中图法分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-21

论文引用格式:Zhou Tao, Chen Zheng, Shen Feiyu, Wan Zhiyu, Lv Ke, Zhou Tao, Hu Fuyuan, Xu Zhenghua, Li Chen. A review of debiasing methods for medical image recognition[J/OL]. Journal of Image and Graphics, XXXX:1-21. DOI: 10.11834/jig.260378. (周涛, 陈铮, 沈飞宇, 万之瑜, 吕科, 周涛, 胡伏原, 许铮铧, 李晨. 面向医学图像识别的去偏见方法综述[J/OL]. 中国图象图形学报, XXXX:1-21. DOI: 10.11834/jig.260378.)[DOI:10.11834/jig.260378]

面向医学图像识别的去偏见方法综述

周涛[1,2],陈铮[1,2],沈飞宇[1,2],万之瑜[3],吕科[4],周涛[5],胡伏原[6],许铮铧[7],李晨[8]

1. 北方民族大学计算机科学与工程学院,宁夏银川　750021; 2. 北方民族大学图像图形智能处理国家民委重点实验室,宁夏银川　750021; 3. 上海科技大学生物医学工程学院,上海　201210; 4. 中国科学院大学工程科学学院,北京　100049; 5. 南京理工大学计算机科学与工程学院,江苏南京　210094; 6. 苏州科技大学电子与信息工程学院,江苏苏州　215009; 7. 河北工业大学生命科学与健康工程学院,天津　300401; 8. 东北大学医学与生物信息工程学院,辽宁沈阳　110819

摘　要: 医学图像识别是医学人工智能的重要研究方向,已广泛应用于病灶检测、疾病分类、器官分割和辅助诊断等任务。然而,在医学图像识别中,病灶区域占比小、疾病类别分布不均、跨机构成像差异和标注标准不一致等因素,可能使模型对小病灶、罕见疾病、特定患者群体或特定医疗中心的识别性能下降,并使模型依赖背景、设备和人群属性等与诊断目标无直接关系的信息。针对上述问题,本文围绕医学图像识别中的去偏见方法进行总结。首先,从偏见出现的位置出发,将相关问题归纳为像素级偏见、样本级偏见、特征级偏见和算法级偏见;然后,针对如何提升小病灶的权重的问题,像素级去偏归纳为像素重加权去偏,边界重加权去偏和区域重加权去偏;针对如何提高少数类样本的训练贡献,样本级去偏归纳为样本重平衡去偏和样本扩充与生成去偏;针对如何解决跨域特征分布不均的问题,特征级去偏归纳为特征对齐去偏,特征解耦去偏和知识蒸馏去偏;针对如何解决模型对非目标属性的依赖的问题,算法级去偏归纳为因果约束去偏;最后,本文总结医学图像识别的公平性问题研究在偏见来源、去偏与诊断信息的平衡、生成的医学图像的真实性、跨中心公平性评价体系等方面面临的挑战。本文为医学图像识别中偏见缓解方法的研究与应用提供参考。

关键词: 公平性问题;像素级偏见;样本级偏见;特征级偏见;算法级偏见

A review of debiasing methods for medical image recognition

Zhou Tao[1,2], Chen Zheng[1,2], Shen Feiyu[1,2], Wan Zhiyu[3], Lv Ke[4], Zhou Tao[5], Hu Fuyuan[6], Xu Zhenghua[7], Li Chen[8]

1. School of Computer Science and Engineering, North Minzu University, Yinchuan, Ningxia 750021, China; 2. Key Laboratory of Image And Graphics Intelligent Processing of State Ethnic Affairs Commission, North Minzu University, Yinchuan, Ningxia 750021, China; 3. School of Biomedical Engineering, Shanghai University of Science and Technology, Shanghai 201210, China; 4. School of Engineering Science, University of Chinese Academy of Sciences, Beijing 100049, China; 5. School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu 210094, China; 6. School of Electronics and Information Engineering, Suzhou University of Science and Technology, Suzhou, Jiangsu 215009, China; 7. School of Life Sciences and Health Engineering, Hebei University of Technology, Tianjin 300401, China; 8. School of Medicine and Biological Information Engineering, Northeastern University, Shenyang,

收稿日期:2026-06-29;**修回日期**:2026-08-28

*通信作者:陈铮,男,硕士生,研究方向为计算机辅助诊断。E-mail:15832676278@163.com

基金项目:国家自然科学基金(项目编号:No. 62561002 and No. 62576009);宁夏自然科学基金项目(项目编号:No. 2025AAC020006);宁夏科技创新领军人才培养项目(项目编号:No. 2025GKLRLX26);

Supported by: This research was funded by National Natural Science Foundation of China (No. 62561002 and No. 62576009); Ningxia Natural Science Foundation Project (No. 2025AAC020006); Ningxia Leading Talent Cultivation Project for Science and Technology (No. 2025GKLRLX26).

Liaoning 110819, China

Abstract: Medical image recognition is a crucial research direction in medical artificial intelligence. It is widely applied in clinical workflows such as computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, dermoscopy, X-ray photography, and pathological sections. These workflows involve lesion detection, disease classification, organ segmentation, and auxiliary diagnosis tasks. However, several factors degrade model recognition performance on small lesions, rare diseases, specific patient subgroups, or specific medical centers. These factors include small lesion area proportions, imbalanced disease category distributions (including long-tailed category distributions), cross-institutional imaging discrepancies, and inconsistent annotation standards. To address these issues, this paper systematically summarizes debiasing methods in medical image recognition. These methods are categorized into four primary classes: pixel-level debiasing methods, sample-level debiasing methods, feature-level debiasing methods, and algorithm-level debiasing methods. Firstly, pixel-level debiasing methods enhance the training contribution and loss weight of small lesions and low-contrast regions in dense prediction tasks, including segmentation and detection. Consequently, small lesion regions receive sufficient attention relative to background regions during network optimization. Pixel-level debiasing methods are categorized into pixel re-weighting debiasing methods, boundary re-weighting debiasing methods, and region re-weighting debiasing methods. Specifically, pixel re-weighting debiasing methods assign different training contributions to individual pixels during pixel-wise loss computation. This strategy weakens the dominant role of massive background pixels in the total loss. Boundary re-weighting debiasing methods introduce ground-truth boundaries, distance fields, or boundary discrepancy regions into the loss function. Thus, models during training not only focus on whether pixel categories are correct, but also pay close attention to spatial offsets between predicted contours and ground-truth contours. Region re-weighting debiasing methods adjust loss contributions based on the overall relationship between predicted regions and ground-truth regions. These adjustments utilize region overlap, category volume, false-detection and missed-detection costs, or region size to alleviate lesion area bias in segmentation tasks. Secondly, sample-level debiasing methods enhance the training contribution of minority-class samples under data imbalance conditions. Sample-level debiasing methods are categorized into sample rebalancing debiasing methods and sample expansion and generation debiasing methods. Specifically, sample rebalancing debiasing methods adjust sample extraction methods, screening methods, or training stage schedules without altering original image content. This improves the effective contribution of minority-class samples during model training. Sample expansion and generation debiasing methods synthesize realistic medical images that preserve semantic information for minority-class samples. This alleviates sample scarcity without violating biological logic or radiological rationality. Thirdly, feature-level debiasing methods resolve uneven cross-domain feature distributions. These distributions are caused by non-disease factors, including hospital sources, equipment types, and demographic group attributes. Feature-level debiasing methods are categorized into feature alignment debiasing methods, feature decoupling debiasing methods, and knowledge distillation debiasing methods. Specifically, feature alignment debiasing methods constrain distribution discrepancies across different domains in high-level feature space. This enables similar diseases across different hospitals, equipment, or datasets to have closer representations. Feature decoupling debiasing methods separately model disease-relevant information from domain styles, color backgrounds, or sensitive attributes during feature learning. Consequently, models rely more on stable disease features. Knowledge distillation debiasing methods utilize teacher models to supervise student models. This design allows student models to absorb more stable category relationships, structural relationships, or fairness information while learning primary tasks. Fourthly, algorithm-level debiasing methods reduce model reliance on non-object attributes during algorithm training. Algorithm-level debiasing methods are summarized as causal constraint debiasing methods. Causal constraint debiasing methods rectify and constrain feature selection from a causal perspective. This guides models to learn true pathological causal relationships and reduces reliance on biased attributes during prediction. Furthermore, this paper analyzes benchmark evaluation experiments across these four categories of debiasing methods. It systematically summarizes commonly used benchmark datasets and evaluation metrics. Finally, this paper summarizes four critical challenges currently confronting fairness research in medical image recognition: accurately identifying bias sources such as hospital origins, equipment types, and demographic group attributes remains difficult; achieving an optimal bal-

ance between debiasing and preserving clinically relevant diagnostic information is challenging; the authenticity, biological validity, and privacy compliance of generated medical images still require improvement; and cross-center fairness evaluation systems remain incomplete. This paper provides a valuable reference for the research and application of debiasing methods in medical image recognition.

Key words: fairness issue; pixel-level bias; sample-level bias; feature-level bias; algorithm-level bias

论文引用格式:[DOI:10.11834/jig.260378]

0 引　言

医学图像识别是计算机视觉与临床医学交叉的重要研究方向。近年来,卷积神经网络、Transformer及多模态基础模型被广泛用于皮肤病分类、眼底疾病筛查、肺结节检测、肿瘤分割、脑部影像分析和病理图像识别等任务。深度模型能够自动学习灰度、纹理、边界、形态和空间结构等多层次信息,为临床辅助诊断提供技术支持。

然而,模型在封闭测试集上的高性能并不等同于其在真实临床环境中稳定可靠。医学图像受到采集设备、扫描协议、医院流程、患者群体、病灶形态和标注标准等因素影响,这些因素可能与疾病标签共同出现,并被模型错误利用。已有研究表明,医学图像模型可能在不同性别、种族或患者亚群中表现不均衡(Larrazabal 等,2020; Seyyed-Kalantari 等,2021),也可能利用背景、设备痕迹或采集方式等非病理线索形成捷径学习(Degrave 等,2021; Oakden-Rayner 等,2020)。因此,医学图像识别中的公平性不仅涉及不同群体间的性能差异,还涉及模型诊断依据是否稳定可靠。

通用图像识别领域已有研究从数据、特征和模型等方面总结公平性问题及缓解方法(Wang 等,2024)。医学领域相关综述主要关注患者群体间的性能差异、公平性评价与缓解方法(Ricci Lara 等,2022; Xu 等,2024),或医院、设备和扫描协议变化引起的域适应问题(Guan 和 Liu,2022)。与上述综述不同,本文从偏见出现和方法介入的位置出发,将相关方法归纳为像素级、样本级、特征级和算法级,并结合分类、检测和分割任务分析其医学适用性与局限。

医学图像识别中的偏见具有多层次特征。在图像内部,小病灶、血管和边界区域容易被大量背景像素掩盖;在数据集层面,罕见疾病、早期病变和少数亚型的样本通常不足;在特征层面,不同医院、设备、患者群体和扫描协议可能造成域差异;在模型学习层面,模型还可能利用背景、设备或人群属性等非目标线索完成预测。随着医学基础模型的发展,特征空间中的群体偏差和跨域风险也需要进一步关注(Glocker 等,2023)。

