计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.010132] [CSTR: 32024.14.csa.010132]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于多模态增强融合与多分支蒸馏的内窥镜异常检测模型①

罗逸飞[1,2], 林清华[3], 陈　健[1,2], 郑文斌[1,2], 李佐勇[4,5]

[1](福建理工大学 电子电气与物理学院, 福州 350118)
[2](福建理工大学 数智系统与装置研究院, 福州 350118)
[3](复旦大学 生物医学工程与技术创新学院, 上海 200433)
[4](闽江学院 计算机与大数据学院, 福州 350121)
[5](闽江学院 福建省信息处理与智能控制重点实验室, 福州 350121)
通信作者: 陈　健, E-mail: jianchen@fjut.edu.cn; 李佐勇, E-mail: fzulzytdq@126.com

摘　要: 内窥镜影像为胃癌的筛查与诊断提供重要依据. 然而, 传统内窥镜检查准确率有限. 为此, 多模态融合异常检测方法被引入内窥镜影像分析, 但仍面临模态偏差与配对数据稀缺等问题. 针对这些问题, 本文提出一种基于多模态增强融合与多分支蒸馏的内窥镜异常检测模型. 首先, 设计交叉掩码注意力跨模态融合模块, 通过局部特征重建与交叉注意力机制挖掘模态间的潜在关系. 其次, 提出一种多分支跨模态蒸馏架构, 由多模态教师网络和两个独立学生分支组成. 该架构仅教师网络需配对数据训练, 学生分支则完全无需配对数据. 这一设计降低模型对配对数据的依赖并有效缓解模态偏差. 最后, 引入全局余弦相似度损失以增强多模态特征的一致性表示. 在真实公开数据集上进行的大量实验表明, 本文方法在多模态内窥镜异常检测任务中取得领先的性能. 本文的源码将公开在: https://github.com/LuoYifei-xs/CEMD.

关键词: 异常检测; 多模态融合; 内窥镜影像; 知识蒸馏

引用格式: 罗逸飞,林清华,陈健,郑文斌,李佐勇.基于多模态增强融合与多分支蒸馏的内窥镜异常检测模型.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/10132.html

Multimodal Enhanced Fusion and Multi-branch Distillation Based Model for Endoscopic Anomaly Detection

LUO Yi-Fei[1,2], LIN Qing-Hua[3], CHEN Jian[1,2], ZHENG Wen-Bin[1,2], LI Zuo-Yong[4,5]

[1](School of Electronic, Electrical Engineering and Physics, Fujian University of Technology, Fuzhou 350118, China)
[2](Institute for Digital Intelligence Systems and Devices, Fujian University of Technology, Fuzhou 350118, China)
[3](College of Biomedical Engineering, Fudan University, Shanghai 200433, China)
[4](School of Computer and Data Science, Minjiang University, Fuzhou 350121, China)
[5](Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou 350121, China)

Abstract: Endoscopic images provide a critical foundation for the screening and diagnosis of gastric cancer. However, the accuracy of traditional endoscopic examinations remains limited. To address this issue, multimodal fusion-based anomaly detection techniques have been applied to endoscopic image analysis. However, they still suffer from modality bias and the scarcity of paired data. To overcome these limitations, this study proposes an endoscopic anomaly detection model integrating multimodal enhanced fusion and multi-branch knowledge distillation. The framework incorporates a cross-masked attention cross-modal fusion module that explores latent inter-modal relationships through local feature

① 基金项目: 国家自然科学基金 (62471207, 61972187); 福建省自然科学基金 (2024J02029, 2023R1050, 2023J011390, 2020J02024); 福建省卫生健康委员会科技重大专项 (2021ZD01004); 福建省医疗大数据工程重点实验室开放项目 (KLKF202301); 福建中医药大学高层次人才研究创业基金 (NX2020005-Talent)
收稿时间: 2025-09-18; 修改时间: 2025-10-09, 2025-11-11; 采用时间: 2025-11-24; csa 在线出版时间: 2026-03-02

reconstruction and cross-attention mechanisms. Furthermore, a multi-branch cross-modal distillation architecture is established, comprising a multimodal teacher network and two independent student branches. This design requires only the teacher network to be trained on paired data while enabling the student branches to learn without any paired data, thus significantly reducing dependency on paired samples and effectively mitigating modality bias. Additionally, a global cosine similarity loss is introduced to enhance consistency in multimodal feature representation. Extensive experiments on public real-world datasets demonstrate that the proposed method achieves competitive performance in multimodal endoscopic anomaly detection tasks. Code will be released at: https://github.com/LuoYifei-xs/CEMD.

