计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(11): 3017−3031, 2026

面向皮肤病诊断去偏的增强型平衡增量蒸馏网络

罗义琴[1] 古天龙[1] 郝峰锐[1] 常 亮[2]
[1](可信人工智能教育部工程研究中心(暨南大学) 广州 510632)
[2](广西可信软件重点实验室(桂林电子科技大学) 广西桂林 541004)
(lyq_ah@163.com)

Enhanced Balanced Incremental Distillation Network for Debiased Dermatological Disease Diagnosis

Luo Yiqin[1], Gu Tianlong[1], Hao Fengrui[1], and Chang Liang[2]
[1] (*Engineering Research Center of Trustworthy AI* (*Jinan University*), *Ministry of Education*, *Guangzhou* 510632)
[2] (*Guangxi Key Laboratory of Trusted Software* (*Guilin University of Electronic Technology*), *Guilin*, *Guangxi* 541004)

Abstract Deep learning technology has been increasingly empowering the practice of dermatological disease diagnosis, leading to a steady improvement in its diagnostic performance. However, such data-driven deep learning models are plagued by severe bias issues. If systems with these biases are deployed in clinical applications, they are bound to cause unfair treatment towards specific groups, such as underrepresented groups. Ensuring the fairness of decision-making has become a critical and urgent problem that needs to be addressed. Existing research has indicated that enhancing fairness often comes at the cost of reduced diagnostic performance. Therefore, to achieve a better trade-off between diagnostic performance and fairness, while simultaneously alleviating the unfair treatment of underrepresented groups, we propose an enhanced balanced incremental distillation network (EBID-Net). Specifically, with the aid of balanced memory, the model leverages representative demographic groups to assist underrepresented groups in knowledge learning, while being incrementally trained by integrating distributions across different groups. Additionally, we incorporate global information into the contextual attention mechanism to capture correlated interactions between features across different spatial locations, thereby obtaining more robust feature representations during incremental learning. Experimental results show that EBID-Net outperforms other methods in fairness criteria and in the trade-off between fairness and diagnostic performance.

Key words fairness; dermatological disease; incremental learning; trade-off; attention mechanism

摘 要 深度学习技术日益赋能皮肤病诊断应用,使其诊断性能稳步提升。然而,这类数据驱动的深度学习模型存在严重的偏见问题。若将存在偏见的系统部署于临床应用,势必会对特定群体(如非代表性群体)造成不公平对待。确保决策的公平性已成为一个至关重要且亟待解决的问题。现有研究表明,提升公平性往往以降低诊断性能为代价。因此,为了在诊断性能和公平性之间取得更优权衡,同时减轻对非代表性群体的不公平对待,提出一个增强型平衡增量蒸馏网络(enhanced balanced incremental distillation network, EBID-Net)。具体而言,该模型借助平衡记忆内存整合不同群体的分布开展增量训练,促使代表性群体辅助非代表性群体学习知识。此外,将全局信息融入上下文注意力机制,捕捉不同空间位置特征间的交互关系,从而在增量学习过程中获得更稳健的特征表示。实验结果显示,EBID-Net 在公平性以及

收稿日期: 2025-11-03; **修回日期**: 2026-06-04
基金项目: 国家自然科学基金项目(U22A2099, 62336003); 中央高校基本科研业务费专项资金(88022124)
This work was supported by the National Natural Science Foundation of China (U22A2099, 62336003) and the Fundamental Research Funds for the Central Universities (88022124).
通信作者: 古天龙(gutianlong@jnu.edu.cn)

诊断性能和公平性间的权衡方面均优于其他方法。

关键词 公平;皮肤病;增量学习;权衡;注意力机制

中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202550799 **CSTR:** 32373.14.issn1000-1239.202550799

皮肤病是全球最常见的健康问题之一。然而,皮肤病的准确诊断仍颇具挑战性。该诊断过程需要整合多种视觉线索,如病变外观、大小分布、颜色、鳞屑特征和排列等[1]。此外,皮肤病在全球范围内的发病率日益攀升,加之皮肤科医生短缺,促使研究人员探索更高效的诊断方法[2]。人工智能辅助诊断在提升全科医生诊断能力方面展现出巨大潜力[3-4]。例如,Liu 等人[5]开发了一个深度学习系统,该系统能够诊断 26 种常见皮肤病,其准确率可与获得认证的皮肤科医生相媲美。在一项真实世界临床试验中,Papachristou 等人[6]证明,人工智能驱动的决策支持工具在检测黑色素瘤方面具有较高的诊断性能,表明其对初级保健医生具有重要的临床价值。此外,Huang 等人[7]还成功开发了一种用于评估实际临床环境中银屑病严重程度的模型。

尽管这类数据驱动的深度学习模型前景广阔,但高度依赖数据质量,因而极易受到数据偏见的影响[8-10]。研究人员发现,这些模型在疾病分类时往往会过度关注人口统计学特征(如肤色和性别)[11]。这种倾向可能导致模型出现偏见,具体表现为不同人口统计学亚组间的诊断性能存在显著差异[12-13]。在临床应用中部署此类存在偏见的系统,将不可避免地对某些特定人口统计学群体造成不公平对待。鉴于医疗应用中对公平性的迫切需求以及医学数据的独特性,探索有效策略以减轻医疗人工智能中的此类偏见显得至关重要。本研究聚焦于人工智能辅助医学图像诊断,特别是皮肤病诊断中的偏见缓解问题。

收集平衡且具有代表性的数据是防止偏见的最有效方法,但在实际操作中,这仍是最具挑战性的任务之一。为了在数据分布不平衡的情况下保证公平性,研究人员在机器学习生命周期的不同阶段提出了各种方法[9,14]。这些方法的核心原则是从数据分布的内在不平衡中学习并适应。具体而言,其实现途径为调整数据分布以实现平衡,或者优化算法本身。事实上,域增量学习也遵循这一原则:不断融入新信息,同时保留先前获取的知识,从而促进不同域的平衡学习[15]。

基于此,本文提出一个面向皮肤病诊断的公平域增量学习网络 EBID-Net(enhanced balanced incremental distillation network)。该网络结合重放、知识蒸馏以及扩展的上下文注意力机制实现不同域的平衡学习。在增量学习过程中,随着新信息的融入,模型会逐渐丢失已学习的旧知识[16]。为应对该挑战,EBID-Net 利用旧模型作为教师模型来提取关键响应信息,帮助新模型在学习新知识的同时,最大限度降低遗忘旧知识的风险。然而,由于提取的响应信息中新旧数据域之间存在不平衡,这一策略可能会引入额外的偏见。针对该问题,本文首先将一半记忆内存用于存储域平衡样本以缓解旧知识的遗忘。其次,将剩余记忆内存用于存储目标类平衡样本,并结合统计均等差异约束,以缓解目标类在不同域之间的预测偏见。研究表明,公平增强方法在提升模型公平性的过程中,往往会导致模型诊断性能下降[17]。为缓解这一问题,本文引入上下文注意力(contextual attention, ConAtt)机制[18],以增强模型对判别性特征的捕获能力。然而,ConAtt 更侧重于局部信息交互。鉴于全局信息与局部信息在模型决策过程中均具有重要作用[19-20],本文将全局信息融入 ConAtt,进一步保证模型性能。本研究的主要贡献如下:

1)提出了 EBID-Net,该网络基于感知的数据分布变化来平衡不同群体的学习。

2)设计了一个平衡记忆组件,其由类平衡记忆模块和域平衡记忆模块组成,并集成差异知识蒸馏模块以促进知识迁移并强化公平性约束。

3)扩展了 ConAtt 机制,将全局信息融入局部静态上下文模型中以捕捉不同空间位置特征间的交互。

4)在 2 个皮肤病学数据集上的实验表明,EBID-Net 在确保公平性的同时,在诊断性能和公平性之间实现了更好的权衡。

1 相关工作

1.1 域增量学习

域增量学习是增量学习的一种重要应用场景[21],致力于在任务恒定的情况下,应对样本属性分布的动态变化[15]。增量学习面临的主要挑战之一是灾难性遗忘,即深度神经网络在获取新知识时容易遗忘

已学的知识，导致难以识别旧域中的样本[22]。为解决这一问题，研究人员提出了3种主要策略：正则化[23]、重放[24-25]以及网络扩展[26]。

正则化方法通过限制权重变化来防止旧知识的过度遗忘；重放方法利用少量旧数据巩固模型记忆；网络扩展方法则通过动态调整网络架构并增加参数来适应新知识，从而避免遗忘。在正则化方法中，域增量学习已被证明能有效平衡不同领域数据的学习过程[15]。基于此，本研究探索了基于重放机制的方法在缓解皮肤病诊断不公平性方面的应用。

1.2 公平增强方法

在机器学习算法的整个生命周期中，预处理、处理中及后处理方法均能用于增强模型决策的公平性[14,27-28]。

预处理方法倾向于修改数据以防止预测模型中潜在的歧视。例如，通过数据集组合[29]构建公平数据集；利用生成对抗网络[30]或草图模型[31]生成额外的图像；直接对训练集进行重采样[32]。不过，鉴于医疗数据的稀缺性和珍贵性，此类方法需要投入大量精力来确保数据的质量和多样性。

处理中方法尝试通过修改和改变学习算法来增强模型训练过程中的公平性[17,33]。FairAdaBN(fair adaptive batch normalization)[17]通过共享部分网络参数动态整合来自子群体的共同信息。QP-Net(Quasi-Pareto net)[33]基于准帕累托改进方法，结合多任务学习和域自适应，提升模型在非代表性群体中的表现。与之不同，本文旨在利用代表性子群体的学习能力来增强模型公平性。

后处理方法作为补救策略，通过对已训练模型或待预测数据实施特定处理，以消除训练中可能存在的潜在偏见，以及训练过程中可能引入的不公平性因素[14]。剪枝策略是一种有效的改进方法，在医疗应用中更受青睐[34-35]。例如，SCP(sensitive channel pruning)[35]提出了基于软最近邻损失的通道剪枝框架，该框架通过选择性修剪对敏感属性敏感的信道实现公平性。此外，还存在其他优化策略，如DNE(debiased noise editing)[36]生成噪声来掩盖医学图像中敏感属性与疾病标签间的虚假相关性，从而降低模型对敏感属性的依赖。