本文所讨论的偏见,是指数据分布、图像构成或模型学习过程中出现的系统性差异;当这些差异使特定病灶、疾病类别、患者群体或医疗中心的识别性能持续下降时,进一步表现为不公平性问题。小病灶、类别长尾和跨中心差异属于可能引起偏见的因素,而像素级、样本级、特征级和算法级框架主要用于说明偏见出现和方法介入的位置。

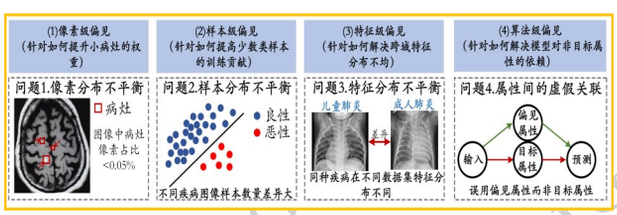

图1　四种偏见

Fig. 1　Four types of bias

如图1所示,像素级偏见表现为病灶区域占比较小,训练过程容易被背景或正常组织主导;样本级偏见表现为不同疾病或患者群体的样本数量不均;特征级偏见表现为同种疾病在不同医院、设备或群体中的特征分布不同;算法级偏见表现为模型预测依赖背景、设备或人群属性等非目标信息。不同偏见在实际任务中可能同时存在。

针对上述问题,医学图像识别去偏研究已由单一类别重平衡扩展到损失设计、样本处理、特征学习和因果约束等层面。如图2所示,像素级去偏通过像素、边界和区域重加权调整损失贡献;样本级去偏通过样本重平衡或补充尾部样本缓解长尾分布;特征级去偏通过特征对齐、特征解耦和知识蒸馏减弱域差异及敏感属性干扰;算法级去偏通过因果约束限制模型对非目标属性的依赖。

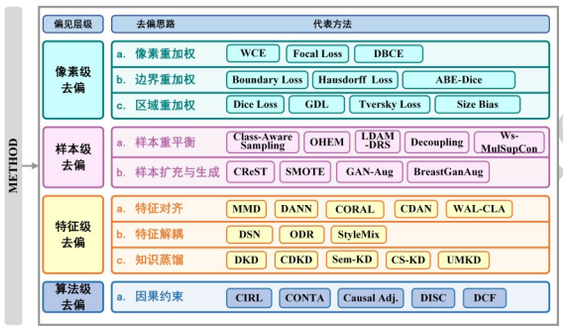

图2　去偏方法

Fig. 2　Debiasing method

本文所称医学图像识别主要包括分类、检测和分割任务。分类任务用于判断疾病类别、病理亚型或风险等级;检测任务用于定位病灶或异常区域;分割任务用于获得器官、组织或病灶的像素级、体素级范围。图像生成仅在用于少数类样本扩充、跨域转换或反事实分析时讨论,重建、配准和压缩等任务不在本文综述范围内。不同任务的主要输出、偏见层级和典型医学问题见表1。

表1　医学图像识别任务及其主要偏见层级

Table 1　Medical image recognition tasks and their main levels of bias

任务	主要输出	主要涉及的偏见层级	典型医学问题
分类	疾病类别、病理亚型或风险等级	样本级、特征级、算法级	罕见疾病长尾、患者亚群差异、设备或中心捷径
检测	病灶位置、目标框或候选区域	像素级、样本级、算法级	小目标漏检、困难病例不足、背景上下文依赖
分割	器官、组织或病灶的像素、体素区域	像素级、样本级、特征级	小病灶、弱边界、体积差异及跨设备域偏移

与自然图像相比,医学图像的外观更容易受到设备型号、扫描协议、层厚、重建方法、序列参数和染色流程等因素影响,部分灰度、信号强度和纹理变化还可能与病理状态有关。器官、病灶和血管具有较明确的解剖位置和空间关系,CT、MRI等三维图像还存在层间连续性和体素各向异性。因此,将通用去偏方法用于医学数据时,不能只消除外观差异,还应检查疾病信息和解剖结构是否得到保留。

医学数据还具有患者数量有限、罕见疾病较少、多种疾病共现以及同一患者包含多幅图像等特点。数据集应按患者划分,样本重平衡需考虑真实患病率和疾病共现,生成图像需检查病灶形态及周围组织是否合理。年龄、性别、医院和设备信息既可能被模型错误利用,也可能包含真实临床信息,因此不能简单删除,应结合外部中心和患者亚组结果评价去偏效果。

本文主要工作包括:

1)区分医学图像识别中的偏见来源与不公平性表现,明确分类、检测和分割任务的综述范围;

2)按照偏见出现和方法介入的位置,将现有方法归纳为像素级去偏、样本级去偏、特征级去偏和算法级去偏;

3)结合医学图像的数据和成像特点,分析各类方法的适用条件与局限,并总结偏见来源识别、诊断信息保留、生成图像真实性和跨中心评价等问题。

1　像素级去偏

像素级偏见是指在医学图像分割任务中,由于病灶、血管、结节等目标区域在图像或三维体数据中占比较小,模型训练过程容易被大量背景像素和正常组织像素主导,从而削弱对少数前景区域、模糊边界和困难像素的学习。其主要特点是偏见并不只来源于样本类别数量差异,而是直接存在于单幅图像内部的像素分布和损失累加过程中,即背景像素虽然诊断价值相对有限,却会因数量优势在总损失和梯度更新中占据更高比例。

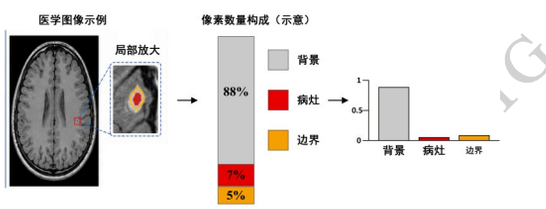

图3　像素级偏见的来源

Fig. 3　The source of pixel-level bias

如图3所示,背景像素在图像中占比较高,而病灶和边界像素较少。在等权训练下,背景像素会因数量优势形成主要优化信号,病灶和边界区域的损失贡献则容易被稀释。因此,像素级去偏主要通过重新分配不同像素、边界和区域的损失权重,使模型

更多关注病灶及关键结构。

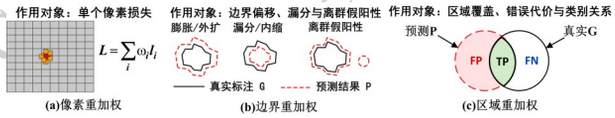

图4　像素级去偏的三类重加权策略

Fig. 4　Three types of pixel-level debiasing reweighting strategies ((a) Pixel reweighting; (b) Boundary reweighting; (c) Region reweighting)

按照重加权对象和优化尺度的不同，像素级去偏方法可分为三类，如图4所示。第一类是像素重加权，通过类别频率、空间先验或预测置信度调整单个像素的损失贡献；第二类是边界重加权，通过距离场、轮廓距离或边界差异约束预测边界；第三类是区域重加权，根据预测区域与真实区域的重叠、体积以及误检漏检代价调整优化目标。三类方法分别作用于像素误差、边界位置和整体区域质量。

1.1　基于像素重加权的像素级去偏

基于像素重加权的像素级方法是指在逐像素损失计算过程中，为不同像素分配不同训练贡献，以削弱大量背景像素对总损失的主导作用。该类方法的特点是保留逐像素监督形式，但通过类别频率、预测置信度或损失排序改变像素的损失权重，使模型更关注病灶、边界和错分区域。如图5所示，该类策略主要包括三种代表性方法：图5(a)为基于类别频率和空间先验的加权交叉熵方法(Ronneberger等，2015)，图5(b)为基于预测置信度的焦点损失方法(Lin等，2017)，图5(c)为基于结构感知权重的膨胀平衡交叉熵方法(Hosseini和Soleymani Baghshah，2026)。

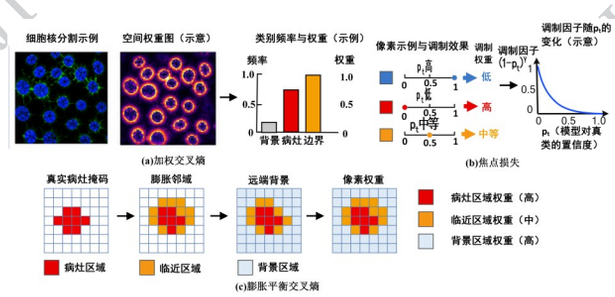

图5　基于像素重加权的像素级去偏

Fig. 5　Pixel-level debiasing based on pixel reweighting ((a) Loss based on reweighted loss and spatial reweighting; (b) Focal loss based on hard sample reweighting; (c) Dilated balanced cross entropy loss based on noise and class reweighting)

1)Ronneberger等人(2015)提出带空间权重图的加权交叉熵。如图5(a)所示，该方法同时根据类别频率和空间位置设置像素权重，使病灶、边界及相邻目标间像素获得更高训练贡献，从而减弱大面积背景对参数更新的主导作用。

2)Lin等人(2017)提出焦点损失。如图5(b)所示，该方法根据预测置信度动态调节损失，高置信度易分背景像素被降权，低置信度病灶、边界和错分像素则保留较高权重。其核心形式为

$$FL(p_t)=-\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{1}$$

式中，α_t用于类别平衡，γ控制易分像素损失衰减程度。公式中的调制因子$(1-p_t)^{\gamma}$是其缓解易分背景主导训练的关键。

3)Hosseini和Soleymani Baghshah(2026)提出膨胀平衡交叉熵。如图5(c)所示，该方法根据真实病灶及其膨胀邻域设置权重，区分病灶内部、邻近区域和远端背景的损失贡献。该设计在提高小病灶权重的同时，避免仅按类别频率统一放大少数类像素造成的过度补偿。

综上，像素重加权方法通过类别频率、预测难度和空间位置调节单个像素的损失，适合前景比例较低的医学分割任务。该类方法易于与现有网络结合，但难以直接约束轮廓位置和区域完整性，通常需要与边界或区域损失配合使用。

1.2　基于边界重加权的像素级去偏

基于边界重加权的方法是指将真实边界、距离场或边界差异区域引入损失函数，使模型在训练时不仅关注像素类别是否正确，还关注预测轮廓与真实轮廓之间的空间偏移。该类方法的特点是把边界位置、边界距离和边界不一致区域转化为损失权重，从而增强小病灶、边界和离群假阳性区域的优化约束。如图6所示，该类策略主要包括三种代表性方法：图6(a)为基于有符号距离场的边界损失方法(Kervadec等，2021)，图6(b)为基于边界距离惩罚的豪斯多夫损失方法(Karimi和Salcudean，2020)，图6(c)为基于自适应边界增强Dice损失方法(Zheng等，2025)。

1)Kervadec等人(2021)提出边界损失。如图6(a)所示，该方法由真实标注生成有符号距离场，并利用预测前景概率和距离值构造损失，使内部漏分和外部误分均推动预测轮廓向真实边界收敛。其损

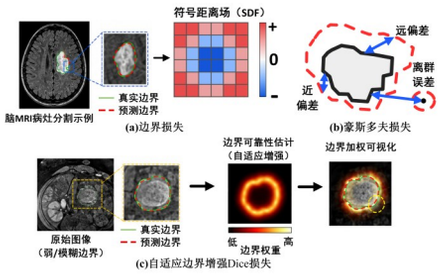

图6　基于边界重加权的像素级去偏

Fig. 6　Pixel-level debiasing based on boundary reweighting ((a) Boundary loss applied to highly imbalanced boundaries; (b) Boundary loss aimed at reducing Hausdorff distance; (c) Adaptive boundary-enhanced Dice loss)

