) (Selvaraju 等, 2017) 对 ERM、AL 和 LCDP 在胸部 X 射线图像上的特征响应区域进行可视化比较, 结果如图 5 所示。

由图 5 可见, ERM 的高响应区域较为分散, 在胸腔边缘和背景区域也存在较强响应, AL 减少了部分外围区域的响应, 但部分样本中仍存在较明显的背景响应。相比之下, LCDP 的高响应区域相对集中于与相应疾病相关的胸部解剖结构和潜在异常区域, 对外围背景的响应有所减弱。上述结果从可视化角度表明, LCDP 能够减少模型对背景及无关区域的关注, 同时保留与疾病分类相关的特征响应。

2.7 交叉子群性能与特征分布分析

为进一步分析 LCDP 对人口统计学交叉子群的

影响, 本文分别考察弱势交叉子群的诊断性能以及去偏前后的潜在特征分布。

2.7.1 弱势交叉子群性能分析

总体公平性指标能够反映交叉子群间的最大性能差异, 但不能直接呈现具体弱势子群的诊断表现。本文根据 ERM 在验证集上的表现确定各病理任务的弱势交叉子群, 并比较 ERM、FCRO 和 LCDP 在相应子群上的诊断性能, 结果如表 6 所示。LCDP 在 4 项病理任务的弱势交叉子群上均取得最高的 AUC, 并在心脏肥大、胸腔积液和肺炎任务中取得最高的 TPR。其中, 胸腔积液任务的子群 TPR 和 AUC 分别由 ERM 的 0.305 和 0.652 提高至 0.777 和 0.894, 肺炎任务的子群 AUC 由 0.641 提高至 0.842。在骨折任务中, LCDP 的子群 AUC 由 0.701 提高至 0.814, TPR 为 0.609, 低于 FCRO 的 0.654。上述结果表明, LCDP 在保持整体诊断性能的同时, 能够提高弱势交叉子群的疾病判别能力, 并减小不

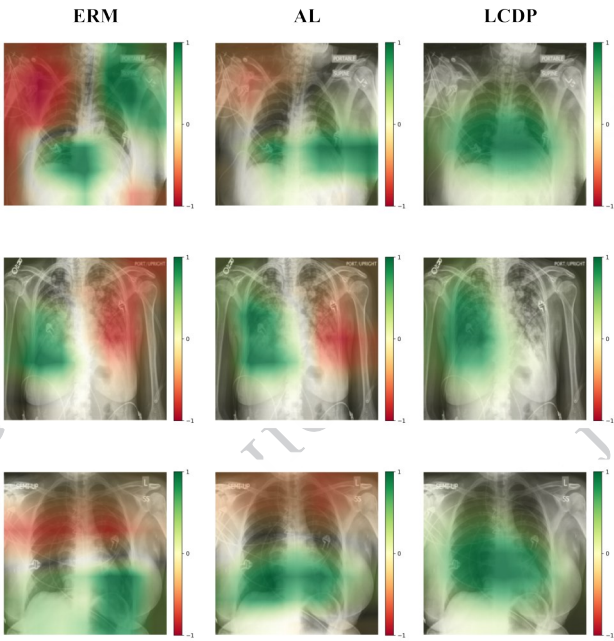


图5 ERM、AL 与 LCDP 在胸部 X 射线图像上的 Grad-CAM 特征响应对比

Fig. 5 Grad-CAM comparison of feature response maps produced by ERM, AL, and LCDP on chest X-ray images

同交叉子群间的性能差异。

2.7.2 特征空间可视化分析

为观察去偏前后潜在特征中人口统计学属性信息 与疾病判别结构的变化,本文以胸腔积液任务为

例,采用 t-SNE 将 ERM 与 LCDP 学习到的高维病理 特征投影至二维空间,并分别按照人口统计学交叉 子群和疾病标签进行着色,结果如图 6 所示。

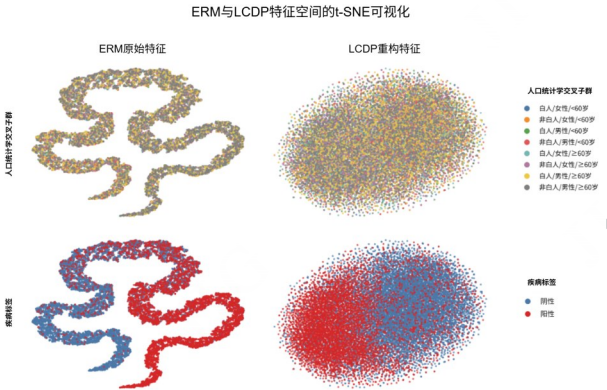


图6 ERM 与 LCDP 在胸腔积液任务中的 t-SNE 特征可 视化

Fig. 6 t-SNE visualization of representation spaces learned by ERM and LCDP in the pleural effusion task