2 方　　法

2.1 模型概述

在域增量学习过程中，数据分批次输入模型，每个批次对应新域或新域组。模型依据当前批次输入的数据训练，完成训练后再进入下一阶段。新域指当前批次的数据，旧域则涵盖已学习的所有数据。训练集与测试集的数据域同步更新，模型训练后需在混合测试集上评估，确保既学习新域知识又不遗忘旧域知识。本文中涉及的域由敏感属性(或人口统计属性)定义。对于每个属性，数据被划分为不同的域组。例如，性别划分为男性域和女性域。

设定一个医学成像数据集S，其中包含N个样本。数据集由3部分组成：输入图像X、敏感属性A以及真实病症标签Y。本文的核心目标是在确保模型诊断性能的同时，极力缩小不同属性域在性能上的差距。

EBID-Net的架构如图1所示，主要由特征提取模块(ResNet和扩展的上下文注意力机制)、平衡记忆更新模块(balanced memory updating module, BMU-Module)和差异知识蒸馏模块(disparity knowledge distillation module, DKD-Module)组成。每一个训练阶段，学生模型结合新域样本和内存中的旧域样本协同训练。教师模型仅提供学生模型训练过程中需要的关于旧域的知识而不更新参数。每一阶段训练结束后平衡更新内存，并复制当前模型作为下一阶段的教师模型。需要注意的是，在初始的第1阶段，记忆内存为空，无教师模型。

2.2 扩展的上下文注意力机制

注意力机制能够从图像中提取关键特征，进而保障模型性能[37]。在传统自注意力机制里，所有成对的查询-键(query-key)关系均是在孤立状态下独立学习的。为探索这些关系中丰富的上下文信息，ConAtt机制利用输入键之间的上下文信息引导注意力矩阵的学习，以此增强基于2维特征图的自注意力学习在视觉表征学习中的效能[18]。ConAtt在计算过程大量采用1×1卷积，这使其更倾向于捕捉局部空间关系。然而，在模型决策过程中，全局信息同样重要。因此，为了增强全局-局部交互，本文扩展了ConAtt，将全局信息引入局部静态上下文建模过程，形成了扩展的上下文注意力(extended contextual attention, EConAtt)，如图2所示。

在EConAtt中，输入特征为$\boldsymbol{X}_{\text{in}} \in \mathbb{R}^{H\times W\times C}$，并定义了3个变量：键($\boldsymbol{K}$)、查询($\boldsymbol{Q}$)和值($\boldsymbol{V}$)($\boldsymbol{K}=\boldsymbol{X}_{\text{in}}$；$\boldsymbol{Q}=\boldsymbol{X}_{\text{in}}$；$\boldsymbol{V}=\boldsymbol{X}_{\text{in}}\boldsymbol{W}_{\text{v}}$，其中的$\boldsymbol{W}_{\text{v}}$为嵌入矩阵)。首先，对$\boldsymbol{K}$应用一个$k\times k$的分组卷积，以提取包含局部上下文信息的特征表示$\boldsymbol{K}^1$，这与局部信息的静态建模类似。

接下来，拼接$\boldsymbol{K}^1$与输入特征$\boldsymbol{X}_{\text{in}}$的全局平均池化(global average pooling, GAP)的输出，从而引入全局特征$\boldsymbol{G}$。随后，将$\boldsymbol{Q}$与$\boldsymbol{K}^1$的拼接结果与全局特征$\boldsymbol{G}$

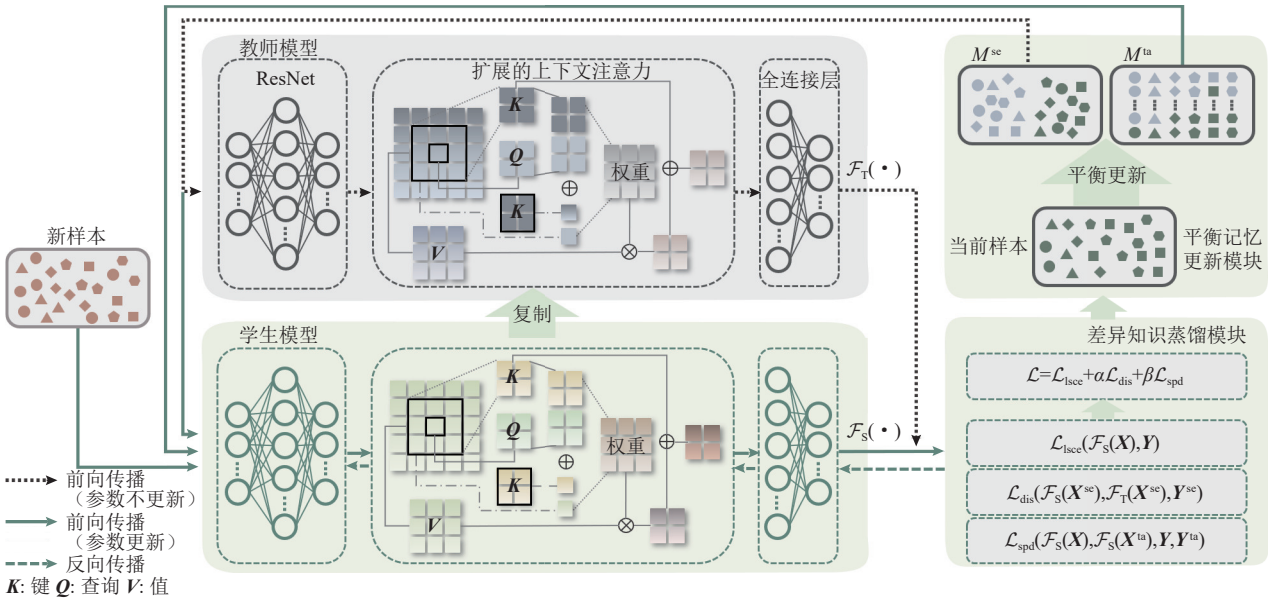


Fig. 1 The main framework of EBID-Net

图 1 EBID-Net 主要框架

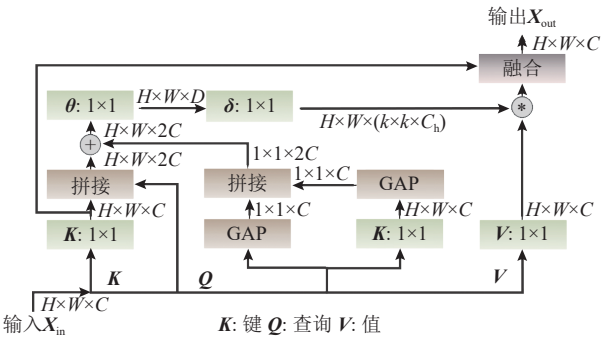


Fig. 2 Illustration of EConAtt

图 2 EConAtt 示意图

进行逐元素相加并经过 2 个连续的卷积层得到 2 维注意力图。然后，将得到的注意力图与 $\boldsymbol{V}$ 相乘，获得动态上下文建模特征图。最后，通过融合局部静态上下文建模与全局静态上下文建模，生成输出特征图。

$$\boldsymbol{G}=Concat(GAP(\boldsymbol{K}\boldsymbol{W}_{\mathrm{k}}),GAP(\boldsymbol{X}_{\mathrm{in}}))\,, \tag{1}$$

$$\boldsymbol{D}=Addition(Concat(\boldsymbol{K}\boldsymbol{W}_{\mathrm{k}},\boldsymbol{Q}),\boldsymbol{G})\boldsymbol{W}_{\theta}\boldsymbol{W}_{\delta}\,, \tag{2}$$

$$\boldsymbol{X}_{\mathrm{out}}=Fusion(\boldsymbol{D},\boldsymbol{V}\circledast\boldsymbol{D})\,, \tag{3}$$

其中 $\boldsymbol{W}_{\mathrm{k}}$, $\boldsymbol{W}_{\theta}$, $\boldsymbol{W}_{\delta}$ 均为嵌入矩阵。

2.3 平衡记忆更新模块

在增量学习中，为防止灾难性遗忘，部分研究引入记忆内存组件[38]以保留少量旧样本用于新域的学习[39]。由于记忆内存中存储的旧域样本数量有限，因此在当前训练阶段中，旧域训练样本数量会远少于新域训练样本。为减轻因训练数据分布不平衡而产生的额外偏见，通常会在训练阶段结束时应用平衡微调组件[38]。该组件采用平衡采样技术生成一个平衡的子集，并使用较小的学习率对模型进行微调。

本文整合增量记忆内存组件与平衡微调组件的优势，提出了一种平衡记忆内存组件设计方案。考虑到数据集中敏感属性域和目标类别均存在分布不平衡的情况，采用平衡记忆内存组件由内存大小相同的目标类别平衡记忆模块 M^{ta} 和敏感属性域平衡记忆模块 M^{se} 组成。该设计旨在实现更为平衡且有效的模型训练，从而提高模型的公平性和诊断性能。

$$M^{\mathrm{ta}}=\left\{\left(x_i^{\mathrm{ta}},a_i^{\mathrm{ta}},y_i^{\mathrm{ta}}\right)_i^{B}\right\},\left|X_y\right|=B/2|U|\,, \tag{4}$$

$$M^{\mathrm{se}}=\left\{\left(x_i^{\mathrm{se}},a_i^{\mathrm{se}},y_i^{\mathrm{se}}\right)_i^{B}\right\},|X_a|=B/2|R|\,, \tag{5}$$