失为

$$\mathcal{L}_B = \int_{\Omega} \phi_G(q) s_\theta(q) dq \tag{2}$$

通常目标内部 $\phi_G(q) < 0$，外部 $\phi_G(q) > 0$。因此，内部漏分时提高 $s_\theta(q)$ 可降低损失，外部误分时降低 $s_\theta(q)$ 可减少惩罚，从而使预测轮廓向真实边界收敛。

2）Karimi 等人（2020）提出豪斯多夫损失。如图6（b）所示，该方法将预测误差与其到预测边界和真实边界的距离结合，重点惩罚远离真实轮廓的离群预测和严重边界偏移。预测误差越大且距真实边界越远，损失权重越高。

3）Zheng 等人（2025）提出自适应边界增强 Dice 损失。如图6（c）所示，该方法在 Dice 优化中加入自适应概率调节和边界增强，根据预测状态提高弱边界和局部偏移区域的训练贡献，使边界权重随预测难度变化，从而改善低对比度和不规则轮廓的分割。

综上，边界重加权方法利用距离场、轮廓距离或边界概率加强边缘位置约束，适合边界模糊和形状不规则的病灶。其效果依赖边界标注质量，且局部轮廓改善不一定能够减少整体漏分，因此通常与区域重叠损失联合使用。

1.3　基于区域重加权的像素级去偏

基于区域重加权的方法是指从预测区域与真实区域的整体关系出发，通过区域重叠、类别体积、误检漏检代价或区域大小调整损失贡献。该类方法减弱大量背景像素对逐点分类损失的影响，使小病灶通过区域级约束获得更多训练关注。如图7所示，本文选取四种代表性方法进行说明：图7（a）为区域重叠损失（Milletari 等，2016），图7（b）为多类别体积损失（Sudre 等，2017），图7（c）为基于误检和漏检代价调节的非对称惩罚损失（Salehi 等，2017），图7（d）为区域大小偏差分析方法（Liu 等，2024）。此外，近期医学图像分割研究还通过多尺度损失加权，加强不同大小病灶区域的训练。

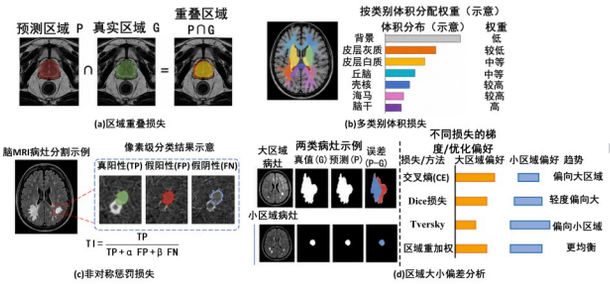

图7　基于区域重加权的像素级去偏

Fig. 7　Pixel-level debiasing based on regional reweighting ((a) Dice loss for regional reweighting; (b) Generalized Dice loss for highly imbalanced segmentation; (c) Tversky loss combining high-foreground and low-foreground penalty terms; (d) Loss with region-size bias reweighting)

1）Milletari 等人（2016）提出区域重叠损失方法。如图7（a）所示，该方法不再逐像素等权累加分类误差，而是直接衡量预测区域与真实区域的重叠程度。其核心形式为

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2} \tag{3}$$

式中，p_i 表示第 i 个像素的预测概率，g_i 表示真实标签。大量背景像素即使预测正确也不会直接提高前景重叠度，因此 Dice Loss 能够避免背景正确分类掩盖小病灶漏分。

2）Sudre 等人（2017）提出多类别体积损失方法。如图7（b）所示，该方法在 Dice 重叠度基础上引入类别体积权重，使体积较小的病灶类别在总损失中获得更高贡献。其形式为

$$\mathcal{L}_{GDL} = 1 - 2\frac{\sum_l w_l \sum_n r_{ln} p_{ln}}{\sum_l w_l \sum_n (r_{ln} + p_{ln})} \tag{4}$$

$$w_l = \frac{1}{\left(\sum_n r_{ln}\right)^2} \tag{5}$$

式中，l 为类别索引，n 为像素或体素索引，r_{ln} 和 p_{ln} 分别表示真实标签和预测概率。真实体积越小，w_l 越

大，从而放大小病灶或细小结构的区域误差。

3）Salehi 等人（2017）提出非对称惩罚损失方法。如图 7（c）所示，该方法将预测区域分为真正例、假阳性和假阴性，并用不同权重调节误检与漏检代价。其核心形式为

$$\mathcal{L}_{Tversky} = 1 - \frac{TP}{TP + \alpha FP + \beta FN} \tag{6}$$

式中，α 和 β 分别控制假阳性和假阴性的惩罚强度。对于微小病灶，可设置 $\alpha > \beta$ 以提高漏检代价，使模型更重视减少病灶漏分。

4）Liu 等人（2024）提出区域大小偏差分析方法。如图 7（d）所示，该方法指出病灶大小变化会影响不同损失产生的梯度和分割误差，部分损失可能更有利于大病灶或小病灶。因此，评价区域损失时除平均 Dice 外，还应分别比较不同大小病灶的结果。

近期医学分割研究还关注不同尺度输出的损失分配。刘建明等人（2026）根据不同层级特征的训练状态调节多尺度损失权重，使不同大小病灶参与网络优化。该方法主要用于提高医学图像分割性能，其加权思路可为小病灶训练提供参考。

综上，区域重加权通过区域重叠、类别体积和误检漏检代价调整训练贡献，适合前景较小和病灶体积差异明显的分割任务。但其对局部边界和困难像素位置的约束有限，实际应用中常与像素或边界损失联合使用。

1.4　医学适用性与局限

像素级去偏主要用于医学图像分割，也可用于小病灶检测中的位置或区域加权。早期病变、微小结节和细小血管虽然面积较小，却可能具有较高诊断价值，因此需要避免其损失被背景像素掩盖。对于 CT 和 MRI 三维数据，权重设置还应考虑体素间距、层间连续性和病灶体积，不能只依据单张切片的像素比例。

过度提高小区域或边界权重可能增加假阳性，而边界差异也可能来自成像分辨率和标注差异。因此，像素级方法应同时报告小病灶召回率、假阳性数量、区域重叠和边界指标。

2　样本级去偏

样本级偏见是指不同疾病类别、病灶亚型或患者群体的样本数量不均衡，使训练过程更容易受到头部类别主导，而尾部类别学习不足。该类偏见主要体现在不同类别进入训练批次的频率、参与参数更新的次数以及可提供的形态变化范围上。

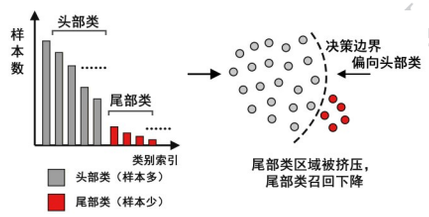

图 8　样本级偏见的来源

Fig. 8　Sources of sample-level bias

如图 8 所示，自然采样下头部类别样本反复参与模型更新，决策边界容易向尾部类别区域挤压，使罕见病灶、早期病变和少数亚型的召回率下降。因此，样本级去偏通过调整样本使用频率或补充尾部样本，使少数类别获得更充分的训练。

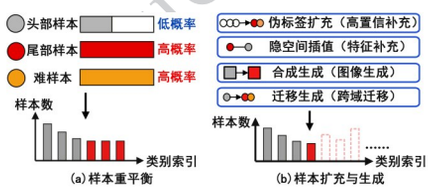

图 9　样本级去偏的两种策略

Fig. 9　Two strategies for sample-level debiasing ((a) Sample rebalancing; (b) Sample augmentation and generation)

围绕上述问题，样本级去偏方法可分为两类，如图 9 所示。第一类是样本重平衡，通过调整采样概率、困难样本选择和训练阶段，提高已有尾部样本的贡献；第二类是样本扩充与生成，通过伪标签、少数类插值和图像生成增加尾部类别的样本或特征变化。前者强调合理使用已有样本，后者强调增加尾部类别的信息来源。

2.1　基于样本重平衡的样本级去偏

基于样本重平衡的样本级去偏方法是指在不改变原始图像内容的前提下，通过调整训练样本的抽取方式、筛选方式或训练阶段安排，提高尾部类别和困难样本在训练中的有效贡献。该类方法的特点是主要作用于数据加载和训练策略层面，通过改变“哪些样本更频繁参与训练”来减弱头部类别数量优势。如图 10 所示，该类策略主要包括五种方法：图 10（a）为基于类别频率的类别均衡采样（Huang 等，2016），图 10（b）为基于损失大小的难样本采样（Shrivastava

等，2016)，图 10(c)为前期自然采样、后期强化尾部样本的延迟重采样(Cao 等，2019)，图 10(d)为先学表征、再重训分类器的解耦分类器重训(Kang 等，2020)，图 10(e)为基于疾病标签组成和共病关系的加权分层对比学习(Lin 和 Chen，2025)。

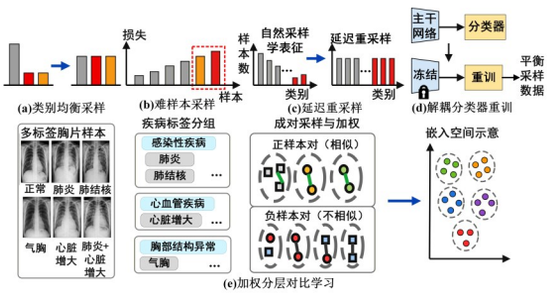

图 10　基于样本重平衡的样本级去偏

Fig. 10　Sample-level debiasing based on sample rebalancing ((a) Class rebalancing based on class re-sampling and class re-weighting; (b) Hard sample re-sampling based on loss magnitude; (c) Margin loss combining known distribution re-sampling and unknown distribution margin adjustment; (d) Decoupled re-weighting for feature representation and classifier; (e) Weighted stratification combining re-weighting and multi-label stratified re-sampling)

1)Huang 等人(2016)提出类别均衡采样。如图 10(a)所示，该方法先抽取类别，再从所选类别中抽取样本，使不同类别获得较接近的访问机会。尾部类别被选中的概率提高，头部类别的数量优势则受到限制。

2)Shrivastava 等人(2016)提出难样本采样。如图 10(b)所示，该方法根据当前训练损失优先选择高损失样本参与反向传播，使模型更多关注尾部病例、低对比度病灶和易混淆样本，减少简单头部样本对训练资源的占用。

3)Cao 等人(2019)提出延迟重采样。如图 10(c)所示，该方法前期使用自然采样学习稳定表征，后期提高尾部类别进入训练批次的比例；同时，LDAM 根据类别样本数设置分类间隔，样本越少，分类间隔越大，从采样频率和分类边界两方面缓解长尾偏见。

4)Kang 等人(2020)提出表征学习与分类器校准解耦方法。如图 10(d)所示，该方法先在自然采样数据上训练主干网络，再冻结主干并使用较均衡数据重训分类器，以减少长尾分布对最终分类边界的影响。

5)Lin 和 Chen(2025)提出加权分层多标签对比学习。如图 10(e)所示，该方法将单疾病与多疾病样本分开处理，并根据疾病出现频率和共病关系设置对比学习权重，使罕见疾病及相关样本获得更高训练贡献。