在按照人口统计学交叉子群着色的 t-SNE 投影 中,ERM 特征呈现较明显的局部聚集趋势,而 LCDP 特征中的不同交叉子群混合程度有所提高。 按照疾病标签着色时,LCDP 特征中的疾病阳性与 阴性样本仍保持一定的分布差异。该可视化结果表 明,LCDP在减弱人口统计学交叉子群聚集趋势的同

表 6 四种病理任务中弱势交叉子群性能对比

Table 6 Performance comparison of disadvantaged intersectional subgroups across four pathology tasks													
病理(弱勢交叉子群)	方法	子群体 TPR		子群体 AUC		总体 AUC		总体 Acc		$\Delta E O (\downarrow)$		$\Delta AUC (\downarrow)$	
胸腔积液 白人-男性-低龄组	ERM	0.305 $\pm$ 0.018		0.652 $\pm$ 0.014		0.886	0.008	0.796 $\pm$ 0.011		0.091 $\pm$ 0.010		0.047 $\pm$ 0.007	
	FCRO	0.628 $\pm$ 0.015		0.844 $\pm$ 0.010		0.879 $\pm$ 0.009		0.793 $\pm$ 0.012		0.081 $\pm$ 0.009		0.042 $\pm$ 0.006	
	LCDP	0.777	0.012	0.894	0.007	0.879 $\pm$ 0.010		0.804	0.010	0.073	0.008	0.041	0.006
心脏肥大 白人-男性-高龄组	ERM	0.727 $\pm$ 0.017		0.839 $\pm$ 0.011		0.846 $\pm$ 0.011		0.798 $\pm$ 0.017		0.176 $\pm$ 0.014		0.068 $\pm$ 0.008	
	FCRO	0.727 $\pm$ 0.016		0.800 $\pm$ 0.013		0.835 $\pm$ 0.012		0.792 $\pm$ 0.013		0.133 $\pm$ 0.011		0.082 $\pm$ 0.010	
	LCDP	0.798	0.013	0.862	0.009	0.869	0.011	0.805	0.009	0.119	0.009	0.046	0.006
肺炎 白人-女性-低龄组	ERM	0.475 $\pm$ 0.022		0.641 $\pm$ 0.017		0.778 $\pm$ 0.014		0.691 $\pm$ 0.018		0.197 $\pm$ 0.018		0.096 $\pm$ 0.013	
	FCRO	0.603 $\pm$ 0.019		0.686 $\pm$ 0.015		0.715 $\pm$ 0.016		0.780 $\pm$ 0.015		0.193 $\pm$ 0.017		0.118 $\pm$ 0.015	
	LCDP	0.762	0.014	0.842	0.010	0.851	0.009	0.801	0.012	0.136	0.013	0.090	0.011
骨折 白人-男性-低龄组	ERM	0.603 $\pm$ 0.020		0.701 $\pm$ 0.016		0.757 $\pm$ 0.015		0.674 $\pm$ 0.020		0.220 $\pm$ 0.019		0.164 $\pm$ 0.017	
	FCRO	0.654	0.018	0.736 $\pm$ 0.014		0.717 $\pm$ 0.017		0.759 $\pm$ 0.016		0.252 $\pm$ 0.022		0.129 $\pm$ 0.014	
	LCDP	0.609 $\pm$ 0.017		0.814	0.011	0.829	0.008	0.772	0.011	0.105	0.009	0.053	0.007

注:加粗字体表示各列最优结果;↓ 表示数值越小性能越优。

时,保留了与疾病分类相关的特征结构。

2.8 分布外泛化与人口统计学属性特征分析

2.8.1 分布外诊断性能与公平性分析

不同医疗机构的成像条件和患者构成差异可能影响模型的外部诊断性能与公平性(Zech等,2018)。为考察 LCDP 的跨数据集泛化能力,本文将 CheXpert 训练模型直接用于 MIMIC-CXR(Johnson等,2019)测试集,不作额外训练或参数微调,4项病理任务的结果见表7和表8。