其中 X_y 表示目标类别为 y 的子集，X_a 表示属性为 a 的子集，B 为记忆容量大小，U 和 R 分别对应目标类别集合与敏感属性域集合。

算法 1. 平衡记忆内存更新。

输入：第 i 个学习阶段 T_i，阶段 T_i 新域数据集 XB_i，阶段 T_i 记忆内存中的旧域数据集 XM_i，目标类别数量 $|U|$，记忆内存大小 B，阶段 T_i 旧域数量 R_{i-1} 和新域数量 R_i，阶段 T_i 目标类别平衡记忆内存 M_i^{ta}，T_i 阶段敏感属性域平衡记忆内存 M_i^{se}；

输出：样本集 XM_{i+1}。

① for y in U

② $XB_i^{y}\leftarrow$ 选择样本(XB_i,y);

③ $M_i^{\mathrm{ta},y}\leftarrow$ 选择样本(M_i^{ta},y);

④ $n_{\mathrm{ta}}=(B/2|U|)$;

⑤ $count = |M_i^{\mathrm{ta},y}|$;
⑥ for 元素 in XB_i^y
⑦ $count++$;
⑧ 以概率$\min(1, n_{\mathrm{ta}}/count)$将元素存入$M_i^{\mathrm{ta},y}$（蓄水池采样）;
⑨ end for
⑩ end for
⑪ $M_{i+1}^{\mathrm{ta}} = M_i^{\mathrm{ta},0} \cup M_i^{\mathrm{ta},2} \cup \cdots \cup M_i^{\mathrm{ta},(|U|-1)}$;
⑫ for a in $R_{i-1} \cup R_i$
⑬ $n_{\mathrm{se}} = B/2(|R_{i-1}| + |R_i|)$;
⑭ if $a \in R_{i-1}$
⑮ $M_i^{\mathrm{se},a} = \{x \in M_i^{\mathrm{se}} | \text{域}(x) == a\}$;
⑯ $M_i^{\mathrm{se},a} \leftarrow$ 随机采样$(M_i^{\mathrm{se},a}, n_{\mathrm{se}})$;
⑰ else
⑱ $XB_i^a = \{x \in XB_i | \text{域}(x) == a\}$;
⑲ $M_i^{\mathrm{se},a} \leftarrow$ 随机采样$(XB_i^a, n_{\mathrm{se}})$;
⑳ end if
㉑ end for
㉒ $M_{i+1}^{\mathrm{se}} = M_i^{\mathrm{se},0} \cup M_i^{\mathrm{se},1} \cup \cdots \cup M_i^{\mathrm{se},(|R_{i-1}|+|R_i|-1)}$;
㉓ return $M_{i+1}^{\mathrm{ta}} \cup M_{i+1}^{\mathrm{se}}$。

2.4 差异知识蒸馏模块

来自同一类别的样本在特征空间中应呈现出相似的嵌入表示。因此，学生模型产生的响应信息应与教师模型中同一类别的响应信息保持一致。蒸馏损失用于衡量学生模型与教师模型输出分布之间的差异，促使学生模型的输出分布向教师模型对齐。在本文设定中，每个新学习阶段开始时，学生模型与教师模型的参数初始化为相同状态。然而，随着学生模型开始接触并学习新域数据，其参数会迭代更新。为确保学生模型在有效学习新域知识的同时保留旧域知识，本文利用蒸馏损失来缓解遗忘并促进知识迁移。

在此约束中，特别指定输入数据为M^{se}。M^{se}中的数据在已学习域间保持平衡。这为学生模型构建了一个仅包含旧域且数据分布平衡的学习环境，使其能够更好地巩固已学知识。通过这种方式确保学生模型保留对旧知识的记忆，从而促进知识的有效迁移与整合。蒸馏损失的计算公式为：

$$\mathcal{L}_{\mathrm{dis}} = -\sum_{i=1}^{N}\sum_{j=1}^{U} q_{\mathrm{t}}(x'_{i,j}) \ln\left(q_{\mathrm{s}}(x'_{i,j})\right), \tag{6}$$

$$q_{\mathrm{t}}(x'_{i,j}) = \frac{\left(q_{\mathrm{t}}(x_{i,j})\right)^{\frac{1}{T}}}{\sum_{i}\left(q_{\mathrm{t}}(x_{i,j})\right)^{\frac{1}{T}}}, \tag{7}$$

$$q_{\mathrm{s}}(x'_{i,j}) = \frac{\left(q_{\mathrm{s}}(x_{i,j})\right)^{\frac{1}{T}}}{\sum_{i}\left(q_{\mathrm{s}}(x_{i,j})\right)^{\frac{1}{T}}}, \tag{8}$$

其中N表示样本数量，U表示目标类别数，T表示蒸馏温度系数（默认设置为2），$q_{\mathrm{t}}(x_{i,j})$和$q_{\mathrm{s}}(x_{i,j})$分别表示第i样本为j类病症的教师模型和学生模型的预测概率。

本文的主要目标是增强学生模型的公平性。因此，需要保证模型能够公平地对待不同的敏感属性域。统计均等性差异（statistical parity disparity）特指在模型预测为正类（或目标类）的实例中，不同属性群体所占比例的差异[40]。为此，本文引入统计均等性差异损失作为公平性约束条件。该损失函数旨在最小化模型对不同属性群体目标类预测结果的差异。为进一步提升学生模型的学习能力，此约束中的输入数据不仅包含当前域数据，还纳入了来自M^{ta}的数据。M^{ta}中的数据在目标类间平衡。统计均等性差异损失的计算公式为：

$$\mathcal{L}_{\mathrm{spd}} = \sum_{y=1}^{U} \left\| \mathbb{E}_{\boldsymbol{X}_i \sim S_{A=0}} \left| (q(\boldsymbol{X}_i) = y) - \mathbb{E}_{\boldsymbol{X}_i \sim S_{A=1}} \left| (q(\boldsymbol{X}_i) = y) \right\|^2 \right. \right. 。 \tag{9}$$

2.5 总体目标

综上所述，EBID-Net的总训练目标表示为

$$\mathcal{L} = \mathcal{L}_{\mathrm{lsce}} + \alpha\mathcal{L}_{\mathrm{dis}} + \beta\mathcal{L}_{\mathrm{spd}}, \tag{10}$$

其中权重因子α和β分别用于控制知识迁移与公平性约束的强度。标签平滑[41]是神经网络中常用的一种正则化方法，旨在防止模型过拟合。标签平滑交叉熵损失用于衡量输出期望概率分布与标签平滑后实际概率分布之间的差异。因此，损失函数$\mathcal{L}_{\mathrm{lsce}}$负责整体性能优化，其计算方式为：

$$\mathcal{L}_{\mathrm{lsce}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{U} \tilde{p}(x_{i,j}) \ln\left(\tilde{q}(x_{i,j})\right), \tag{11}$$

$$\tilde{p}(x_{i,j}) = \begin{cases} 1-\epsilon, & \text{if } i = j, \\ \epsilon/U - 1, & \text{其他。} \end{cases} \tag{12}$$

其中$\tilde{p}(x_{i,j})$和$\tilde{q}(x_{i,j})$分别表示经标签平滑处理后的真实概率与预测概率。ϵ是一个较小的常数（设$\epsilon = 0.01$），用于防止概率出现极端值（即1或0）。

3 实验设置

3.1 数据集

在公平皮肤病诊断领域的研究中，数据集的选择至关重要，其需要具备高质量的图像和属性信息以及广泛的认可度。基于此，本文与现有研究一致，选用

了 Fitzpatrick17k[42] 与 ISIC 2019[43-45] 这 2 个数据集。

Fitzpatrick17k 数据集是皮肤病诊断领域内重要的通用公开数据集。它包含 16 577 张皮肤病学图像，这些图像涉及 9 个诊断类别，且每张图像都附有 Fitzpatrick 皮肤类型标签。皮肤类型从Ⅰ~Ⅵ共分为 6 个等级，等级越低，肤色越浅。数据处理时，本文保留了带有皮肤类型标签的图像，并将Ⅰ~Ⅲ型皮肤设置为浅色皮肤，Ⅳ~Ⅵ型皮肤设置为深色皮肤。最终，浅色图像与深色图像的比例约为 7∶3。值得注意的是，尽管浅色皮肤图像数量较多，但在标准条件下训练的模型，在深色皮肤图像上的准确率反而更高，因此选择深色群体为代表性群体。

ISIC 2019 数据集是国际皮肤影像协作组织主办的皮肤影像分析挑战赛的竞赛数据集。它包含 25 331 张皮肤病学图像，涉及 8 个诊断类别。本文排除了没有性别标签的图像，并将性别作为敏感属性。处理后，女性与男性图像的比例约为 4.5∶5.5。使用数据集训练的基准模型在女性图像上表现出更高的准确率，因此选择女性群体为代表性群体。

3.2 评估指标

为了全面评估模型性能，本文从诊断性能、公平性以及诊断性能和公平性的权衡 3 个维度进行评估。在评估诊断性能时，采用的指标为精确率、召回率和 $F1$ 值。关于公平指标，本文引入平等机会(equal opportunity，$EOpp$)[46] 和均衡概率(equalized odds，$EOdd$)[46]。基于实际标签，平等机会考虑了 $EOpp0$ 和 $EOpp1$ 这 2 种情况。

$$EOpp0=|P(\hat{Y}=0\mid Y=0,A=1)-P(\hat{Y}=0\mid Y=0,A=0)|, \tag{13}$$

$$EOpp1=|P(\hat{Y}=1\mid Y=1,A=1)-P(\hat{Y}=1\mid Y=1,A=0)|, \tag{14}$$

$$EOdd=|P(\hat{Y}=1\mid Y=y,A=1)-P(\hat{Y}=1\mid Y=y,A=0)|,\ y\in\{0,1\}, \tag{15}$$

其中 A，Y，$\hat{Y}$ 分别表示敏感属性、真实标签和预测标签。

为了评估性能和公平性之间的权衡关系，本文采用 $FATE$[17](fairness-accuracy trade-off efficiency)指标。该指标可量化公平性标准化提升与诊断性能降低之间的权衡情况。其公式如下：

$$FATE_{FC}=\frac{ACC_{\mathrm{e}}-ACC_{\mathrm{b}}}{ACC_{\mathrm{b}}}-\lambda\frac{FC_{\mathrm{e}}-FC_{\mathrm{b}}}{FC_{\mathrm{b}}}, \tag{16}$$