综上，样本重平衡通过调整采样频率、困难样本选择、训练阶段和分类器校准，提高尾部类别的有效训练贡献。其实现成本较低、训练过程相对稳定，但不能增加尾部病例的真实数量和形态变化。当少数类样本极少或类内差异较大时，还需结合样本扩充与生成方法。

2.2　基于样本扩充与生成的样本级去偏

基于样本扩充与生成的样本级去偏方法是指通过伪标签筛选、插值生成、图像合成或跨类迁移等方式，为尾部类别补充新的样本或特征变化。该类方法的特点是增加尾部类别可供模型学习的变化范围，使模型不再只依赖少量原始病例。如图 11 所示，该类策略主要包括四种方法：图 11(a)为基于无标签数据的伪标签扩充方法(Wei 等，2021)，图 11(b)为基于少数类近邻插值的插值示生成方法(Chawla 等，2002)，图 11(c)为基于生成器的图像/样本合成生成方法(Goodfellow 等，2014)，图 11(d)为类别可控的乳腺超声生成增强方法(Chen 等，2025)。

1)Wei 等人(2021)提出伪标签扩充。如图 11(a)所示，该方法利用当前模型为无标签数据生成伪

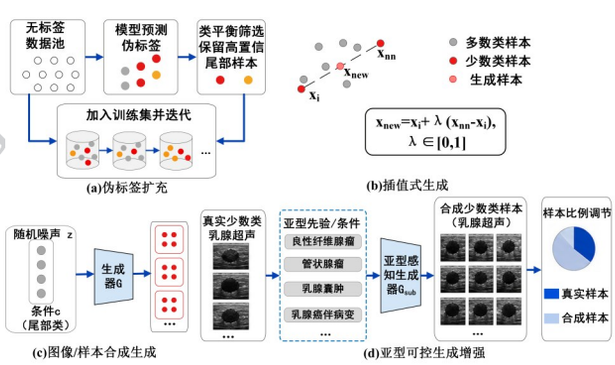

图 11　基于样本扩充与生成的样本级去偏

Fig. 11　Sample-level debiasing based on sample augmentation ((a) Conventional data augmentation based on traditional techniques; (b) Interpolation data generation based on minority class interpolation; (c) Generative augmentation based on generative adversarial networks; (d) Few-shot augmentation based on diffusion models)

标签，并通过类别均衡筛选优先保留高置信尾部样本。新增样本加入训练集后，可减少尾部类别对少量有标签病例的依赖。

2）Chawla等人（2002）提出插值式生成方法。如图11（b）所示，该方法在少数类样本及其同类近邻之间插值生成新样本，而不是简单复制原有样本。其形式为

$$\boldsymbol{x}_{\text{new}} = \boldsymbol{x}_i + \lambda(\boldsymbol{x}_{\text{nn}} - \boldsymbol{x}_i), \quad \lambda \in [0,1] \tag{7}$$

式中，$\boldsymbol{x}_i$为少数类样本，$\boldsymbol{x}_{\text{nn}}$为同类近邻，$\lambda$为随机插值系数。生成样本增加了少数类在特征空间中的密度，使尾部类分布更连续。

3）Goodfellow等人（2014）提出基于生成器的图像合成。如图11（c）所示，该方法根据随机噪声和类别条件生成新的尾部类图像，用于补充纹理、形态和背景变化。医学应用中需要保证生成病灶结构和成像规律合理，避免引入新的偏见。

4）Chen等人（2025）针对乳腺超声亚型长尾问题提出亚型导向生成增强。如图11（d）所示，该方法结合类别条件和病灶结构先验生成少数亚型图像，并调节真实样本与合成样本的使用比例，在补充尾部图像变化的同时减少过多合成样本对分类器的影响。

综上，样本扩充与生成能够增加尾部类别的病例、特征点和外观变化，弥补真实样本数量和类内变化不足。但其效果依赖伪标签准确性、语义合理性和医学结构真实性。错误标签、失真病灶或固定生成痕迹可能引入新的偏见，因此还需结合生成质量和下游任务结果进行评价。

2.3 医学适用性与局限

医学图像中的样本重平衡应以患者为基本单位。同一患者的多次检查、连续切片或不同视图不能分别进入训练集和测试集，否则容易造成数据泄漏。多标签任务还需考虑疾病共现关系，简单复制少数类样本可能改变原有标签组合。类别不均衡有时反映真实患病率，因此重采样后的模型还应在自然分布测试集上评价校准性能。对于生成扩充方法，除分类或分割结果外，还应检查合成病灶的形态、位置、内部纹理以及与周围组织的关系，防止模型转而学习生成图像中的固定痕迹。

3 特征级去偏

特征级偏见是指模型在特征学习过程中，将医院来源、设备类型、扫描协议、成像风格或患者属性等非疾病因素混入疾病特征，导致模型在跨中心、跨设备或跨人群测试时性能下降。其主要特点是疾病信息与域风格或敏感属性在特征空间中发生耦合。

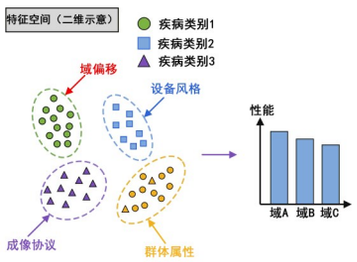

图12 特征级偏见的来源

Fig. 12 Sources of feature-level bias

如图12所示，同类疾病可能因设备或群体差异被分散到不同特征区域，不同疾病也可能因相似成像风格而过度接近。当模型依赖这些域相关线索时，可能在某一域表现较好，而迁移到其他域后性能下降。因此，特征级去偏需要调整特征空间中的表示关系，使模型更多依据稳定的病理结构和疾病语义进行判断。

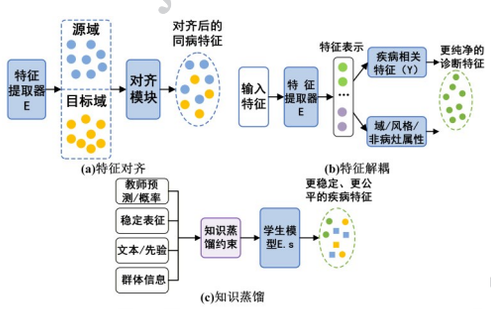

图13 特征级去偏的三种策略

Fig. 13 Three strategies for feature-level debiasing ((a) Feature alignment; (b) Feature disentanglement; (c) Knowledge distillation)

根据特征处理方式的不同，特征级去偏方法分为三类，如图13所示。第一类是特征对齐，通过统计距离、对抗学习或类别条件约束缩小域差异；第二类是特征解耦，将疾病特征与成像风格或敏感属性分开表示；第三类是知识蒸馏，利用教师预测、结构

关系或语义知识约束学生模型。

3.1 基于特征对齐的特征级去偏

基于特征对齐的特征级去偏方法是指通过约束不同域在高层特征空间中的分布差异，使不同医院、设备或数据集中的同类疾病具有更接近的表示。该类方法的特点是不改变图像内容，而是利用统计距离、域判别器、协方差或类别条件减少域相关信息的影响。如图14所示，该类策略主要包括五种方法：图14(a)为分布对齐方法(Long等，2015)，图14(b)为对抗域对齐方法(Ganin等，2016)，图14(c)为协方差对齐方法(Sun和Saenko，2016)，图14(d)为类别条件对齐方法(Long等，2018)，图14(e)为医学多标签类别级对齐方法(Liu等，2025)。此外，医学目标检测研究还从特征白化和噪声抑制等方面处理训练域与测试域之间的差异。

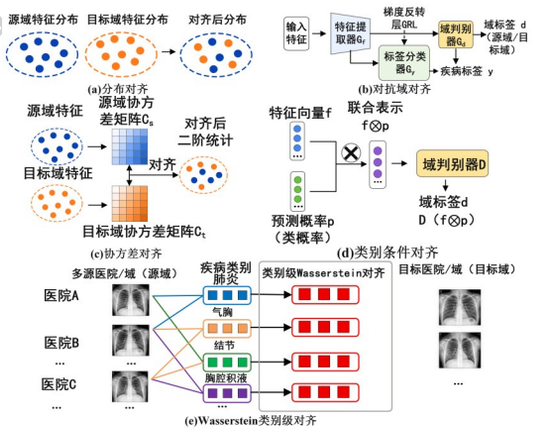

图14 基于特征对齐的特征级去偏

Fig. 14 Feature-level debiasing based on feature alignment ((a) Feature alignment; (b) Adversarial feature alignment; (c) Correlation alignment; (d) Class-level reweighted alignment; (e) Multi-label cross-domain class-level alignment)

1)Long等人(2015)提出分布对齐。如图14(a)所示，该方法通过缩小源域与目标域在高层特征空间中的分布距离，使同类疾病样本不再因医院、设备或协议差异而明显分离。

2)Ganin等人(2016)提出对抗域对齐。如图14(b)所示，该方法由特征提取器、标签分类器和域判别器组成，并通过梯度反转学习域不可分表示，使模型在保留疾病判别信息的同时削弱数据来源信息。

3)Sun等人(2016)提出协方差对齐。如图14(c)所示，该方法通过对齐源域和目标域的协方差矩阵，使两域特征的二阶统计结构更加接近。其损失函数为

$$\mathcal{L}_{CORAL}=\frac{1}{4d^2}\left\|\boldsymbol{C}_s-\boldsymbol{C}_t\right\|_{\mathrm{F}}^2 \tag{8}$$

式中，$\boldsymbol{C}_s$和$\boldsymbol{C}_t$分别为源域和目标域协方差矩阵，d为特征维度，$\|\cdot\|_{\mathrm{F}}$为Frobenius范数。该机制可降低模型对特定纹理组合和成像风格的依赖。

4)Long等人(2018)提出类别条件对齐。如图14(d)所示，该方法将深层特征与类别预测组成联合表示，并在类别条件下进行对抗对齐，使同类疾病跨域特征更接近，同时避免不同疾病类别被整体拉近。

5)Liu等人(2025)提出面向医学多标签域适应的类别级对齐方法。如图14(e)所示，该方法利用Wasserstein对抗学习缩小整体域差异，并在疾病类别层面对同类样本进行对齐，从而减少不同疾病特征被错误拉近的问题。

在医学检测任务中，史彩娟等人(2024)通过实例特征白化和噪声抑制，降低训练域特有噪声对乳腺肿瘤检测的影响，说明特征层面的域差异处理还需结合实例位置和多尺度特征。

综上，特征对齐适用于跨中心、跨设备和跨数据集任务，可减少模型对域风格、医院来源和设备类型的依赖。但当疾病信息与域因素高度耦合时，整体分布对齐可能同时削弱有效诊断特征，因此还需进一步进行特征解耦。

3.2 基于特征解耦的特征级去偏

基于特征解耦的特征级去偏方法是指在特征学习过程中，将疾病相关信息与域风格、颜色背景或敏感属性分开建模，使模型更多依赖稳定的疾病特征。该类方法的特点是不仅缩小不同域的整体差异，还调整了特征内部的组织方式。如图15所示，该类策略主要包括三种方法：图15(a)为基于共享/私有表示分离的域分离网络方法(Bousmalis等，2016)，图15(b)为利用正交约束分离疾病特征与敏感属性的正交表示学习方法(Sarhan等，2020)，图15(c)为风格混合增强的医学特征解耦方法(Cai等，2025)。