LCDP 在4项任务中均取得最低的 ΔAUC ,其平均值为0.085,平均 ΔEO 为0.166,心脏肥大和肺炎任务的 ΔEO 最优,心脏肥大和骨折任务的 Acc 分别达到最高值0.724和0.844。虽然 AUC 并非各任务最优,但交叉子群 AUC 差异均有所减小,表明 LCDP 具有较好的跨数据集公平性保持能力。

2.8.2 人口统计学属性探针分析

为检验人口统计学属性条件是否会在去噪重构过程中重新强化人口统计学信息,本文采用线性分类器构建人口统计学属性探针。该探针以 LCDP 提取的潜在特征为输入,以性别、年龄和种族组合形成的8个人口统计学交叉子群为预测类别,用于衡量基于潜在特征预测人口统计学交叉子群的能力。

在每个交叉验证折次中,分别使用 Z_T 和 Z_p 在相应折的 CheXpert 训练集上训练结构与超参数相同的线性探针,并在 CheXpert 固定独立测试集和 MIMIC-CXR 测试集上评价。

分别计算8个交叉子群的一对其余类别 AUC,并取其算术平均值作为探针 AUC。探针 AUC 越接近随机预测水平,表示基于潜在特征区分人口统计学交叉子群的能力越弱,结果如表9所示。

经过条件扩散重构后,分布内测试中的平均探针 AUC 由0.750降至0.560,降低0.190,分布外测试中的平均探针 AUC 由0.719降至0.564,降低0.155,且4项病理任务均呈下降趋势。上述结果表明,重构后的病理特征具有更低的人口统计学属性线性可分性,人口统计学属性条件未增强重构特征中的人口统计学信息。

2.9 计算开销分析

为评价 LCDP 引入的计算开销,本文在相同实验配置下,对 ERM、FCRO、FairREAD 和 LCDP 的模型参数量、平均单折训练时间、训练峰值显存及单图推理时延进行比较,结果如表10所示。

表7 心脏肥大与胸腔积液任务上的 MIMIC-CXR 分布外测试结果

Table 7 OOD test results on Cardiomegaly and Pleural Effusion tasks in MIMIC-CXR					
病理	方法	Accuracy	AUC	ΔAUC (↓)	ΔEO (↓)
心脏肥大	ERM	0.656	0.787	0.132	0.471
	FCRO	0.716	0.798	0.156	0.395
	LNTL	0.676	0.772	0.134	0.359
	FairAdaBN	0.712	0.659	0.134	0.536
	AL	0.722	0.770	0.155	0.380
	FairREAD	0.722	0.790	0.113	0.378
	Ada-ABC	0.706	0.766	0.107	0.261
	本文方法	0.724	0.759	0.082	0.139
胸腔积液	ERM	0.773	0.904	0.140	0.304
	FCRO	0.813	0.905	0.127	0.193
	LNTL	0.806	0.899	0.115	0.254
	FairAdaBN	0.803	0.846	0.085	0.268
	AL	0.806	0.910	0.142	0.229
	FairREAD	0.823	0.911	0.086	0.166
	Ada-ABC	0.817	0.878	0.087	0.086
	本文方法	0.808	0.895	0.050	0.111

注:加粗字体表示各列最优结果;↓表示数值越小性能越优。

LCDP 的模型参数量、平均单折训练时间和训练峰值显存分别为45.044 M、0.88 h 和17.3 GB,均高于其他方法,表明双分支特征编码和条件去噪网络增加了模型规模与训练开销。LCDP 的单图推理时延为3.875 ms,较表中最低的 FCRO 增加0.704 ms。总体而言,LCDP 增加了一定的模型规模和训练开销,但推理阶段的额外计算代价相对有限。

3 结 论

针对医学图像中病理信息与人口统计学属性相关信息相互耦合所导致的子群性能差异问题,本文提出潜空间条件扩散去偏框架 LCDP。与通过生成医学图像调整训练数据分布的方法不同,LCDP 将扩散过程作用于编码后的潜在病理特征,通过双空间正交约束降低病理特征与人口统计学特征的相关

表 8 骨折与肺炎任务上的 MIMIC-CXR 分布外测试结果
Table 8 OOD test results on Fracture and Pneumonia tasks in MIMIC-CXR