其中 FC 取值为 $EOpp0$，$EOpp1$，$EOdd$ 中的 1 种。ACC 代表诊断性能，后续采用精确率进行比较分析。下标 e 和 b 用于区分公平增强模型和基线模型。λ 是加权因子，用于调整用户预定义的公平性要求。在实验评估中，λ 设置为 1.0。

3.3 实验细节

与文献 [17] 的处理方式一致，本文将数据集按照 6∶2∶2 的比例随机划分为训练集、验证集和测试集。随后，采用与文献 [32] 相同的数据增强技术，具体包括随机翻转、旋转和缩放操作。图像裁剪为 224×224 像素大小。本文用 ResNet18 作为主干网络，并使用 AdamW 优化器开展训练。其中，批量大小设置为 128，学习率设置为 1E−4，缓冲区大小设置为 600。对于不同的数据集，系数设置有所不同。在 Fitzpatrick17k 数据集中，α 和 β 分别设置为 0.6 和 1。对于 ISIC 2019 数据集，系数 α 和 β 分别设置为 0.8 和 1.25。所有实验均在 NVIDIA RTX A5000 显卡上进行，所有 $FATE$ 指标以 ResNet18 作为基线进行评估。

3.4 对比方法

EBID-Net 将与基准(ResNet18)和其他公平性增强模型进行比较。具体包括：重采样(resampling)[32]，GroupDRO(group distributionally robust optimization)[47]，EnD(entangling and disentangling)[48]，HSIC(Hilbert-Schmidt norm)[49]，MFD(MMD-based fair distillation)[50]，ME[51]，FairAdaBN[17]，QP-Net[33]，SCP[35]，DNE[36]，FairMoE[52]。

4 结果与分析

4.1 与公平增强模型的对比

1)Fitzpatrick17k 数据集上的实验结果

表 1 展示了在 Fitzpatrick17k 数据集上的诊断性能、公平性指标及二者的权衡效果。首先，EBID-Net 在所有公平性指标上均优于其他对比方法。其次，尽管与最优诊断性能结果相比 EBID-Net 存在轻微差距，但相较于公平性表现次优的 FairAdaBN 模型，其在精确率、召回率和 $F1$ 值方面均有小幅提升。上述结果表明，在该数据集上，公平性的提升未以显著牺牲诊断性能为代价，二者之间的冲突较弱。在诊断性能与公平性的权衡方面，EBID-Net 在所有指标上均取得最高值。这些结果表明，EBID-Net 在保持诊断性能的同时，有效提升了模型的公平性。

2)ISIC 2019 数据集上的实验结果

与 Fitzpatrick17k 数据集相比，ISIC 2019 数据集多了约 9 000 张图像，这增加了模型同时保证诊断性能和公平性的难度。尽管 ISIC 2019 数据集中敏感属性的分布更为平衡，但如表 2 所示，各方法在该数据

Table 1 Results of Diagnostic Performance, Fairness and Trade-off of Various Methods on Fitzpatrick17k Dataset

表 1 不同方法在 Fitzpatrick17k 数据集上的诊断性能、公平性及权衡结果

方法	肤色	诊断性能 /%			公平性			权衡		
		精确率↑	召回率↑	*F*1 值↑	*EOpp*0↓	*EOpp*1↓	*EOdd*↓	$FATE_{EOpp0}$↑	$FATE_{EOpp1}$↑	$FATE_{EOdd}$↑
ResNet18	深色	82.86±0.013	82.97±0.012	81.80±0.013	1.03±0.101	9.88±1.468	10.38±1.791	0.00±0.000	0.00±0.000	0.00±0.000
	浅色	75.46±0.003	76.64±0.001	74.94±0.003						
	总体	77.68±0.003	78.62±0.004	77.13±0.002						
Resampling[32]	深色	83.38±0.009	83.60±0.008	82.63±0.007	0.88±0.167	11.01±1.191	11.66±1.370	15.66±0.155	−10.67±0.113	−11.59±0.125
	浅色	76.18±0.008	77.16±0.005	75.68±0.003						
	总体	78.28±0.006	79.17±0.005	77.86±0.003						
GroupDRO[47]	深色	80.38±0.007	79.87±0.011	79.72±0.008	0.94±0.179	9.23±0.334	9.77±0.398	6.58±0.173	4.74±0.033	4.07±0.037
	浅色	74.54±0.008	74.71±0.011	74.07±0.009						
	总体	76.25±0.007	76.32±0.011	75.83±0.008						
EnD[48]	深色	79.90±0.003	79.74±0.003	79.45±0.003	1.05±0.142	8.31±1.567	8.67±1.753	−5.26±0.133	12.53±0.163	13.19±0.173
	浅色	72.99±0.007	72.71±0.010	72.59±0.008						
	总体	75.10±0.004	74.91±0.007	74.74±0.005						
HSIC[49]	深色	78.76±0.005	79.34±0.006	78.69±0.006	0.81±0.052	8.06±0.651	8.20±0.720	18.68±0.052	15.46±0.063	18.03±0.066
	浅色	73.89±0.002	74.95±0.003	74.15±0.002						
	总体	75.35±0.003	76.32±0.004	75.56±0.003						
MFD[50]	深色	82.72±0.004	82.87±0.003	81.87±0.002	0.90±0.045	9.34±0.308	9.38±0.334	12.73±0.043	5.89±0.033	10.06±0.034
	浅色	76.10±0.002	76.96±0.001	76.12±0.002						
	总体	78.02±0.001	78.81±0.001	77.94±0.001						
ME[51]	深色	80.71±0.002	80.71±0.002	80.32±0.002	0.95±0.012	11.25±0.324	11.77±0.349	5.40±0.011	−16.52±0.033	−16.11±0.034
	浅色	73.49±0.002	74.25±0.002	73.67±0.001						
	总体	75.59±0.001	76.20±0.001	75.68±0.001						
FairAdaBN[17]	深色	77.50±0.006	78.67±0.005	77.05±0.005	0.66±0.078	7.10±0.202	7.37±0.090	28.75±0.071	21.25±0.023	22.18±0.005
	浅色	70.26±0.005	72.23±0.004	69.89±0.002						
	总体	72.36±0.005	74.24±0.004	72.14±0.003						
QP-Net[33]	深色	79.13±0.006	77.84±0.004	77.96±0.004	0.74±0.060	8.41±1.189	8.66±1.146	24.25±0.056	10.68±0.123	12.34±0.113
	浅色	72.46±0.003	71.61±0.006	71.73±0.004						
	总体	74.40±0.002	73.56±0.003	73.66±0.002						
SCP[35]	深色	80.96±0.001	81.54±0.004	80.73±0.002	0.76±0.043	8.45±0.830	8.85±0.813	25.41±0.053	13.71±0.073	13.91±0.068
	浅色	75.30±0.013	76.22±0.016	75.15±0.012						
	总体	77.06±0.009	77.88±0.012	76.92±0.008						
DNE[36]	深色	80.65±0.003	79.91±0.001	80.06±0.002	0.69±0.107	8.82±0.496	9.26±0.544	30.40±0.105	8.44±0.047	8.47±0.050
	浅色	73.97±0.003	73.82±0.007	73.81±0.005						
	总体	75.91±0.003	75.72±0.005	75.74±0.004						
FairMoE[52]	深色	75.99±0.019	76.25±0.020	75.52±0.023	0.79±0.026	7.68±0.312	7.86±0.344	15.45±0.032	14.74±0.018	16.78±0.017
	浅色	70.17±0.011	70.45±0.016	69.69±0.019						
	总体	71.83±0.013	72.26±0.018	71.49±0.020						
EBID-Net（本文）	深色	78.52±0.002	77.51±0.005	77.65±0.004	0.59±0.107	6.92±0.156	7.22±0.210	39.43±0.108	26.96±0.012	27.44±0.019
	浅色	74.26±0.003	74.75±0.005	74.37±0.004						
	总体	75.38±0.003	75.61±0.005	75.37±0.004						

注：“↑”表示该指标值越高越好，“↓”表示该指标值越低越好。“±”前数值为平均值，“±”后数值为标准差。

Table 2 Results of Diagnostic Performance, Fairness and Trade-off of Various Methods on ISIC 2019 Dataset