1)Bousmalis等人(2016)提出域分离网络。如图15(a)所示，该方法利用共享编码器提取跨域一致信息，利用私有编码器保存域特有风格，并通过差异约束减少两类表示的混叠。该方法将疾病结构和来源相关因素分开存放，降低医院、设备和协议差异对预测的影响。

2)Sarhan等人(2020)提出正交解耦表示学习方

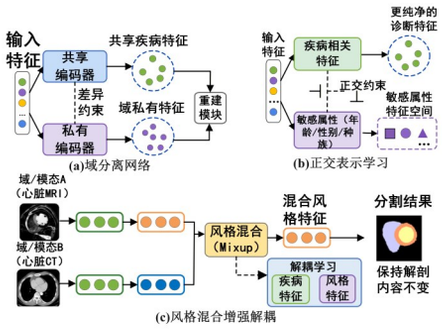

图15　基于特征解耦的特征级去偏

Fig. 15　Feature-level debiasing based on feature disentanglement ((a) Shared and private feature disentanglement; (b) Orthogonal disentanglement; (c) Style disentanglement)

法。如图15(b)所示，该方法将特征分为疾病相关表示z_y和敏感属性表示z_s，并通过正交约束减少二者信息重叠：

$$z_y^{\mathrm{T}} z_s \approx 0 \tag{9}$$

若两类表示方向接近，说明疾病特征中仍混入敏感属性信息；使二者近似正交后，分类器更难借助年龄、性别或种族等群体属性完成预测，从而降低不公平判别。

3)Cai等人(2025)提出风格混合增强解耦方法。如图15(c)所示，该方法将图像特征划分为解剖内容和成像风格，并混合不同样本的风格表示，要求模型在风格变化后保持分割结构一致，从而减少扫描协议和成像外观对结果的影响。

综上，特征解耦能够将可变成像因素与稳定解剖和病理结构分开表示，适合域风格、肤色背景和敏感属性影响明显的任务。但该类方法通常依赖明确的域信息、敏感属性或专门约束，偏见因素难以标注时效果可能受限。

3.3　基于知识蒸馏的特征级去偏

基于知识蒸馏的特征级去偏方法是指利用教师模型、跨域结构关系、文本语义或多个专家为学生模型提供监督，使学生特征不完全由当前训练集的外观分布决定。该类方法的特点是通过外部模型或语义参照约束学生特征，使学生在学习主任务的同时吸收更稳定的类别关系、结构关系或公平性信息。如图16所示，该类策略主要包括五种方法：图16(a)为利用教师软输出传递类别关系的暗知识蒸馏(Wang等，2021)，图16(b)为蒸馏跨域结构关系的跨域关系蒸馏(Qi等，2024)，图16(c)为借助文本语义校正图像特征的语义正则蒸馏(Huang等，2023)，图16(d)为保留跨风格稳定结构的跨风格分层蒸馏(Yi等，2024)，图16(e)为不确定性感知多专家知识蒸馏(Tong等，2025)。

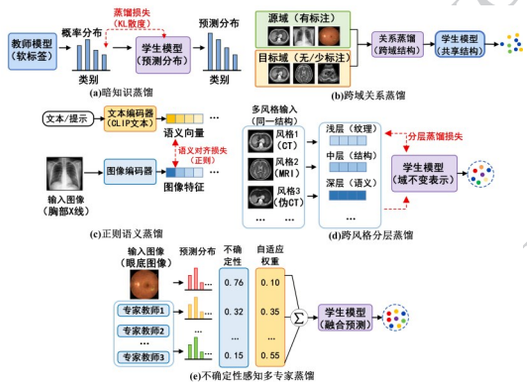

图16　基于知识蒸馏的特征级去偏

Fig. 16　Feature-level debiasing based on knowledge distillation ((a) Dark knowledge distillation; (b) Cross-domain relationship distillation; (c) Prompt-guided distillation; (d) Cross-style hierarchical distillation; (e) Uncertainty-aware multi-expert knowledge distillation)

1)Wang等人(2021)提出暗知识蒸馏。如图16(a)所示，教师模型利用软输出传递类别关系，并通过梯度过滤削弱不稳定或域特异信息，使学生模型减少对源域硬标签和外观捷径的依赖。

2)Qi等人(2024)提出跨域关系蒸馏。如图16(b)所示，该方法将多域特征组织为结构关系并传递给学生模型，使学生学习器官、病灶和区域之间较稳定的相对关系，而不是只模仿单幅图像的预测结果。

3)Huang等人(2023)提出语义正则蒸馏。如图16(c)所示，该方法利用文本语义Prompt或先验知识校正图像特征，使学生表示更加接近疾病名称、病灶语义和诊断概念，减少对颜色、背景和采集风格的依赖。

4)Yi等人(2024)提出跨风格分层蒸馏。如图16(d)所示，该方法构造不同成像风格下的特征表示，并利用教师模型约束各层特征的一致性，使学生保留跨风格稳定的解剖和病理结构。

5)Tong等人(2025)提出不确定性感知多专家知识蒸馏。如图16(e)所示，该方法利用多个专家向学生模型传递知识，并根据预测不确定性调节各专家的蒸馏权重，减少单一域或头部类别教师对学生模型的影响。

综上，知识蒸馏能够为跨域和类别不平衡任务

提供较稳定的监督，但学生模型也可能继承教师模型的偏见。因此，需要评价教师知识的可靠性，并同时报告不同疾病类别、患者群体和外部域的结果。

3.4 医学适用性与局限

医院、设备、扫描协议和患者群体并非完全等价的域因素。设备和协议主要改变成像外观，而患者属性还可能与真实疾病风险和病理表现相关，因此特征对齐或解耦不能机械删除全部域信息。

特征对齐应检查疾病判别信息是否保留，特征解耦应明确域因素和敏感属性的定义，知识蒸馏则需防止教师偏见向学生模型传递。相关方法应结合外部中心、患者亚组和不同成像条件进行综合评价。且已有医学研究多集中于跨域分类和分割，直接用于群体公平性的特征解耦和蒸馏方法仍需要更多外部数据验证。

4 算法级去偏

算法级偏见是指模型在预测过程中放大背景、设备、上下文或患者属性等非目标因素的作用，使预测依据依赖不稳定的虚假关联。其主要特点是模型形成了不稳定的预测路径：背景组织、设备痕迹、医院来源、成像协议、图像噪声或人群属性等因素可能与疾病标签在训练集中共同出现，使模型将这些非目标属性误当作诊断依据。

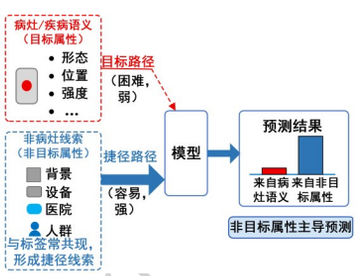

图17 算法级偏见的来源

Fig. 17 Sources of algorithmic bias

如图17所示，模型可能沿着与标签相关但缺乏诊断意义的捷径完成预测，而病灶、组织异常和疾病语义未成为主要依据。

算法级去偏的目标是使预测路径由非目标属性主导转向目标属性主导。与样本和特征层方法相比，该类方法更关注模型为什么作出预测，主要通过因果因素分离、混杂干预、概念干预和反事实一致性约束减少虚假关联。

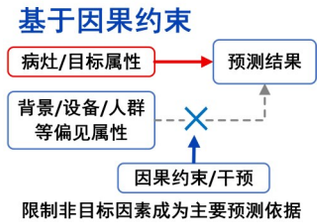

图18 算法级去偏策略

Fig. 18 Algorithm-level debiasing strategy

如图18所示，算法级去偏主要采用因果约束，限制背景、设备或人群属性等非目标因素对疾病预测的影响，使模型学习更稳定的诊断依据。

4.1 基于因果约束的算法级去偏

基于因果约束的算法级去偏方法是指利用因果表示、混杂干预、概念干预或反事实一致性约束，减少模型对非目标因素的依赖。该类方法的特点是从因果关系上修正和约束特征的选择，使模型学习真正的病理因果关系，避免模型预测过程中对偏见属性的依赖。如图19所示，该类策略主要包括五种方法：图19(a)为因果因素分离方法(Lv等，2022)，图19(b)为上下文混杂干预方法(Zhang等，2020)，图19(c)为因果混杂校正方法(Pi等，2026)，图19(d)为发现并干预不稳定概念的概念级干预方法(Wu等，2023)，图19(e)为通过反事实图像约束预测一致性的反事实一致方法(Kumar等，2023)。

1)Lv等人(2022)提出因果因素分离方法。如图19(a)所示，该方法将图像表示分为因果因素和

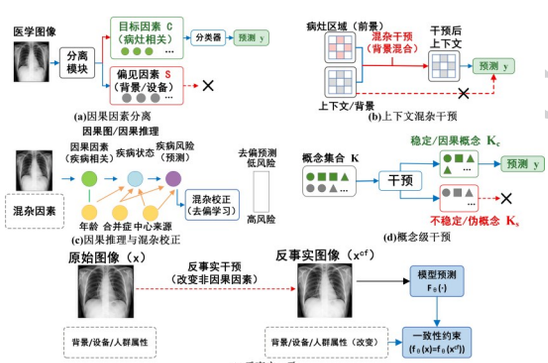

图19 基于因果约束的算法级去偏

Fig. 19 Algorithm-level debiasing based on causal constraints ((a) Causal representation; (b) Backdoor adjustment causal intervention; (c) Causal de-confounding; (d) Counterfactual intervention; (e) Invariance constraint)

非因果因素，将病灶形态、组织异常和疾病语义视为较稳定因素，将设备风格、背景纹理和域来源视为可能的非因果因素。该方法通过表示分解和独立性约束学习支持疾病预测的因果表示。

2）Zhang 等人（2020）提出上下文混杂干预方法。如图 19（b）所示，该方法将上下文视为可能影响类别判断的混杂因素，并通过干预改变其作用路径。在弱监督分割或分类中，该方法可减少模型对扫描边缘、组织背景和设备痕迹的依赖。

3）Pi 等人（2026）提出因果推理与混杂校正方法。如图 19（c）所示，该方法将影像特征、慢性合并症和目标风险纳入因果分析，并校正可能影响预测的混杂因素，以减少特定队列中的非目标关联对外部人群预测的影响。

4）Wu 等人（2023）提出概念级干预方法。如图 19（d）所示，该方法将虚假关联表示为可解释概念，先识别不同环境中不稳定的设备伪影、扫描纹理、肤色背景或非病灶组织概念，再对其进行干预，使模型更多依赖稳定疾病概念。

5）Kumar 等人（2023）提出反事实一致方法。如图 19（e）所示，该方法构造反事实图像，在保持疾病相关标记基本不变的前提下改变背景、伪影或设备因素，并约束模型输出一致。其一致性损失为

$$\mathcal{L}_{cf}=\left\| f_\theta(\boldsymbol{x})-f_\theta(\boldsymbol{x}^{\mathrm{cf}})\right\|_2^2 \tag{10}$$

式中，$\boldsymbol{x}$ 为原始图像，$\boldsymbol{x}^{\mathrm{cf}}$ 为反事实图像，$f_\theta(\cdot)$ 为模型预测。若病灶状态未变而预测明显变化，说明模型仍依赖非病灶线索；该约束可提升预测稳定性。