Key words: anomaly detection; multimodal fusion; endoscopic image; knowledge distillation

随着医疗技术的飞速发展, 内窥镜检查已逐渐成为现代医院不可或缺的医疗技术手段[1]. 通过高清摄像头, 医生能够直观地观察到胃黏膜的微小变化, 从而及时识别出可能的癌变区域. 然而, 最近的研究表明, 由于医学影像分析工作量大, 经验丰富的内镜医师不可避免地会出现误诊和漏诊[2]. 传统内窥镜检查的检测准确性仅为 69%–79%[3]. 为应对这些挑战, 基于多模态融合的异常检测方法被引入到内窥镜检查中.

白光图像 (white light image, WLI) 和窄带光图像 (narrow band image, NBI) 是当前内镜诊断中最常用且技术成熟的成像模式, 具有一定的临床适用基础. WLI 能够清晰地展现黏膜的整体形态与颜色特征, 常用于胃部等区域的初步筛查与全面评估. NBI 通过增强黏膜表层的微血管和微结构对比度, 在食管等部位的早期病变识别与边界判定中表现出显著优势. 因此, 对这两种模态进行融合, 有助于模型进一步结合宏观形态与微观结构特征, 从而更完整地模拟临床诊断路径, 提升异常检测性能.

最近, 知识蒸馏在基于多模态融合的异常检测方法中得到广泛运用[4–6]. Lu 等人[7]提出一种用于胃镜图像异常检测的多模态融合网络, 旨在通过结合胃镜图像和病理文本信息提高检测准确性. 针对测试阶段病理文本缺失问题, 研究引入仅含胃镜图像的分支, 并借助知识蒸馏技术加以引导. Sun 等人[8]提出一种无监督多模态异常检测方法, 采用轻量级师生网络和带符号距离函数, 分别从 RGB 图像和 3D 点云中学习并融合异常信息. Gu 等人[9]提出多模态反向蒸馏范式, 其中冻结的教师编码器利用暹罗架构从不同模态提取互补视觉特征并融合为蒸馏目标, 可学习的学生解码器则从中学习模态相关的先验知识并交互形成多模态表征. 通过教师网络和学生网络之间的交互, 共同形成用于目标重建的多模态表示.

虽然上述工作通过引入知识蒸馏提升多模态融合异常检测方法的性能, 但该领域的发展仍面临若干根本性挑战. 1) 多模态融合方法中存在严重的模态偏差现象. Zheng 等人[10]指出, 在图文融合任务中, 由于数据分布不均与训练目标设计缺陷, 模型往往过度依赖文本模态而抑制视觉模态. 目前主流的内窥镜成像分别为 WLI 和 NBI: NBI 利用窄带光谱成像特性, 能够呈现高对比度的黏膜表面纹理与微血管结构; WLI 因其宽光谱成像方式, 往往导致组织边界模糊, 对比度相对较低. 这种差异使模型在优化过程中更倾向于 NBI 的图像模式, 导致严重的模态偏差现象. 2) 多模态数据稀缺性及方法对配对数据的依赖, 限制了多模态融合性能. 临床实践中难以获取像素级匹配的 WLI-NBI 图像对, 而大多数多模态方法依赖于配对的多模态数据, 这一数据获取的局限性进一步制约了多模态融合异常检测算法的性能.

为解决上述问题, 本文提出一种基于多模态增强融合与多分支蒸馏的内窥镜无监督异常检测模型 (multimodal enhanced multi-branch distillation, MEMD). 首先, 为缓解内窥镜多模态融合中的模态偏差问题, 提出交叉掩码注意力跨模态融合模块, 通过掩码注意力重建模块细化不同模态特征之间的局部特征, 挖掘出不同模态特征之间的潜在关系, 并通过跨模态交叉注意力融合模块对不同模态的特征进行跨模态融合. 其次, 通过计算融合前后特征的全局余弦相似度, 从而强化多模态特征的一致性表示. 此外, 为应对内镜场景下 WLI-NBI 配对样本稀缺的问题, 引入 WtNGAN[11], 以真实 WLI 为基准生成高质量的配对 NBI, 用于补足训练数据. 最后, 为摆脱对配对多模态数据的依赖, 设计多分支跨模态蒸馏架构, 该架构包括多模态教师网络,

WLI 学生分支和 NBI 学生分支. 多模态教师网络通过配对多模态数据挖掘模态间的互补信息. 在此基础上, 学生分支通过知识蒸馏继承教师的多模态表征, 实现对 WLI 和 NBI 的独立优化. 训练完成后, 学生分支可独立进行推断, 无需依赖配对数据集. 该架构有效缓解了模态偏差并降低了对配对数据的依赖, 从而显著增强了模型在真实临床场景中的适用性. 本文在公开的 Kvasir V2[12]肠胃医学图像数据集对模型进行评估, 实验结果表明, 与主流的异常检测方法相比, MEMD 的性能具有显著优势. 本文的主要贡献如下.