表 2 不同方法在 ISIC 2019 数据集上的诊断性能、公平性及权衡结果

方法	性别	诊断性能/%			公平性			权衡		
		精确率↑	召回率↑	$F1$ 值↑	$EOpp0$↓	$EOpp1$↓	$EOdd$↓	$FATE_{EOpp0}$↑	$FATE_{EOpp1}$↑	$FATE_{EOdd}$↑
ResNet18	男性	80.61±0.005	80.76±0.006	80.36±0.006	0.81±0.025	5.73±0.772	5.50±0.936	0.00±0.00	0.00±0.00	0.00±0.00
	女性	83.02±0.005	83.19±0.005	82.92±0.005						
	总体	81.69±0.005	81.91±0.006	81.58±0.006						
Resampling[32]	男性	82.02±0.001	82.30±0.001	81.99±0.001	0.87±0.046	5.86±0.548	5.73±0.475	−5.77±0.057	−1.10±0.096	−2.89±0.087
	女性	83.47±0.001	83.76±0.002	83.36±0.001						
	总体	82.69±0.001	82.99±0.001	82.65±0.001						
GroupDRO[47]	男性	80.02±0.010	79.64±0.014	79.45±0.012	0.78±0.063	6.08±0.436	5.90±0.728	3.17±0.085	−6.99±0.071	−8.15±0.129
	女性	81.99±0.008	81.84±0.010	81.70±0.008						
	总体	80.92±0.009	80.68±0.012	80.53±0.010						
EnD[48]	男性	78.90±0.004	78.76±0.007	78.58±0.007	0.86±0.033	5.18±0.454	4.71±0.493	−8.45±0.037	7.73±0.074	12.44±0.085
	女性	81.69±0.005	81.38±0.009	81.26±0.008						
	总体	80.17±0.004	80.00±0.007	79.85±0.007						
HSIC[49]	男性	80.18±0.003	80.52±0.003	80.11±0.003	0.64±0.061	5.22±0.257	5.34±0.122	20.07±0.031	7.51±0.041	1.52±0.018
	女性	81.14±0.004	81.45±0.004	81.10±0.004						
	总体	80.60±0.003	80.96±0.004	80.59±0.004						
MFD[50]	男性	79.48±0.007	79.55±0.009	79.30±0.007	0.72±0.022	5.15±0.435	5.10±0.382	9.59±0.031	8.78±0.079	5.69±0.075
	女性	81.64±0.002	81.70±0.003	81.61±0.002						
	总体	80.45±0.004	80.57±0.006	80.41±0.005						
ME[51]	男性	80.53±0.005	80.45±0.006	80.40±0.005	0.59±0.008	4.63±0.455	4.34±0.610	26.13±0.009	18.11±0.078	20.06±0.110
	女性	81.33±0.004	81.14±0.006	81.15±0.005						
	总体	80.85±0.004	80.78±0.006	80.75±0.005						
FairAdaBN[17]	男性	76.52±0.005	77.14±0.005	76.57±0.005	0.60±0.005	4.25±0.250	3.87±0.208	20.98±0.004	20.53±0.042	4.28±0.036
	女性	78.37±0.003	78.91±0.003	78.33±0.003						
	总体	77.32±0.002	77.98±0.003	77.41±0.002						
QP-Net[33]	男性	78.64±0.011	78.08±0.016	78.17±0.014	0.86±0.063	4.38±0.065	4.26±0.115	−10.00±0.073	20.09±0.009	19.13±0.010
	女性	79.40±0.011	78.74±0.019	78.94±0.017						
	总体	78.90±0.011	78.39±0.017	78.50±0.015						
SCP[35]	男性	80.65±0.001	80.77±0.001	80.59±0.001	0.65±0.045	4.27±0.136	4.03±0.169	19.27±0.056	25.35±0.022	26.66±0.029
	女性	82.80±0.005	83.00±0.007	82.82±0.005						
	总体	81.63±0.002	81.83±0.004	81.65±0.003						
DNE[36]	男性	80.39±0.002	80.62±0.002	80.44±0.002	0.60±0.024	4.10±0.098	4.39±0.139	15.18±0.031	27.64±0.016	19.37±0.024
	女性	81.88±0.001	82.01±0.001	81.89±0.001						
	总体	81.08±0.001	81.28±0.001	81.14±0.001						
FairMoE[52]	男性	81.62±0.002	81.70±0.003	81.21±0.003	0.60±0.042	3.81±0.214	3.82±0.361	26.84±0.052	34.01±0.037	31.11±0.064
	女性	82.71±0.003	82.84±0.002	82.36±0.002						
	总体	82.10±0.002	82.24±0.001	81.76±0.002						
EBID-Net（本文）	男性	80.17±0.004	80.16±0.005	80.05±0.005	0.56±0.024 1	3.65±0.376 6	3.50±0.441	29.47±0.031	34.85±0.066	34.97±0.080
	女性	81.06±0.002	81.13±0.003	81.00±0.003						
	总体	80.56±0.003	80.62±0.003	80.50±0.003						

注：“↑”表示该指标值越高越好，“↓”表示该指标值越低越好。“±”前数值为平均值，“±”后数值为标准差。

集上的性能表现并不优于在 Fitzpatrick17k 数据集上的表现。此外，还有多个公平增强模型在公平性指标上表现逊于基础模型。上述现象表明，数据规模的扩大不仅增加了优化难度，也加剧了公平性与诊

断性能之间的冲突。在此环境下，EBID-Net 取得了最优的公平性表现。FairMoE 模型分别在 3 项公平性指标上获得次优结果。但值得注意的是，EBID-Net 的诊断性能略低于多数对比模型。这一现象表明，在该数据集上，公平性与诊断性能之间存在明显的权衡关系：公平性的显著提升以诊断性能的轻微下降为代价。尽管如此，在评估权衡指标时，EBID-Net 仍取得了最优结果。

4.2 消融实验

1）训练顺序的影响

实验结果表明，增量学习的有效性与其训练顺序密切相关。如表 3 和表 4 所示，采用“深色—浅色”训练顺序时，模型在不同域上的性能表现及平衡效果均优于“浅色—深色”训练顺序。类似地，与“男性—女性”训练顺序相比，采用“女性—男性”训练顺序时，模型的性能及跨域平衡性更佳。在此情况下，深色群体和女性群体可被视为代表性群体。上述规律表明，从代表性群体到非代表性群体的训练顺序，有助于实现公平性与诊断性能的权衡优化。

本文的核心设计理念在于：从代表性群体到非代表性群体的训练过程中学习与决策相关的知识。在数据分布不平衡的数据集中，数据量较大的域群体通常被视为代表性群体。然而，对于 Fitzpatrick17k 数据集而言，深色图像的实际数量少于浅色图像。因此，域群体是否具有代表性，并不仅仅取决于该域群体的数据量。

2）EConAtt 的影响

注意力机制能增强输出特征图的表达能力。如表 3 和表 4 所示，与未使用注意力机制的 EBID-Net 相比，使用注意力机制的 EBID-Net 在整体及跨域诊断性能上均有一定提升。这表明特征表达能力的增强对下游任务具有正向促进作用，且有利于公平性与诊断性能之间权衡。此外，使用 EConAtt 的 EBID-Net 在“深色—浅色”“女性—男性”训练顺序下，于非代表性域中取得了更优的诊断性能，公平性也得到提升。这证明了引入全局信息对知识迁移的有效性。

3）蒸馏损失的影响

在总目标损失中，蒸馏损失系数α用于控制蒸馏损失对整体损失的贡献。图 3(a)~(c)分别展示了在 Fitzpatrick17k 数据集上，不同α值对诊断性能、公平

Table 3 Impact of Training Order and Attention Mechanism on Diagnostic Performance, Fairness and Trade-off on Fitzpatrick17k Dataset

表 3 Fitzpatrick17k 数据集上训练顺序与注意力机制对诊断性能、公平性及权衡关系的影响

训练顺序	注意力机制	肤色	诊断性能/%			公平性			权衡		
			精确率↑	召回率↑	*F*1 值↑	*EOpp*0↓	*EOpp*1↓	*EOdd*↓	$FATE_{EOpp0}$↑	$FATE_{EOpp1}$↑	$FATE_{EOdd}$↑
深色—浅色	w/o	深色	77.31	75.85	76.30	0.73	7.27	7.31	24.99	22.28	25.44
		浅色	73.41	73.79	73.46						
		总体	74.47	74.43	74.32						
	w/ ConAtt	深色	77.53	77.05	77.09	0.68	7.17	7.28	31.21	24.66	27.09
		浅色	74.93	75.92	75.11						
		总体	75.53	76.27	75.72						
	w/ EConAtt	深色	78.29	76.95	77.29	0.62	6.72	6.93	36.48	28.66	29.91
		浅色	73.95	74.10	73.94						
		总体	75.10	74.99	74.95						
浅色—深色	w/o	深色	81.00	81.54	81.03	1.33	12.17	13.32	−33.97	−28.03	−33.17
		浅色	70.82	72.15	71.29						
		总体	73.91	75.09	74.33						
	w/ ConAtt	深色	81.90	82.44	81.94	1.37	11.20	12.22	−37.42	−17.77	−22.14
		浅色	70.91	71.97	71.33						
		总体	74.25	75.24	74.64						
	w/ EConAtt	深色	80.41	81.04	80.40	0.98	11.74	12.46	0.91	−22.76	−23.97
		浅色	72.04	73.11	72.33						
		总体	74.62	75.58	74.87						

注：“↑”表示该指标值越高越好，“↓”表示该指标值越低越好。

Table 4 Impact of Training Order and Attention Mechanism on Diagnostic Performance, Fairness and Trade-off on ISIC 2019 Dataset

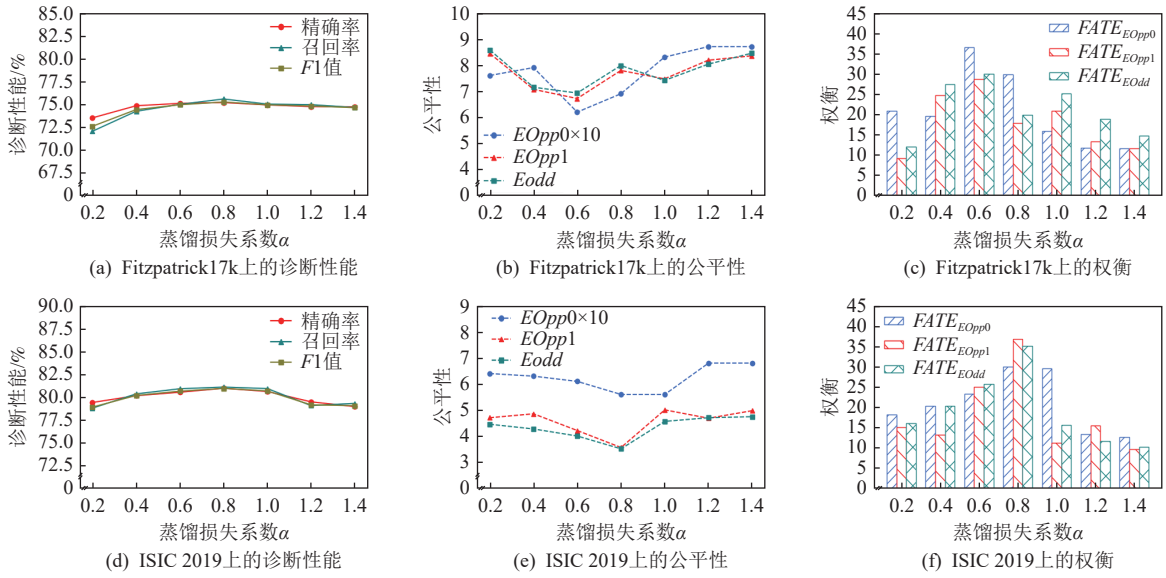
表 4 ISIC 2019 数据集上训练顺序与注意力机制对诊断性能、公平性及权衡关系的影响