综上，因果约束方法从预测依据和作用路径层面减少非目标属性的影响，适合背景捷径、设备痕迹和临床混杂明显的任务。但其效果依赖因果变量、混杂因素及干预方式的合理定义，错误假设可能削弱真实诊断信息。

4.2 医学适用性与局限

因果约束方法需要事先确定哪些变量可能影响图像和诊断结果，这在医学场景中并不容易。年龄、性别、合并症和医疗中心既可能造成偏差，也可能包含真实的临床风险信息，因此不能简单从模型表示中全部删除。反事实图像还应保持解剖结构和疾病状态合理，否则预测一致并不能说明模型真正摆脱了非目标线索。此类方法除比较总体性能外，还应报告外部中心、患者亚群和校准结果，并通过消融实验说明因果变量和干预方式对结果的影响。

5 医学图像去偏方法的比较与评价

前文从像素级、样本级、特征级和算法级四个层面对医学图像识别中的去偏方法进行了总结。由于现有研究涉及不同任务、成像模态、网络结构和实验条件，不同方法之间通常不存在完全统一的评价环境。因此，本章基于公开论文中的实验结果，对同一实验设置下的代表性方法进行比较。

需要说明的是，表中数据均来源于对应论文公开实验结果。同一实验设置中的不同方法采用相同数据集、网络结构和评价指标，可以进行横向比较；不同论文之间由于数据划分和实验条件不同，不进行绝对性能排序。

5.1 常用数据集及偏见特征

医学图像去偏研究涉及结肠镜、皮肤镜、CT、胸部X线、乳腺超声、眼底和病理图像等多种模态。分割数据主要存在病灶区域较小、边界不清和器官尺度差异；长尾分类数据主要存在疾病类别数量不均、尾部类别病例不足和疾病共现；跨域研究主要涉及医院、设备、扫描体位和患者群体变化；算法级研究还需要考虑患者选择、合并症和临床混杂因素。如下表2总结了相关研究常用的数据集及偏见特征。

5.2 医学图像去偏评价指标

医学图像去偏不仅需要提高总体任务性能，还应减少不同疾病类别、患者群体和医疗中心之间的性能差异。分类任务通常采用 AUC、F1 和 Macro-F1，其中 Macro-F1 对各类别赋予相同权重，更适合评价长尾类别；分割任务通常采用 mDice、mIoU 和 HD95，分别反映区域重叠和边界误差。

公平性评价应在性别、年龄、患者群体、医院或设备等亚组上分别计算任务指标，并采用 ΔAUC、ΔF1 或 Equal Opportunity Difference 衡量组间差异。Worst-group Performance 用于反映表现最差亚组的结果，避免总体平均值掩盖少数群体性能不足。对于风险预测和多中心应用，还可采用 ECE 评价概率校准，并报告 External-center Performance 判断模型的跨中心稳定性。

5.3 不同偏见层级方法的实验性能比较

5.3.1 像素级去偏方法比较

像素级去偏主要通过损失函数调整病灶、背景

表2　医学图像去偏研究中的常用数据集及偏见特征

Table 2　Common datasets and bias characteristics in medical image debiasing studies

数据集	图像模态	主要任务	主要偏见问题	对应层级	代表性研究
Kvasir-SEG、CVC-ClinicDB	结肠镜图像	息肉分割	小病灶、弱边界、跨数据集差异	像素级、特征级	Hosseini和Soleymani Baghshah(2026)
ISIC 2018	皮肤镜图像	皮肤病变分割	病灶大小差异、边界不清	像素级	Zheng等(2025)
Synapse	腹部CT	多器官分割	器官大小差异	像素级	Hosseini和Soleymani Baghshah(2026)
ChestX-ray14	胸部X线	多标签分类	疾病长尾、疾病共现	样本级	Lin和Chen(2025)；Liu等(2025)
MIMIC-CXR	胸部X线	多标签分类	患者来源差异	样本级	Lin和Chen(2025)
CheXpert	胸部X线	多标签分类	人群和设备差异	特征级	Liu等(2025)
Breast-LT-8	乳腺超声	亚型分类	病变亚型长尾、尾部类别样本不足	样本级	Chen等(2025)
SICAPv2	病理图像	前列腺疾病分级	病理等级分布不均、跨域差异	特征级	Tong等(2025)
APTOS	眼底图像	疾病分级	疾病严重程度不均	特征级	Tong等(2025)
MACE筛查队列	胸部X线	心血管风险预测	人群选择、合并症及外部人群差异	算法级	Pi等(2026)

表3　医学图像去偏研究中的主要评价指标

Table 3　Main evaluation metrics for medical image debiasing studies

指标类别	主要指标	评价目的
分类性能	AUC、F1	评价疾病区分和综合分类能力
类别均衡性能	Macro-F1、mAcc	评价少数类别表现
分割性能	mDice、mIoU、HD95	评价区域重叠和边界误差
亚组差异	ΔAUC、ΔF1、EOD	评价群体差异
最差组性能	Worst-group Performance	评价最弱亚组表现
校准与泛化	ECE、External-center Performance	评价概率可靠性和跨中心稳定性

和边界区域的训练贡献。不同损失通常可在相同网络和数据集下直接替换，因此具有较好的组内可比性。表4汇总了Focal、Dice、Tversky和DBCE在医学图像分割任务中的结果。

表4显示，不同损失在各数据集上的相对表现并不完全一致。DBCE在多数实验中取得较高的mDice和mIoU，表明结合病灶及其邻域进行加权能够改善区域分割；Focal、Dice和Tversky在不同数据集上的结果存在变化，说明像素级方法的效果还受到病灶尺度、边界特征和数据分布影响。

5.3.2　样本级去偏方法比较

样本级方法主要针对疾病类别长尾、尾部样本不足和疾病共现问题。样本重平衡和样本生成增强的作用机制不同，因此分别采用两组统一实验进行比较。

1）　样本重平衡方法

样本重平衡通过调整样本或标签组合的训练贡献缓解类别长尾。表5汇总了多标签对比学习及加权分层策略在CXR-14和MIMIC数据集上的结果。

表5显示，加权分层多标签对比学习在两个数据集上取得较高的mAUC，并改善部分宏平均指标。该结果说明，在医学多标签任务中同时考虑疾病频率和共病关系，有助于提高尾部标签的训练贡献。但不同精确率和召回率指标的变化并不完全一致，评价时仍需结合具体临床需求。

表4　同一实验设置下像素级去偏方法的分割性能比较

Table 4　Comparison of Segmentation Performance of Pixel-level Debiasing Methods under the Same Experimental Setup

数据集	方法	mDice/%	mIoU/%
CVC-ClinicDB	Focal	85.26(±0.060)	77.96(±0.055)
	Dice	86.04(±0.061)	78.64(±0.048)
	Tversky	83.96(±0.053)	75.65(±0.046)
	DBCE	**87.38**(±0.092)	**80.29**(±0.103)
ISIC 2018	Focal	88.37(±0.044)	80.85(±0.041)
	Dice	**88.90**(±0.045)	**81.71**(±0.031)
	Tversky	87.39(±0.033)	79.41(±0.047)
	DBCE	88.85(±0.045)	81.62(±0.039)
Synapse	Focal	80.46(±0.018)	71.22(±0.022)
	Dice	80.23(±0.013)	70.91(±0.018)
	Tversky	79.59(±0.015)	70.07(±0.010)
	DBCE	**81.68**(±0.018)	**72.81**(±0.020)

注：mDice表示平均Dice系数，mIoU表示平均交并比，二者均越高越好。表中结果应直接取自同一原论文的统一实验，不同数据集之间不比较绝对数值。

2）　样本生成增强方法

样本扩充与生成通过合成尾部类别图像增加其数量和变化范围。表6比较了乳腺超声长尾分类中的未生成增强基线、MGDM、Skin-SDM和BreastGen-Aug。

表6显示，生成增强能够提高尾部类别的Few结果，但不同方法在整体分类性能、尾部类别性能和FID之间存在差异。BreastGenAug在该实验中取得较高的F1、Rec、All和Few，并具有较低FID。需要注意的是，FID只能反映特征分布差异，不能完全证明生成病灶在解剖结构和临床语义上合理，因此还应结合医生评价和外部数据验证。

5.3.3　特征级去偏方法比较

特征级去偏包括特征对齐、特征解耦和知识蒸馏。由于三类方法采用的数据集和任务不同，本节分别整理特征对齐和知识蒸馏的统一实验。特征解耦若缺少与其他正文方法在相同条件下的实验结果，可在5.4节中进行定性比较。

1）　特征对齐方法

特征对齐主要用于减少扫描体位、患者群体或设备差异造成的特征偏移。表7汇总了不同域适应

表5　多标签长尾医学分类中样本重平衡方法的性能比较

Table 5　Performance comparison of sample rebalancing methods for long-tailed multi-label medical image classification

数据集	方法	mAUC/%	mi-F1/%	ma-F1/%	ma-R/%
CXR-14	多标签监督对比学习(MulSupCon)	80.41	14.81	8.55	5.60
CXR-14	软对比学习(SoftCon)	80.11	15.03	8.47	5.51
CXR-14	加权分层多标签监督对比学习(ws-MulSupCon)	81.99	20.44	12.95	8.73
MIMIC	多标签监督对比学习(MulSupCon)	81.04	31.18	16.89	12.91
MIMIC	软对比学习(SoftCon)	81.70	32.49	18.14	13.69
MIMIC	加权分层多标签监督对比学习(ws-MulSupCon)	82.33	34.75	20.48	15.54

表6　乳腺超声长尾分类中的样本生成增强性能比较

Table 6　Performance comparison of generative augmentation methods for long-tailed breast ultrasound classification

方法	F1/%	Rec/%	All/%	Few/%	FID
未使用生成增强(Baseline)	30.94	30.81	70.13	0.00	—
MGDM	32.78	33.41	70.00	12.50	36.28
Skin-SDM	33.30	35.40	68.59	9.38	32.46
亚型导向生成增强(BreastGenAug)	35.23	38.98	72.31	31.25	30.19

注：F1表示各类别分类结果的综合性能；Rec表示Recall；All表示全部类别的总体Accuracy；Few表示尾部类别Accuracy；FID表示真实图像与生成图像特征分布之间的差异，越低越好，其余指标均为越高越好。表中结果均取自Chen等原论文在Breast-LT-8数据集上的统一实验。

方法在ChestX-ray14和CheXpert数据集上的oAUC和cAUC结果。

表7 不同胸部X线数据集中特征对齐方法的性能比较

Table 7 Performance comparison of feature alignment methods on different chest X-ray datasets

/(1)ChestX-ray14数据集

方法	PA→AP cAUC	AP→PA cAUC	M→F cAUC	F→M cAUC
仅源域训练(Source)	0.645	0.643	0.717	0.726
域对抗特征对齐(DANN)	0.681	0.673	0.754	0.741
多核最大均值差异对齐(MK-MMD)	0.620	0.631	0.696	0.695
Wasserstein类别级对齐(WAL-CLA)	**0.706**	**0.685**	**0.761**	**0.752**

注:(2)CheXpert数据集

方法	Fron→Lat cAUC	Lat→Fron cAUC	M→F cAUC	F→M cAUC
仅源域训练(Source)	0.615	0.551	0.749	0.743
域对抗特征对齐(DANN)	0.649	0.627	0.751	0.748
多核最大均值差异对齐(MK-MMD)	0.605	0.601	0.698	0.695
Wasserstein类别级对齐(WAL-CLA)	0.630	0.628	0.769	0.766