病理	方法	Accuracy	AUC	ΔAUC (↓)	ΔEO (↓)
骨折	ERM	0.790	0.656	0.422	0.833
	FCRO	0.730	0.661	0.438	0.857
	LNTL	0.773	0.609	0.450	0.861
	FairAdaBN	0.700	0.596	0.436	0.857
	AL	0.733	0.628	0.418	0.875
	FairREAD	0.783	0.708	0.361	0.833
	Ada-ABC	0.765	0.735	0.159	0.119
	本文方法	0.844	0.622	0.115	0.246
肺炎	ERM	0.690	0.682	0.204	0.490
	FCRO	0.741	0.673	0.205	0.529
	LNTL	0.616	0.700	0.197	0.668
	FairAdaBN	0.670	0.585	0.190	0.680
	AL	0.693	0.644	0.199	0.630
	FairREAD	0.761	0.691	0.146	0.216
	Ada-ABC	0.764	0.655	0.131	0.680
	本文方法	0.719	0.674	0.091	0.169

注:加粗字体表示各列最优结果;↓表示数值越小性能越优。

表 9 LCDP 重构前后人口统计学交叉子群探针平均 AUC 的 ID/OOD 对比

Table 9 ID/OOD comparison of mean intersectional-subgroup probe AUC before and after LCDP reconstruction

病理	ID Z_T	ID Z_p	变化	OOD Z_T	OOD Z_p	变化
心脏肥大	0.756	0.606	-0.150	0.722	0.613	-0.108
胸腔积液	0.743	0.543	-0.201	0.701	0.541	-0.161
肺炎	0.752	0.568	-0.184	0.728	0.580	-0.147
骨折	0.750	0.523	-0.227	0.724	0.521	-0.203
平均值	0.750	0.560	-0.190	0.719	0.564	-0.155

注:ID、OOD 表示分布内、分布外测试, Z_T 、 Z_p 表示重构前、重构后特征;“变化”为重构后与重构前探针 AUC 之差。

性,并利用人口统计学属性和扩散时间步引导病理特征重构。

在 CheXpert 心脏肥大、胸腔积液、肺炎和骨折

4 项分类任务上的五折交叉验证结果表明,LCDP 能够在保持较高诊断性能的同时,整体减小不同人口统计学交叉子群间的性能差异。消融实验验证了双空间正交约束、潜空间扩散重构、人口统计学属性条件和子群自适应阈值的有效性。弱势交叉子群分析进一步表明,与 ERM 相比,LCDP 在 4 项病理任务中均提高了相应弱势交叉子群的 AUC,并在心脏肥大、肺炎和骨折任务中提高了总体 AUC。在 MIMIC-CXR 分布外测试中,LCDP 在 4 项病理任务上均取得最低的 ΔAUC ,表现出较好的跨数据集公平性保持能力。人口统计学属性探针结果显示,基于重构后特征预测人口统计学交叉子群的能力有所降低。综合上述结果,LCDP 能够在保持疾病判别能力的同时,减小人口统计学交叉子群间的诊断性能差异。

本文研究仍存在一定局限。目前实验主要基于 CheXpert 和 MIMIC-CXR 两个胸部 X 射线数据集,该方法在 CT、MRI、超声和病理图像等其他医学影像模态及多中心临床数据中的适用性仍需进一步验证。此外,本文将性别、年龄和种族划分为有限类别,尚未覆盖更细粒度的人口统计学属性划分及更复杂的交叉子群关系,条件去噪网络也增加了模型参数规模和训练阶段的资源需求。后续将开展多模态、多中心及更细粒度人口统计学属性条件下的验证,并通过轻量化去噪网络和模型压缩降低计算开销,以提高 LCDP 在实际医学图像辅助诊断中的适用性。

表 10 不同方法的计算开销对比

Table 10 Computational overhead comparison of different methods

方法	总参数量/M	平均单折训练时间/h	训练峰值显存/GB	单图推理时延/ms
ERM	8.110	0.32	8.0	3.277
FCRO	16.221	0.68	9.6	3.171
FairREAD	9.312	0.62	9.1	3.620
LCDP	45.044	0.88	17.3	3.875

参考文献(Reference)

Brown A, Tomasev N, Freyberg J, Liu Y, Karthikesalingam A and Schrouff J. 2023. Detecting shortcut learning for fair medical AI using shortcut testing. *Nature Communications*, 14 (1) : 4314 [DOI: 10.1038/s41467-023-39902-7].