训练顺序	注意力机制	性别	诊断性能/%			公平性			权衡		
			精确率↑	召回率↑	*F*1 值↑	*EOpp*0↓	*EOpp*1↓	*EOdd*↓	$FATE_{EOpp0}$↑	$FATE_{EOpp1}$↑	$FATE_{EOdd}$↑
女性—男性	w/o	男性	79.86	79.46	80.46	0.67	3.99	3.85	15.08	28.16	27.79
		女性	80.46	80.82	80.41						
		总体	79.89	80.31	79.92						
	w/ ConAtt	男性	79.75	79.97	79.70	0.57	3.88	3.71	27.90	30.56	30.81
		女性	80.97	81.20	80.94						
		总体	80.28	80.55	80.29						
	w/ EConAtt	男性	80.73	80.81	80.68	0.56	3.57	3.52	29.97	36.80	35.10
		女性	81.26	81.41	81.24						
		总体	80.96	81.09	80.95						
男性—女性	w/o	男性	78.07	78.10	77.99	0.92	5.41	5.62	−16.90	2.25	−5.51
		女性	80.04	80.02	80.01						
		总体	78.97	79.01	78.95						
	w/ ConAtt	男性	78.89	79.28	78.91	0.86	5.15	5.25	−8.57	7.72	2.14
		女性	80.78	81.16	80.88						
		总体	79.73	80.17	79.85						
	w/ EConAtt	男性	78.38	78.94	78.51	0.924	4.85	5.26	−15.94	12.99	2.00
		女性	81.34	81.75	81.39						
		总体	79.76	80.27	79.88						

注: “↑”表示该指标值越高越好, “↓”表示该指标值越低越好。

性及二者权衡效果的影响。实验结果显示, 蒸馏损失能够间接影响公平性与诊断性能的权衡关系。随着α逐渐增大至 0.6, 包括精确率、召回率和 *F*1 值在内的诊断性能指标均呈现缓慢同步提升的趋势。公平性指标 *EOpp*0, *EOpp*1, *EOdd* 也同步改善, 并达到最优值。相应地, 图 3(c)中的权衡性能也达到最佳。此外, 图 3(d)~(f)展示了在 ISIC 2019 数据集上的实验结果。随着α逐渐增大至 0.8, 其变化趋势与 Fitzpatrick17k 数据集上观察到的趋势一致。

然而, 当α继续增大时, 网络性能不再提升。这



(a) Fitzpatrick17k上的诊断性能 (b) Fitzpatrick17k上的公平性 (c) Fitzpatrick17k上的权衡

(d) ISIC 2019上的诊断性能 (e) ISIC 2019上的公平性 (f) ISIC 2019上的权衡

Fig. 3 Impact of distillation loss

图 3 蒸馏损失的影响

是因为蒸馏损失在网络中的主要作用是缓解旧知识的遗忘。如图 3(a)所示，当α从 0.6 增大至 1.4 时，模型的诊断性能呈现轻微下降趋势。同时，如图 3(b)所示，公平性指标开始急剧下降。此时，联合损失中负责分类的交叉熵损失和负责公平性的统计均等差异损失的贡献均减小，自然导致模型的分类性能和公平性均下降。因此，模型的权衡性能也迅速下降。值得注意的是，ISIC 2019 数据集上的实验结果与 Fitzpatrick17k 数据集上的结果高度一致。这表明，蒸馏损失对模型性能、公平性以及权衡的影响在不同数据集上具有普适性。

4)统计均等差异损失的影响

在网络中，统计均等差异损失的主要作用是确保不同属性类别间的公平性。在总目标损失中，系数β用于调节统计均等差异损失对整体损失的影响。对于 Fitzpatrick17k 数据集，如图 4(a)所示，随着系数β的变化，模型的诊断性能保持相对稳定。而对于 ISIC 2019 数据集，如图 4(d)所示，诊断性能在β增大初期保持相对稳定，随后略有下降。

2 个数据集的公平性并未随着β的增大而持续改善，而是在达到最优值后呈现先增强后减弱的趋势。因此，统计差异损失所施加的约束虽能提升公平性，但若约束过于严格，则这种提升效果可能受到限制。特别地，与其他 2 项公平性指标相比，*EOpp*0 在达到最优结果后的下降幅度并不显著。综合考虑诊断性能与公平性，模型的权衡效果如图 4(c)(f)所示，呈现先上升后下降的趋势。这验证了公平性与诊断性能之间的权衡关系，当公平约束强度处于合理区间时，二者达到最优权衡，过弱或过强的公平约束均会导致综合性能下降。

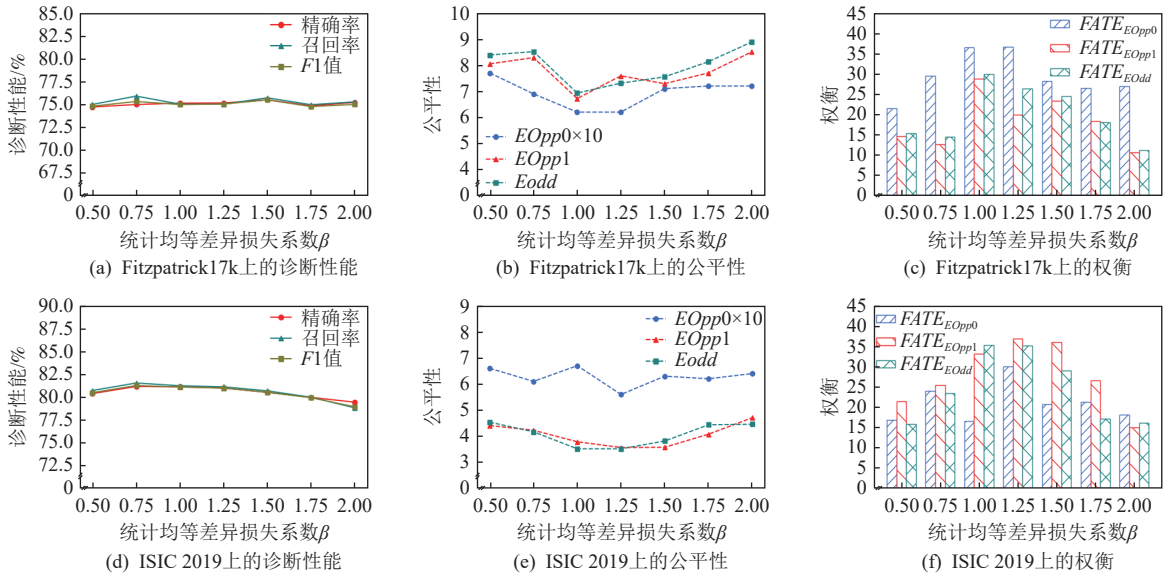


Fig. 4 Impact of statistical parity disparity loss

图 4 统计均等差异损失的影响

5)平衡记忆更新的影响

大多数现有的记忆内存更新策略在数据更新过程中会优先保留代表性样本，这往往导致记忆数据分布出现严重失衡。此类分布失衡现象可能在模型训练阶段引入额外的偏见。如表 5 所示，采用平衡更新策略后，模型在公平性和权衡效果 2 方面均有所提升。这些结果表明，通过平衡更新策略约束数据分布，能够有效缓解潜在偏见，进而优化权衡指标。

6) 记忆内存大小的影响

本节评估了记忆内存大小对模型性能的影响。内存大小被划分为 5 个等级，最小为 200，最大为 1 000，间隔为 200。

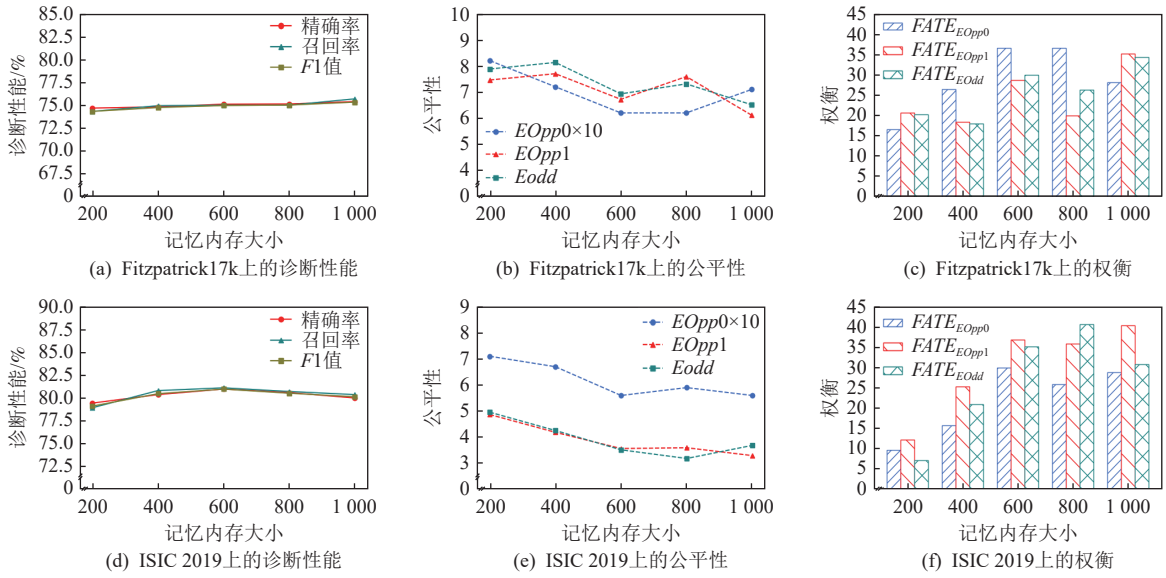
如图 5 所示，随着内存的持续增大，模型的诊断性能保持相对稳定。这种稳定性可能归因于蒸馏损失对灾难性遗忘问题的有效缓解。在公平性方面，随着内存增大，3 项指标的值均呈下降趋势。当内存从 600 增至 1 000 时，ISIC 2019 数据集的公平性指标出现轻微波动。内存的持续增大提供了更多旧属性样本，从而为统计均等差异损失提供了更丰富的信息。因此，增大内存有助于提升公平性。在此过程中，诊断性能与公平性的权衡效果是先提升后趋于稳定。这表明适度的内存大小足以支撑公平性与诊断性能的协同优化，过度增大内存虽能维持公平性，但对综合权衡效果的增益有限。