(1) 针对多模态融合中的模态偏差问题, 提出交叉掩码注意力跨模态融合模块, 使用掩码注意力重建模块和跨模态交叉注意力融合模块分别对多模态特征进行增强和融合, 并使用全局余弦相似度加强多模态特征的一致性表示.

(2) 为减轻多模态模型对配对数据的依赖, 构建一种多分支跨模态蒸馏架构, 通过从多模态教师网络中蒸馏知识至单模态学生分支, 显著降低对配对数据的依赖, 有效缓解模态偏差, 提升模型在临床真实场景中的适用性.

(3) 在公开的肠胃医学图像数据集 Kvasir V2 上进行的广泛实验表明, 本文所提方法相较于其他异常检测方法具有更优的异常检测性能.

1 相关工作

作为异常检测领域的重要研究方向, 多模态异常检测近年来发展迅速. 其核心目标在于融合不同模态的信息, 通过跨模态信息互补解决传统单模态检测的盲区. 相较于单模态方法, 多模态异常检测不仅需要提取单模态特征, 还需获取有效的多模态融合特征表示, 增加了任务的难度和复杂性. 本文将从 3 个方面阐述相关领域的已有工作.

1.1 异常检测方法

现有的异常检测方法主要侧重于无监督的异常检测. 无监督异常检测方法仅通过正常图像训练, 学习正常样本特征分布. 异常样本因分布差异而导致异常偏差, 该误差可作为衡量异常程度的指标.

例如, PatchCL-AE[13]使用基于补丁对比学习的自动编码器模型进行异常检测, 该模型利用对比学习来加强输入-输出补丁之间的语义一致性, 并通过输入图像与输出图像之间的特征差异定位异常. Hetero-AE[14]采用混合 CNN-Transformer 网络作为解码器, 结合基于卷积神经网络的编码器, 形成异构结构. 此外, 通过多尺度稀疏 Transformer 模块处理局部信息和区域信息之间的关联关系. 这一设计有效地提升了模型学习正常数据的内在信息的能力, 并扩大异常样本的差异. Cai 等人[15]利用单类半监督学习, 通过已知的正常图像和未标记图像进行训练, 并基于此提出异常检测的双分布差异.

1.2 多模态融合方法

多模态融合方法在近年来受到广泛关注. 早期多模态融合方法主要集中在简单的特征级融合, 即把不同模态的数据进行拼接后输入传统机器学习模型, 这种方式虽能初步整合多模态信息, 但缺乏对模态间深层次关联的挖掘.

最近, Zhang 等人[16]提出一种更具鲁棒性的 3D 医学图像多模态融合框架用于医学图像分割, 该框架通过共享特征融合模块探索视觉和语义的一致性, 并耦合优化视觉融合损失与病变分割损失, 使视觉相关和语义相关特征相辅相成, 共同提升模型准确性. Zhu 等人[17]针对通用多模态图像融合面临的跨任务差异问题, 提出任务定制混合适配器架构. 该架构以多模态图像作为输入, 通过动态路由适配不同下游任务, 实现一次训练、多任务部署的通用融合. Li 等人[18]提出一种利用光学相干断层扫描和眼底图像模态信息融合的异常诊断模型. 该模型引入具有任务特定令牌的模态共享解码器, 并利用低级提示调谐对多模态信息进行引导, 实现对视网膜动脉阻塞疾病的分割和检测.

1.3 知识蒸馏

知识蒸馏最先由 Hinton 等人[19]提出, 该方法通过训练一个轻量化的学生模型, 使其模拟复杂教师模型的输出分布, 从而实现模型压缩与知识迁移. Romero 等人[20]在此基础上进行改进, 引入中间层蒸馏的概念, 通过传递教师模型的中间层特征来指导学生模型的学习, 进一步提升知识蒸馏的效果. Hsieh 等人[21]提出逐步蒸馏的概念, 利用大模型的推理过程来训练小模型, 使后者在特定任务上反超大模型. Deng 等人[22]首次提出反向蒸馏概念, 设计一种由教师编码器和学生解码器组成的新型师生模型. 学生网络以教师网络的单类嵌入为输入, 重建教师网络的多尺度表示, 从而避免直接处理高维原始图像. Liu 等人[23]在反向蒸馏的基础上进一步改进师生网络, 通过同时蒸馏编码器与解码器, 增强学生网络生成正常特征的性能, 优化教师网络对

正常和异常特征的判别边界.