DeGrave A J, Janizek J D and Lee S I. 2021. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3 (7) : 610-619 [DOI: 10.1038/s42256-021-00338-7].

Deng W, Zhong Y, Dou Q and Li X. 2023. On fairness of medical image classification with multiple sensitive attributes via learning orthogonal representations//*Information Processing in Medical Imaging*. San Carlos de Bariloche, Argentina: Springer: 158-169 [DOI: 10.1007/978-3-031-34048-2_13].

Gao Y, Hao J and Zhou B. 2026. FairREAD: Re-fusing demographic attributes after disentanglement for fair medical image classification. *Medical Image Analysis*, 107: 103858 [DOI: 10.1016/j.media.2025.103858].

Gichoya J W, Banerjee I, Bhimireddy A R, Burns J L, Celi L A, Chen L C, et al. 2022. AI recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4 (6) : e406-e414 [DOI: 10.1016/S2589-7500(22)00063-2].

Glocker B, Jones C, Bernhardt M and Winzeck S. 2023. Algorithmic encoding of protected characteristics in chest X-ray disease detection models. *EBioMedicine*, 89: 104467 [DOI: 10.1016/j.ebiom.2023.104467].

Hardt M, Price E and Srebro N. 2016. Equality of opportunity in supervised learning//*Advances in Neural Information Processing Systems*: Vol. 29. Curran Associates, Inc.: 3315-3323.

Ho J, Jain A and Abbeel P. 2020. Denoising diffusion probabilistic models//*Advances in Neural Information Processing Systems*: Vol. 33. Curran Associates, Inc.: 6840-6851.

Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. 2019. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33 (1) : 590-597 [DOI: 10.1609/aaai.v33i01.3301590].

Jin R, Xu Z, Zhong Y, Yao Q, Dou Q, Zhou S K, et al. 2024. FairMedFM: fairness benchmarking for medical imaging foundation models//*Advances in Neural Information Processing Systems*: Vol. 37. Curran Associates, Inc.: 111318-111357 [DOI: 10.52202/079017-3535].

Johnson A E W, Pollard T J, Berkowitz S J, Greenbaum N R, Lungren M P, Deng C Y, et al. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1): 317 [DOI: 10.1038/s41597-019-0322-0].

Jones C, Castro D C, De Sousa Ribeiro F, Oktay O, McCradden M and Glocker B. 2024. A causal perspective on dataset bias in machine learning for medical imaging. *Nature Machine Intelligence*, 6(2) : 138-146 [DOI: 10.1038/s42256-024-00797-8].

Kim B, Kim H, Kim K, Kim S and Kim J. 2019. Learning not to learn: training deep neural networks with biased data//*Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, USA: IEEE: 9012-9020 [DOI: 10.1109/CVPR.2019.00922].

Ktena I, Wiles O, Albuquerque I, Rebuffi S A, Tanno R, Guha Roy A, et al. 2024. Generative models improve fairness of medical classifiers under distribution shifts. *Nature Medicine*, 30(4) : 1166-1173 [DOI: 10.1038/s41591-024-02838-6].

Larrazabal A J, Nieto N, Peterson V, Milone D H and Ferrante E. 2020. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23) : 12592-12594 [DOI: 10.1073/pnas.1919012117].

Lotter W. 2024. Acquisition parameters influence AI recognition of race in chest x-rays and mitigating these factors reduces underdiagnosis bias. *Nature Communications*, 15 (1) : 7465 [DOI: 10.1038/s41467-024-52003-3].

Luo L, Huang X, Wang M, Wan Z, Ma W and Chen H. 2025. Mitigating medical dataset bias by learning adaptive agreement from a biased council. *Medical Image Analysis*, 105: 103629 [DOI: 10.1016/j.media.2025.103629].

Madras D, Creager E, Pitassi T and Zemel R. 2018. Learning adversarially fair and transferable representations//*Proceedings of the 35th International Conference on Machine Learning*: Vol. 80. Stockholm, Sweden: PMLR: 3384-3393.