Table 5 Ablation Study Results of Balanced Updating Strategy

表 5 平衡更新策略的消融实验结果

数据集	平衡更新	属性	诊断性能/%			公平性			权衡		
			精确率↑	召回率↑	*F*1 值↑	*EOpp*0↓	*EOpp*1↓	*EOdd*↓	$FATE_{EOpp0}$↑	$FATE_{EOpp1}$↑	$FATE_{EOdd}$↑
Fitzpatrick17k	w/o	深色	76.55	73.85	74.67	1.15	11.50	11.59	−17.17	−21.91	−17.18
		浅色	72.26	70.02	70.78						
		总体	73.39	71.22	71.98						
	w/	深色	78.29	76.95	77.29	0.62	6.72	6.93	36.48	28.66	29.91
		浅色	73.95	74.10	73.94						
		总体	75.10	74.99	74.95						
ISIC 2019	w/o	男性	79.35	79.51	79.34	0.79	6.21	5.57	0.44	−10.39	−3.29
		女性	81.13	81.37	81.03						
		总体	80.04	80.39	80.13						
	w/	男性	80.73	80.81	80.68	0.56	3.57	3.52	29.97	36.80	35.10
		女性	81.26	81.41	81.24						
		总体	80.96	81.09	80.95						

注：“↑”表示该指标值越高越好，“↓”表示该指标值越低越好。“±”前数值为平均值，“±”后数值为标准差。



(a) Fitzpatrick17k上的诊断性能 (b) Fitzpatrick17k上的公平性 (c) Fitzpatrick17k上的权衡

(d) ISIC 2019上的诊断性能 (e) ISIC 2019上的公平性 (f) ISIC 2019上的权衡

Fig. 5 Impact of memory size

图 5 记忆内存大小的影响

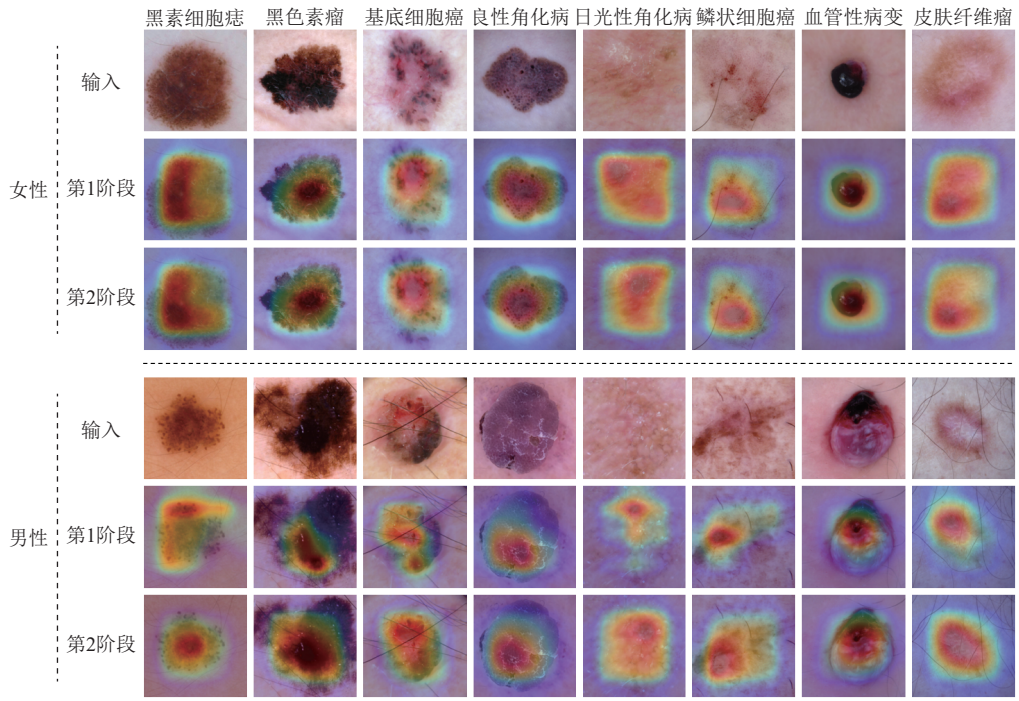
4.3 可视化

增量技术确保特定域不会主导 EBID-Net，并有效地将已学习的特征迁移至后续训练过程中。如图 6 所示，在第 2 阶段结果中，模型对男性和女性受试者的病变区域均实现了精准聚焦，印证了 EBID-Net 的跨域平衡学习能力。对比男性受试者在第 2 阶段与第 1 阶段的注意力可视化结果可知，增量学习后模型对病变区域的关注度明显提升，表明域增量学习机制能够驱动知识从代表性域向非代表性域的有效迁移。

灾难性遗忘是指模型在引入新任务后，其在已学习任务上的性能出现明显下降的现象。对比女性受试者第 2 阶段与第 1 阶段的结果可知，增量学习后模型对病变区域的关注度仅有轻微衰减，且对关键区域的注意力保持稳定。这表明 EBID-Net 有效缓解了灾难性遗忘问题。

5 结 论

本文提出了一个增强型平衡增量蒸馏网络(EBID-Net)以缓解偏见问题。在 2 个皮肤病学数据集上的实验表明，EBID-Net 在提升公平性及权衡方



注：训练顺序为“女性—男性”。

Fig. 6 Attention visualization results of dermatological disease images with different classes on ISIC 2019 dataset

图 6 ISIC 2019 数据集上不同类别皮肤病图像的注意力可视化结果

面具有潜力与有效性。EBID-Net 的另一优势在于，其能够利用在代表性域中所学的知识提升非代表性域的诊断性能，从而改善在涉及某些属性数据缺失的应用中出现性能不佳的情况。未来工作中，我们将进一步优化模型性能，并在多样化临床场景中验证其有效性，从而为构建公平、高效的医疗人工智能系统贡献力量。

作者贡献声明：罗义琴提出了算法思路和实验方案，完成实验并撰写论文；古天龙提出指导意见并修改论文；郝峰锐协助完成论文修改；常亮提出指导意见。

参 考 文 献

[1] Karimkhani C, Dellavalle R P, Coffeng L E, et al. Global skin disease morbidity and mortality: An update from the global burden of disease study 2013[J]. JAMA Dermatology, 2017, 153(5): 406−412

[2] National Engineering Laboratory for Medical Big Data Application Technology (Central South University), Skin Disease Big Data Working Committee, Skin Tumor Research Center, Chinese Society of Dermatology, Chinese Medical Association, Sub-specialty Committee of Skin Tumor, Dermatology Branch, Chinese Medical Doctor Association. Expert consensus on the methodology of dermato-epidemiological investigations[J]. Chinese Journal of Dermatology, 2020, 53(12): 951−961 (in Chinese)
(医疗大数据应用技术国家工程实验室(中南大学)皮肤疾病大数据工作委员会, 中华医学会皮肤性病学分会皮肤肿瘤研究中心, 中国医师协会皮肤科医师分会皮肤肿瘤亚专业委员会. 皮肤病流行病学研究专家共识[J]. 中华皮肤科杂志, 2020, 53(12): 951−961)

[3] Chen Yucong, Dong Xiaohua, Liang Zemin, et al. Research and design of intelligent auxiliary diagnosis and treatment system for skin diseases[J]. Chinese Journal of Health Informatics and Management, 2025, 22(3): 443−449 (in Chinese)
(陈宇聪, 董晓华, 梁泽民, 等. 皮肤病智能辅助诊疗系统的研究与设计[J]. 中国卫生信息管理杂志, 2025, 22(3): 443−449)

[4] Geng Leilei, Xu Zhenxi, Ji Zexuan, et al. Skin lesion segmentation based on topological prior and dual-branch vision Mamba[J]. Journal of Computer Research and Development, 2026, 63(5): 1290−1304 (in Chinese)
(耿蕾蕾, 许振熙, 纪则轩, 等. 基于拓扑先验和双分支视觉 Mamba 的皮肤病变分割[J]. 计算机研究与发展, 2026, 63(5): 1290−1304)

[5] Liu Yuan, Jain A, Eng C, et al. A deep learning system for differential diagnosis of skin diseases[J]. Nature Medicine, 2020, 26(6): 900−908

[6] Papachristou P, Söderholm M, Pallon J, et al. Evaluation of an artificial intelligence-based decision support for the detection of cutaneous melanoma in primary care: A prospective real-life clinical trial[J]. British Journal of Dermatology, 2024, 191(1): 125−133

[7] Huang Kai, Wu Xian, Li Yixin, et al. Artificial intelligence–based psoriasis severity assessment: Real-world study and application[J]. Journal of Medical Internet Research, 2023, 25: e44932

[8] Lin Mingquan, Li Tianhao, Yang Yifan, et al. Improving model fairness in image-based computer-aided diagnosis[J]. Nature Communications, 2023, 14(1): 6261

[9] Seth P, Pai A K. Does the fairness of your pre-training hold up? Examining the influence of pre-training techniques on skin tone bias in skin lesion classification[C]//Proc of the 13th IEEE/CVF Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2024: 570−577

[10] Fan Zhuoya, Meng Xiaofeng. Algorithmic fairness and fairness computing[J]. Journal of Computer Research and Development, 2023, 60(9): 2048−2066 (in Chinese)
(范卓娅, 孟小峰. 算法公平与公平计算[J]. 计算机研究与发展, 2023, 60(9): 2048−2066)