注:cAUC表示class-wise AUC,即分别计算各疾病类别的AUC后取平均,数值越高越好。PA和AP表示ChestX-ray14中的不同胸部X线投照体位;Fron和Lat分别表示CheXpert中的正位和侧位;M和F分别表示男性和女性患者群体。箭头表示由源域向目标域迁移。Source为仅使用源域数据训练的实验基线。表中数值均引自Liu等人(2025)原论文的统一实验结果。

表7显示,类别级对齐在多数迁移任务中提高了oAUC或cAUC,但不同方法在不同迁移方向上的相对表现存在差异。例如,扫描体位变化和患者性别变化所造成的域差异并不相同,因此单一对齐策略未必在所有场景中均占优势。特征对齐方法还应结合具体域因素和疾病类别进行评价。

2) 知识蒸馏方法

知识蒸馏通过教师模型的预测或特征表示约束学生模型学习。在类别不平衡条件下,教师模型自身的类别偏向可能影响知识传递,因此蒸馏方法还需要考虑不同教师知识的可靠性。表8比较了未蒸馏学生模型、经典知识蒸馏和UMKD在前列腺癌及眼底疾病分级任务中的结果。

表8 类别不平衡疾病分级中知识蒸馏方法的性能比较

Table 8 Performance comparison of knowledge distillation methods for imbalanced disease grading

/(1)前列腺癌分级数据集

不平衡设置	方法	mAcc/%	F1/%	MAE
源端不平衡	未蒸馏学生模型(Student)	89.06	89.49	0.1463
源端不平衡	经典知识蒸馏(KD)	88.39	88.97	0.1624
源端不平衡	不确定性感知多专家知识蒸馏(UMKD)	**90.23**	**90.94**	**0.1294**
目标端不平衡	未蒸馏学生模型(Student)	88.11	90.23	0.1318
目标端不平衡	经典知识蒸馏(KD)	89.44	90.90	0.1368
目标端不平衡	不确定性感知多专家知识蒸馏(UMKD)	**90.72**	**91.72**	**0.1199**

(2) 眼底疾病分级数据集

表8显示,经典KD在部分不平衡设置下未超过未蒸馏学生模型,而UMKD在多数实验中取得较高的mAcc和F1以及较低的MAE。该结果表明,知识蒸馏的效果不仅取决于是否引入教师模型,还与教师知识的类别分布和可靠性有关,评价时应同时考

不平衡设置	方法	mAcc/%	F1/%	MAE
源端不平衡	未蒸馏学生模型(Student)	67.18	73.94	0.4192
源端不平衡	经典知识蒸馏(KD)	66.07	72.77	0.4448
源端不平衡	不确定性感知多专家知识蒸馏(UMKD)	**67.33**	74.43	**0.3589**
目标端不平衡	未蒸馏学生模型(Student)	70.56	81.04	0.2521
目标端不平衡	经典知识蒸馏(KD)	71.56	83.50	0.2578
目标端不平衡	不确定性感知多专家知识蒸馏(UMKD)	**74.38**	**84.03**	**0.2476**

注:Student为未使用知识蒸馏的ResNet18学生模型;KD为经典知识蒸馏;UMKD为不确定性感知多专家知识蒸馏。源端不平衡表示教师模型训练数据存在类别不均衡,目标端不平衡表示学生模型训练数据存在类别不均衡。表中结果均引自Tong等原论文的统一实验。

虑分类性能和等级预测误差。

5.3.4　算法级去偏方法比较

算法级去偏主要通过因果推理和混杂因素校正,减少模型对非目标属性或虚假关联的依赖。表9汇总了不同因果建模与混杂校正策略在内部队列、分布偏移急诊队列和外部PCI队列上的AUC结果。

表9　因果与混杂校正模型的性能比较

Table 9　Performance comparison of causal and confounding-adjusted models

模型	6个月MACE AUC(内部)	1年MACE AUC(内部)	Shift ED AUC	Shift PCI AUC
基线混杂模型	0.707	0.702	0.726	0.639
因果模型(含特征)	0.709	0.713	0.732	0.613
因果与混杂模型(无特征)	0.719	0.718	0.725	0.646
因果与混杂模型(含特征)	0.696	0.702	0.753	0.700

注:MACE表示主要不良心血管事件;6个月MACE和1年MACE分别表示预测6个月和1年内MACE发生风险。Shift ED为存在人群分布变化的急诊队列,Shift PCI为外部PCI队列。AUC越高表示模型区分能力越强。

表9显示,不同策略在内部数据和分布变化数据上的表现存在差异。因果模型及因果与混杂组合模型在部分内部任务中取得较高AUC,含特征的组合模型在Shift和外部队列上表现较好。总体来看,因果建模与混杂校正的作用受到特征输入和数据分布影响,仍需结合外部测试和患者亚组结果评价。

5.4　方法特点与医学适用范围

定量结果只能反映特定数据集和实验条件下的性能,实际选择方法时还需考虑偏见来源、医学任务、标注条件和临床环境。

像素级方法适合处理小病灶和弱边界;样本级方法主要针对类别长尾和尾部病例不足;特征级方法用于缓解跨医院、设备和患者群体的特征差异;算法级方法用于减少背景、患者属性和临床混杂因素对预测的影响。各类方法的医学图像应用基础、适用条件和主要局限见表10。部分方法源于通用视觉或机器学习,但已在医学图像分类、检测或分割中得到应用;另一些方法则针对具体医学问题提出。方法在医学数据上取得有效结果,并不意味着其能够直接推广至其他模态、医院或患者群体,因此仍需结合解剖结构、疾病共现、外部中心和亚组结果判断其适用范围。

6　总结与展望

医学图像识别去偏是提升医学人工智能可靠性和临床适用性的重要问题。本文围绕医学图像识别中的公平性与去偏见研究进行综述,分析了偏见在像素分布、样本组成、特征表示和模型决策过程中的主要表现,并将现有方法归纳为像素级、样本级、特征级和算法级四类。像素级方法通过像素、边界和区域重加权缓解前景区域被背景数量优势稀释的问

表10　医学图像识别去偏方法的特点与适用范围

Table 10　Characteristics and application scopes of debiasing methods for medical image recognition

偏见层级	方法类别	代表方法	医学图像应用基础	适用条件	主要局限
像素级	像素重加权	加权交叉熵、Focal Loss、DBCE	前两类已有医学应用；DBCE面向医学分割提出	前景或小病灶占比较低	权重设置可能增加假阳性
	边界重加权	Boundary Loss、Hausdorff Loss、边界增强Dice	已用于医学图像分割	病灶边界模糊或不规则	依赖边界标注质量
	区域重加权	Dice Loss、GDL、Tversky Loss	已广泛用于医学图像分割	病灶或器官区域较小	局部边界约束有限
样本级	样本重平衡	类别均衡采样、延迟重采样、加权分层学习	通用策略已有医学应用；加权分层学习面向医学多标签分类提出	类别长尾或标签组合不均衡	不增加尾部样本多样性
	样本扩充与生成	伪标签、SMOTE、GAN、条件生成	前三类已有医学应用；条件生成已用于医学长尾任务	尾部病例少且类内变化不足	生成样本真实性难保证
特征级	特征对齐	MMD、DANN、CORAL、WAL-CLA	通用对齐方法已有医学应用；WAL-CLA面向医学多标签任务提出	跨医院、设备、体位或群体	过度对齐可能损失疾病信息
	特征解耦	内容—风格解耦、敏感属性解耦	已有医学图像应用，但外部验证有限	疾病特征与成像风格耦合	解耦效果难以直接验证
	知识蒸馏	KD、关系蒸馏、UMKD	通用蒸馏已有医学应用；UMKD面向医学疾病分级提出	类别不平衡或跨域知识传递	可能继承教师模型偏见
算法级	因果约束	因果表示、混杂校正、反事实约束	已有医学图像研究，跨中心验证仍有限	背景捷径或临床混杂明显	依赖因果假设和变量定义

注：医学图像应用基础依据相关方法是否在医学图像数据上进行直接实验划分。“已有医学图像应用”仅表示存在医学任务中的研究实例，不代表已完成临床验证或适用于所有医学图像场景。

题；样本级方法通过样本重平衡和样本扩充提高尾部类别样本的有效学习贡献；特征级方法通过特征对齐、特征解耦和知识蒸馏降低域风格、设备差异和群体属性等干扰；算法级方法通过因果约束降低模型对非目标属性的依赖。

尽管相关研究已取得一定进展，但医学图像识别去偏仍存在以下亟待研究的问题：

1）偏见来源难以准确识别。医院、设备、扫描协议、标注习惯、患者属性和疾病亚型可能同时影响图像，其中部分信息缺失或难以标注。现有方法通常依赖预先定义的域标签或敏感属性，难以发现未知偏见。未来需要研究在偏见标签不完整的条件下识别模型所依赖的非目标信息，并区分数据相关性与真实疾病关联。

2）去偏与诊断信息需要保持平衡。背景、器官位置、灰度变化和病灶周围组织既可能形成捷径，也可能具有诊断价值。过度重加权、对齐或解耦可能削弱真实疾病信息。未来应结合医学知识、解剖结构和临床变量，判断哪些信息应被保留，哪些关联需要限制。

3）生成医学图像的真实性仍需提高。生成图像不仅应在视觉上接近目标类别，还应符合解剖结构、病灶形态和成像规律。常用生成质量指标难以完全反映临床合理性，因此还需结合医生评价、结构一致性、外部数据和下游任务性能进行验证。

4）跨中心公平性评价体系仍不完善。现有研究多使用单中心数据和总体Accuracy、Dice或AUC，可能掩盖不同病灶尺度、疾病类别、患者群体和医疗中心间的性能差异。未来应同时报告总体性能、类别均衡指标、最差组结果、校准误差和外部中心表现，并建立适用于分类、检测和分割任务的评价体系。

综上，医学图像识别偏见不是单一数据不均衡问题，而是由图像结构、样本分布、特征偏移和模型优化机制共同作用形成的系统性问题。未来研究应从数据、模型、评价全过程出发，将多层次去偏方法与医学先验、临床风险和真实应用场景结合起来，使模型减少对非目标属性和虚假关联的依赖，并在外部中心、复杂病例、尾部类别和不同患者群体中保持稳定可靠。

参考文献(References)

Bousmalis K, Trigeorgis G, Silberman N, Krishnan D and Erhan D. 2016. Domain separation networks//Advances in Neural Information Processing Systems. Barcelona: Curran Associates: 343-351

Cai Z T, Xin J M, You C Y, Shi P W, Dong S Y, Dvornek N C, Zheng N N and Duncan J S. 2025. Style mixup enhanced disentanglement learning for unsupervised domain adaptation in medical image segmentation. Medical Image Analysis, 101: 103440 [DOI:10.1016/j.media.2024.103440]

Cao K D, Wei C L, Gaidon A, Arechiga N and Ma T. 2019. Learning imbalanced datasets with label-distribution-aware margin loss// Advances in Neural Information Processing Systems. Vancouver: Curran Associates: 1567-1578

Chawla N V, Bowyer K W, Hall L O and Kegelmeyer W P. 2002. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16: 321-357 [DOI: 10.1613/jair.953]