- Oakden-Rayner L, Dunnmon J, Carneiro G and Ré C. 2020. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging//Proceedings of the ACM Conference on Health, Inference, and Learning. ACM: 151-159 [DOI: 10.1145/3368555.3384468].
- Ricci Lara M A, Echeveste R and Ferrante E. 2022. Addressing fairness in artificial intelligence for medical imaging. *Nature Communications*, 13(1): 4581 [DOI: 10.1038/s41467-022-32186-3].
- Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D and Batra D. 2017. Grad-CAM: visual explanations from deep networks via gradient-based localization//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 618-626 [DOI: 10.1109/ICCV.2017.74].
- Seyyed-Kalantari L, Zhang H, McDermott M B A, Chen I Y and Ghassemi M. 2021. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27 (12) : 2176-2182 [DOI: 10.1038/s41591-021-01595-0].
- Shui C, Szeto J, Mehta R, Arnold D L and Arbel T. 2023. Mitigating calibration bias without fixed attribute grouping for improved fairness in medical imaging analysis//Medical Image Computing and Computer Assisted Intervention - MICCAI 2023. Vancouver, Canada: Springer: 189-198 [DOI: 10.1007/978-3-031-43898-1_19].
- Skorupko G, Osuala R, Szafranowska Z, Kushibar K, Aung N, Petersen S E, et al. 2024. Debiasing cardiac imaging with controlled latent diffusion models[EB/OL]. [2024-03-28]. [2026-08-08].
<https://arxiv.org/pdf/2403.19508.pdf> [DOI: 10.48550/arXiv.2403.19508].
- Wang M, Deng W H and Su S. 2024. Review on fairness in image recognition. *Journal of Image and Graphics*, 29(7): 1814-1833 (王玫, 邓伟洪, 苏森. 2024. 面向图像识别的公平性研究进展. *中国图象图形学报*, 29(7): 1814-1833) [DOI: 10.11834/jig.230226].
- Xing S X, Ju Z H, Liu Z J, Wang Y and Fan F Q. 2023. Multi-label classification of chest X-ray images with pre-trained vision Transformer model. *Journal of Image and Graphics*, 28(4): 1186-1197 (邢素霞, 鞠子涵, 刘子骄, 王瑜, 范福强. 2023. 视觉Transformer预训练模型的胸腔X线影像多标签分类. *中国图象图形学报*, 28(4): 1186-1197) [DOI: 10.11834/jig.220284].
- Xu Z, Li J, Yao Q, Li H, Zhao M and Zhou S K. 2024. Addressing fairness issues in deep learning-based medical image analysis: a systematic review. *npj Digital Medicine*, 7(1): 286 [DOI: 10.1038/s41746-024-01276-5].
- Xu Z, Zhao S, Quan Q, Yao Q and Zhou S K. 2023. FairAdaBN: mitigating unfairness with adaptive batch normalization and its application to dermatological disease classification//Medical Image Computing and Computer Assisted Intervention - MICCAI 2023. Vancouver, Canada: Springer: 307-317 [DOI: 10.1007/978-3-031-43895-0_29].
- Yang Y, Liu Y, Liu X, Gulhane A, Mastrodicasa D, Wu W, et al. 2025. Demographic bias of expert-level vision-language foundation models in medical imaging. *Science Advances*, 11(13): eadq0305 [DOI: 10.1126/sciadv.adq0305].
- Yang Y, Zhang H, Gichoya J W, Katabi D and Ghassemi M. 2024. The limits of fair medical imaging AI in real-world generalization. *Nature Medicine*, 30 (10) : 2838-2848 [DOI: 10.1038/s41591-024-03113-4].
- Zech J R, Badgeley M A, Liu M, Costa A B, Titano J J and Oermann E K. 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLOS Medicine*, 15 (11) : e1002683 [DOI: 10.1371/journal.pmed.1002683].
- Zhang K, Zhou R, Adhikarla E, Yan Z, Liu Y, Yu J, et al. 2024. A generalist vision - language foundation model for diverse biomedical tasks. *Nature Medicine*, 30(11): 3129-3141 [DOI: 10.1038/s41591-024-03185-2].

作者简介

董沛宇,男,硕士研究生,主要研究方向为计算机视觉、联邦学习和医学图像处理。E-mail:974736771@qq.com

翟锐,通信作者,男,副教授,主要研究方向为计算机视觉、联邦学习和医学图像分析。E-mail:zr@henu.edu.cn

王雨森,男,硕士研究生,主要研究方向为计算机视觉、图像处理和目标检测。E-mail:ys.wang0202@gmail.com

白晨希,女,副教授,主要研究方向为软件工程、人工智能和计算机视觉。E-mail:baichenxi@henu.edu.cn

吕科,男,教授,主要研究方向为计算机视觉、三维图像重建和计算机图形学。E-mail:luk@ucas.ac.cn