[11] Zong Yongshuo, Yang Yongxin, Hospedales T. MEDFAIR: Benchmarking fairness for medical imaging[J]. arXiv preprint, arXiv: 2210.01725, 2022

[12] Burgon A, Zhang Yuhang, Petrick N, et al. Bias amplification to facilitate the systematic evaluation of bias mitigation methods[J]. IEEE Journal of Biomedical and Health Informatics, 2025, 29(2): 1444−1454

[13] Abdulqawy A, Sallam E, Elkholy A. UnBias: Unveiling bias implications in deep learning models for healthcare applications[J]. IEEE Journal of Biomedical and Health Informatics, 2025, 29(2): 1464−1471

[14] Mehrabi N, Morstatter F, Saxena N, et al. A survey on bias and fairness in machine learning[J]. ACM Computing Surveys, 2021, 54(6): 1−35

[15] Churamani N, Kara O, Gunes H. Domain-incremental continual learning for mitigating bias in facial expression and action unit recognition[J]. IEEE Transactions on Affective Computing, 2022, 14(4): 3191−3206

[16] Mccloskey M, Cohen N J. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem[M]//Psychology of Learning and Motivation. Amsterdam: Academic, 1989: 109−165

[17] Xu Zikang, Zhao Shang, Quan Quan, et al. FairAdaBN: Mitigating unfairness with adaptive batch normalization and its application to dermatological disease classification[C]//Proc of the 26th Int Conf on Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2023: 307−317

[18] Li Yehao, Yao Ting, Pan Yingwei, et al. Contextual Transformer networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(2): 1489−1500

[19] Tang Xiaoya, Zhang Bodong, Ho M M, et al. Duoformer: Leveraging hierarchical representations by local and global attention vision transformer[J]. arXiv preprint, arXiv: 2506.12982, 2025

[20] Chowdary G J, Yin Zhaozheng. Med-former: A Transformer based architecture for medical image classification[C]//Proc of the 27th Int Conf on Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2024: 448−457

[21] Van de Ven G M, Tuytelaars T, Tolias A S. Three types of incremental learning[J]. Nature Machine Intelligence, 2022, 4(12): 1185−1197

[22] Han Yanan, Liu Jianwei, Luo Xionglin. Research progress of continual learning[J]. Journal of Computer Research and Development, 2022, 59(6): 1213−1239 (in Chinese)
(韩亚楠, 刘建伟, 罗雄麟. 连续学习研究进展[J]. 计算机研究与发展, 2022, 59(6): 1213−1239)

[23] Smith J S, Karlinsky L, Gutta V, et al. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning[C]//Proc of the 36th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 11909−11919

[24] Shen Xiaoyu, Zhang Tao, Wu Hao, et al. An adaptive domain-incremental framework with knowledge replay and domain alignment for specific emitter identification[J]. IEEE Transactions on Information Forensics and Security, 2025, 20: 3519−3533

[25] Chen Jinpeng, Cong R, Luo Yuxuan, et al. Replay without saving: Prototype derivation and distribution rebalance for class-incremental semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(6): 4699−4716

[26] Weng Zhi, Lu Yongzheng, Zheng Zhiqiang, et al. Few-shot class-incremental learning for classification and object detection: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(4): 2924−2945

[27] Gu Tianlong, Li Long, Chang Liang, et al. Fair machine learning: Concept, analysis, and design[J]. Chinese Journal of Computers, 2022, 45(5): 1018−1051 (in Chinese)
(古天龙, 李龙, 常亮, 等. 公平机器学习: 概念, 分析与设计[J]. 计算机学报, 2022, 45(5): 1018−1051)

[28] Chen Jinyin, Chen Yipeng, Chen Yiming, et al. Fairness research on deep learning[J]. Journal of Computer Research and Development, 2021, 58(2): 264−280 (in Chinese)
(陈晋音, 陈奕芃, 陈一鸣, 等. 面向深度学习的公平性研究综述[J]. 计算机研究与发展, 2021, 58(2): 264−280)

[29] Seyyed-Kalantari L, Liu Guanxiong, McDermott M, et al. CheXclusion: Fairness gaps in deep chest X-ray classifiers[C]//Proc of the Pacific Symp (BIOCOMPUTING 2021). Singapore: World Scientific, 2020: 232−243

[30] Joshi N, Burlina P. AI fairness via domain adaptation[J]. arXiv preprint, arXiv: 2104.01109, 2021

[31] Yao Ruichen, Cui Ziteng, Li Xiaoxiao, et al. Improving fairness in image classification via sketching[J]. arXiv preprint, arXiv: 2211.00168, 2022

[32] Puyol-Antón E, Ruijsink B, Piechnik S K, et al. Fairness in cardiac MR image analysis: An investigation of bias due to data imbalance in deep learning based segmentation[C]//Proc of the 24th Int Conf on Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2021: 413−423

[33] Yao Siqiong, Dai Fang, Sun Peng, et al. Enhancing the fairness of AI prediction models by Quasi-Pareto improvement among heterogeneous thyroid nodule population[J]. Nature Communications, 2024, 15(1): 1958

[34] Wu Yawen, Zeng Dewen, Xu Xiaowei, et al. Fairprune: Achieving fairness through pruning for dermatological disease diagnosis[C]//Proc of the 25th Int Conf on Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2022: 743−753

[35] Kong Qingpeng, Chiu Chinghao, Zeng Dewen, et al. Achieving fairness through channel pruning for dermatological disease diagnosis[C]//Proc of the 27th Int Conf on Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2024: 24−34

[36] Jin Ruinan, Deng Wenlong, Chen Minghui, et al. Debiased noise editing on foundation models for fair medical image

classification[C]//Proc of the 27th Int Conf on Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2024: 164−174

[37] Liu Ying, Lei Yanbo, Fan Jiulun, et al. Survey on image classification technology based on small sample learning[J]. Acta Automatica Sinica, 2021, 47(2): 297−315 (in Chinese)
(刘颖, 雷研博, 范九伦, 等. 基于小样本学习的图像分类技术综述[J]. 自动化学报, 2021, 47(2): 297−315)

[38] Castro F M, Marín-Jiménez M J, Guil N, et al. End-to-end incremental learning[C]//Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 233−248

[39] Zhou Dawei, Wang Fuyun, Ye Hanjia, et al. Deep learning for class-incremental learning: A survey[J]. Chinese Journal of Computers, 2023, 46(8): 1577−1605 (in Chinese)
(周大蔚, 汪福运, 叶翰嘉, 等. 基于深度学习的类别增量学习算法综述[J]. 计算机学报, 2023, 46(8): 1577−1605)

[40] Zemel R, Wu Y, Swersky K, et al. Learning fair representations[C]//Proc of the 21st Int Conf on Machine Learning. New York: PMLR, 2013: 325−333

[41] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]// Proc of the 29th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 2818−2826

[42] Groh M, Harris C, Daneshjou R, et al. Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm[J]. Proceedings of the ACM on Human-Computer Interaction, 2022, 6(CSCW2): 1−26

[43] Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions[J]. Scientific Data, 2018, 5(1): 1−9

[44] Codella N C F, Gutman D, Celebi M E, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)[C]//Proc of the 15th IEEE Int Symp on Biomedical Imaging. Piscataway, NJ: IEEE, 2018: 168−172

[45] Combalia M, Codella N C, Rotemberg V, et al. Bcn20000: Dermoscopic lesions in the wild[J]. arXiv preprint, arXiv: 1908.02288, 2019

[46] Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning[J/OL]. Advances in Neural Information Processing Systems, 2016[2026-05-23]. https://proceedings.neurips.cc/paper_files/paper/2016/hash/6a9659feb1216f14f7384ba499518b38-Abstract.html

[47] Sagawa S, Koh P W, Hashimoto T B, et al. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization[J]. arXiv preprint, arXiv: 1911.08731, 2019

[48] Tartaglione E, Barbano C A, Grangetto M. End: Entangling and disentangling deep representations for bias correction[C]//Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 13508−13517

[49] Quadrianto N, Sharmanska V, Thomas O. Discovering fair representations in the data domain[C]//Proc of the 32nd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 8227−8236

[50] Jung S, Lee D, Park T, et al. Fair feature distillation for visual recognition[C]//Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 12115−12124

[51] Chiu Chinghao, Chung Haowei, Chen Yujen, et al. Toward fairness through fair multi-exit framework for dermatological disease diagnosis[C]//Proc of the 26th Int Conf on Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2023: 97−107

[52] Xu Gelei, Duan Yuying, Liu Zheyuan, et al. Incorporating rather than eliminating: Achieving fairness for skin disease diagnosis through group-specific experts[C]//Proc of the 28th Int Conf on Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2025: 284−294



Luo Yiqin, born in 1991. PhD. Student member of CCF. Her main research interests include fair machine learning and trustworthy artificial intelligence.

罗义琴, 1991 年生。博士。CCF 学生会员。主要研究方向为公平机器学习、可信人工智能。



Gu Tianlong, born in 1964. PhD, professor. Senior member of CCF. His main research interests include formal methods, trustworthy artificial intelligence, artificial intelligence safety, and artificial intelligence ethics.

古天龙, 1964 年生。博士, 教授。CCF 高级会员。主要研究方向为形式化方法、可信人工智能、人工智能安全、人工智能伦理。



Hao Fengrui, born in 1995. PhD. Student member of CCF. His main research interests include fair machine learning and trustworthy artificial intelligence.

郝峰锐, 1995 年生。博士。CCF 学生会员。主要研究方向为公平机器学习、可信人工智能。



Chang Liang, born in 1980. PhD, professor. Senior member of CCF. His main research interests include data and knowledge engineering, trustworthy artificial intelligence, artificial intelligence ethics, and data privacy protection.

常　亮, 1980 年生。博士, 教授。CCF 高级会员。主要研究方向为数据与知识工程、可信人工智能、人工智能伦理、数据隐私保护。