Chen S J, Zhou X R, Wang Y H, Huang Y H, Chang A, Ni D and Huang R B. 2025. Subtyping breast lesions via generative augmentation based long-tailed recognition in ultrasound//Medical Image Computing and Computer Assisted Intervention—MICCAI 2025. Cham: Springer Nature Switzerland: 519-529 [DOI:10.1007/978-3-032-04984-1_50]

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: 610-619 [DOI:10.1038/s42256-021-00338-7]

Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, et al. 2016. Domain-adversarial training of neural networks. Journal of Machine Learning Research, 17(59): 1-35

Glocker B, Jones C, Roschewitz M and Winzeck S. 2023. Risk of bias in chest radiography deep learning foundation models. Radiology: Artificial Intelligence, 5(6): e230060 [DOI: 10.1148/ryai.230060]

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. 2014. Generative adversarial nets//Advances in Neural Information Processing Systems. Montreal: Curran Associates: 2672-2680

Guan H and Liu M X. 2022. Domain adaptation for medical image analysis: a survey. IEEE Transactions on Biomedical Engineering, 69(3): 1173-1185 [DOI:10.1109/TBME.2021.3117407]

Hosseini S M and Soleymani Baghshah M. 2026. Dilated balanced cross entropy loss for medical image segmentation. BMC Medical Imaging, 26: 113 [DOI:10.1186/s12880-026-02245-y]

Huang C, Li Y N, Loy C C and Tang X. 2016. Learning deep representation for imbalanced classification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE: 5375-5384 [DOI:10.1109/CVPR.2016.580]

Huang Z Y, Zhou A, Lin Z J, Cai M, Wang H H and Lee Y J. 2023. A sentence speaks a thousand images: domain generalization through distilling CLIP with language guidance//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris: IEEE: 11685-11695 [DOI:10.1109/ICCV51070.2023.01073]

Kang B Y, Xie S N, Rohrbach M, Yan Z C, Gordo A, Feng J S, et al. 2020. Decoupling representation and classifier for long-tailed recognition//International Conference on Learning Representations. Addis Ababa: ICLR

Karimi D and Salcudean S E. 2020. Reducing the Hausdorff distance in medical image segmentation with convolutional neural networks. IEEE Transactions on Medical Imaging, 39(2): 499-513 [DOI:10.1109/TMI.2019.2930068]

Kervadec H, Bouchtiba J, Desrosiers C, Granger E, Dolz J and Ben Ayed I. 2021. Boundary loss for highly unbalanced segmentation. Medical Image Analysis, 67: 101851 [DOI:10.1016/j.media.2020.101851]

Kumar A, Fathi N, Mehta R, Nichyporuk B, Falet J P R, Tsaftaris S A, et al. 2023. Debiasing counterfactuals in the presence of spurious correlations//Clinical Image-Based Procedures, Fairness of AI in Medical Imaging, and Ethical and Philosophical Issues in Medical Imaging. Cham: Springer: 276-286 [DOI:10.1007/978-3-031-45249-9_27]

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 of the United States of America, 117(23): 12592-12594 [DOI:10.1073/pnas.1919012117]

Lin T Y, Goyal P, Girshick R, He K M and Dollár P. 2017. Focal loss for dense object detection//Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE: 2999-3007 [DOI:10.1109/ICCV.2017.324]

Lin Y C and Chen Y S. 2025. Weighted stratification in multi-label contrastive learning for long-tailed medical image classification//Medical Image Computing and Computer Assisted Intervention—MICCAI 2025. Cham: Springer Nature Switzerland: 677-687 [DOI:10.1007/978-3-032-05169-1_65]

Liu B Y, Dolz J, Galdran A, Kobbi R and Ben Ayed I. 2024. Do we really need Dice? The hidden region-size biases of segmentation losses. Medical Image Analysis, 91: 103015 [DOI: 10.1016/j.media.2023.103015]

Liu J M, Cao S H and Zhang Z P. 2026. Medical image segmentation with vision Mamba and adaptive multiscale loss fusion. Journal of Image and Graphics, 31(1): 335-348 (刘建明, 曹圣浩, 张志鹏. 2026. 融合视觉Mamba与自适应多尺度损失的医学图像分割. 中国图象图形学报, 31(1): 335-348) [DOI: 10.11834/jig.250224]

Liu W J, Miao F Y and Wang X. 2025. Solving medical multi-label

domain adaptation via Wasserstein adversarial learning with class-level alignment//Medical Image Computing and Computer Assisted Intervention—MICCAI 2025. Cham: Springer Nature Switzerland: 572-582 [DOI:10.1007/978-3-032-04981-0_54]

Long M S, Cao Y, Wang J M and Jordan M I. 2015. Learning transferable features with deep adaptation networks//Proceedings of the 32nd International Conference on Machine Learning. Lille: PMLR: 97-105

Long M S, Cao Z J, Wang J M and Jordan M I. 2018. Conditional adversarial domain adaptation//Advances in Neural Information Processing Systems. Montreal: Curran Associates: 1640-1650

Lv F R, Liang J, Li S, Zang B, Liu C H, Wang Z T, et al. 2022. Causality inspired representation learning for domain generalization//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE: 8046-8056 [DOI: 10.1109/CVPR52688.2022.00788]

Milletari F, Navab N and Ahmadi S A. 2016. V-Net: fully convolutional neural networks for volumetric medical image segmentation//Proceedings of the 4th International Conference on 3D Vision. Stanford: IEEE: 565-571 [DOI:10.1109/3DV.2016.79]

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. New York: ACM: 151-159 [DOI:10.1145/3368555.3384468]

Pi J L, Farina J M, Chao C J, Ayoub C, Arsanjani R and Banerjee I. 2026. Mitigating bias in opportunistic screening for MACE with causal reasoning. IEEE Transactions on Artificial Intelligence, 7(1): 103-111 [DOI:10.1109/TAI.2025.3567961]

Qi X Q, Wu Z J, Zou W X, Ren M, Gao Y F, Sun M Y, et al. 2024. Exploring generalizable distillation for efficient medical image segmentation. IEEE Journal of Biomedical and Health Informatics, 28(7): 4170-4183 [DOI:10.1109/JBHI.2024.3385098]

Ricci Lara M A, Echeveste R and Ferrante E. 2022. Addressing fairness in artificial intelligence for medical imaging. Nature Communications, 13: 4581 [DOI:10.1038/s41467-022-32186-3]

Ronneberger O, Fischer P and Brox T. 2015. U-Net: convolutional networks for biomedical image segmentation//Navab N, Hornegger J, Wells W M, et al., eds. Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015. Cham: Springer: 234-241 [DOI:10.1007/978-3-319-24574-4_28]

Salehi S S M, Erdogmus D and Gholipour A. 2017. Tversky loss function for image segmentation using 3D fully convolutional deep networks//Machine Learning in Medical Imaging. Cham: Springer: 379-387 [DOI:10.1007/978-3-319-67389-9_44]

Sarhan M H, Navab N, Eslami A and Albarqouni S. 2020. Fairness by learning orthogonal disentangled representations//Computer Vision—ECCV 2020. Cham: Springer: 746-761 [DOI: 10.1007/978-3-030-58526-6_44]

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]

Shi C J, Zheng Y F, Ren B J, Kong F Y and Duan C Y. 2024. Single-domain generalized breast tumor detection in X-ray images. Journal of Image and Graphics, 29(3): 725-740 (史彩娟, 郑远帆, 任弼娟, 孔凡跃, 段昌钰. 2024. 单域泛化X-ray乳腺肿瘤检测. 中国图象图形学报, 29(3): 725-740) [DOI:10.11834/jig.230279]

Shrivastava A, Gupta A and Girshick R. 2016. Training region-based object detectors with online hard example mining//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE: 761-769 [DOI:10.1109/CVPR.2016.89]

Sudre C H, Li W Q, Vercauteren T, Ourselin S and Cardoso M J. 2017. Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations//Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Cham: Springer: 240-248 [DOI:10.1007/978-3-319-67558-9_28]

Sun B C and Saenko K. 2016. Deep CORAL: correlation alignment for deep domain adaptation//Computer Vision—ECCV 2016 Workshops. Cham: Springer: 443-450 [DOI:10.1007/978-3-319-49409-8_35]

Tong S, Gao S D, Liu K, Huang Z H, Xu H X, Ying H C and Wu J. 2025. Uncertainty-aware multi-expert knowledge distillation for imbalanced disease grading//Medical Image Computing and Computer Assisted Intervention—MICCAI 2025. Cham: Springer Nature Switzerland: 624-634 [DOI: 10.1007/978-3-032-05169-1_60]

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]

Wang Y F, Li H L, Chau L P and Kot A C. 2021. Embracing the dark knowledge: domain generalization using regularized knowledge distillation//Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM: 2595-2604 [DOI: 10.1145/3474085.3475434]

Wei C, Sohn K, Mellina C, Yuille A and Yang F. 2021. CReST: a class-rebalancing self-training framework for imbalanced semi-supervised learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE: 10857-10866 [DOI:10.1109/CVPR46437.2021.01071]

Wu S, Yuksekgonul M, Zhang L J and Zou J. 2023. Discover and cure: concept-aware mitigation of spurious correlation//Proceedings of the 40th International Conference on Machine Learning. Honolulu: PMLR: 37765-37786

Xu Z K, Li J, Yao Q S, Li H, Zhao M Y and Zhou S K. 2024. Addressing fairness issues in deep learning-based medical image analysis:

a systematic review. npj Digital Medicine，7：286［DOI：10.1038/s41746-024-01276-5］

Yi J J，Bi Q，Zheng H，Zhan H L，Ji W，Huang Y W，et al. 2024. Hallucinated style distillation for single domain generalization in medical image segmentation//Medical Image Computing and Computer Assisted Intervention—MICCAI 2024. Cham：Springer：438-448［DOI：10.1007/978-3-031-72117-5_41］

Zhang D，Zhang H W，Tang J H，Hua X S and Sun Q. 2020. Causal intervention for weakly-supervised semantic segmentation//Advances in Neural Information Processing Systems. Vancouver：Curran Associates：655-666

Zheng Y Y，Tian B H，Yu S C，Yang X G，Yu Q X，Zhou J，Jiang G Q，Zheng Q X，Pu J T and Wang L. 2025. Adaptive boundary-enhanced Dice loss for image segmentation. Biomedical Signal Processing and Control，106：107741［DOI：10.1016/j.bspc.2025.107741］

作者简介

沈飞宇，男，硕士生，研究方向为计算机辅助诊断。E-mail：shenfeiyu16@stu.nmu.edu.cn

万之瑜，男，研究员，主要研究方向为智能医学、机器学习与人工智能。E-mail：wanzhy@shanghaitech.edu.cn

吕科，男，教授，主要研究方向为人工智能、数字图像处理技术。E-mail：luk@ucas.ac.cn

周涛，男，教授，主要研究方向为医学图像处理与分析、AI健康、计算机视觉。E-mail：taozhou@njust.edu.cn

胡伏原，男，教授，主要研究方向为图像处理、模式识别、人工智能。E-mail：fuyuanhu@usts.edu.cn

许铮铧，男，教授，主要研究方向为智能医学影像与健康大数据分析、深度强化学习。E-mail：zhenghua.xu@hebut.edu.cn

李晨，男，副教授，主要研究方向为病理显微图像分析、多模态医学影像分析。E-mail：lichen@mbie.neu.edu.cn