2 方 法

本文针对内窥镜图像提出一种多模态增强融合多分支蒸馏无监督异常检测模型 MEMD, 其模型架构如图 1 所示. 模型主要由多模态教师网络、WLI 学生分支和 NBI 学生分支组成. 多模态教师网络主要由编码器和交叉掩码注意力跨模态融合模块组成, 用于模态间的互补信息交互, 并获得多模态融合特征. WLI 学生分支和 NBI 学生分支分别负责学习 WLI 和 NBI 的单模态特征, 同时通过跨模态知识蒸馏从教师网络中获取多模态融合知识.

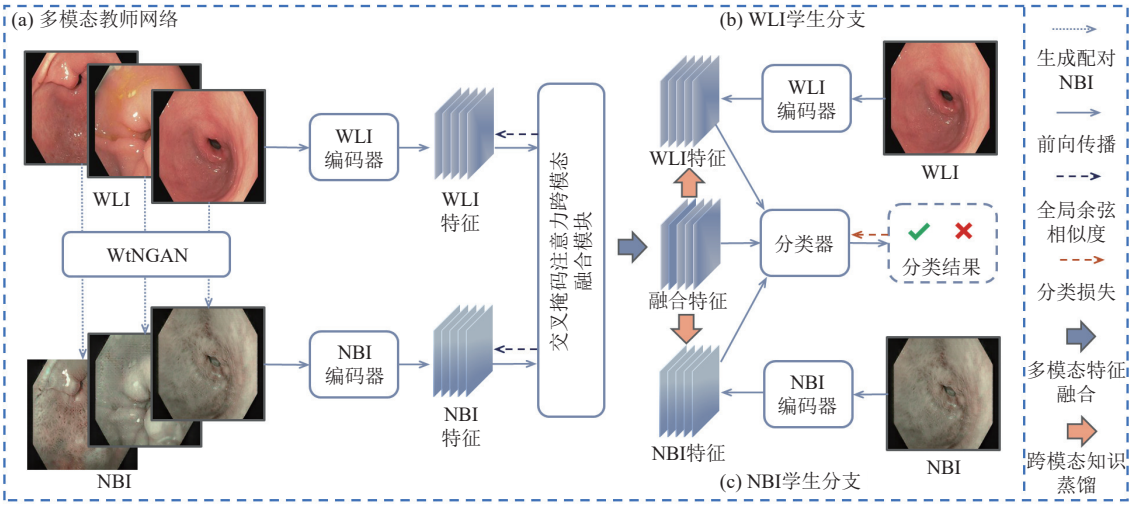


图 1 MEMD 模型图

2.1 问题定义

对于一组多模态数据样本D, 每个样本$x \in D$包含配对的 WLI (x_{WLI}) 和 NBI (x_{NBI}). 其中, x_{WLI}是真实的白光内窥镜图像, x_{NBI}是将x_{WLI}输入 WtNGAN 中生成的配对窄带光内窥镜图像. 多模态数据样本D划分为训练数据集D_{train}、验证数据集D_{val}、测试数据集D_{test}. 本文的多模态无监督异常检测任务定义如下, 多模态教师网络以D_{train}作为训练数据集进行训练, 学习跨模态的特征一致性表示, 并使用D_{val}进行验证. WLI 学生分支和 NBI 学生分支通过知识蒸馏的方式从多模态教师网络中获取跨模态的融合特征, 最终在测试数据集D_{test}上评估模型对于异常检测的性能表现.

2.2 交叉掩码注意力跨模态融合模块

如图 2 所示, 为确保多模态数据的精准对齐, WLI 图像通过 WtNGAN 模型生成严格像素对齐的 NBI 图像, 进而利用预训练的 ResNet50[24]模型进行高效的特征提取. 为使多模态教师网络深入挖掘 WLI 与 NBI 模态间的一致性表示并缓解模态偏差, 引入交叉掩码注意力跨模态融合模块, 在特征空间层面显式建模 WLI 和 NBI 模态间的相关性, 从而促进跨模态特征的一致性表示学习. 具体而言, 首先利用掩码注意力重建模块对单模态的局部特征进行掩码卷积增强并通过注意力机制提取关键特征. 受 Cross-attention[25]启发, 跨模态交叉注意力融合模块提出一种双向查询-键值 (Q-KV) 注意力机制, 用于构建模态间的特征交互, 其中查询向量来自 WLI 模态, 键值对来自 NBI 模态, 这种非对称的注意力设计既保留模态特异性, 又实现跨模态特征的深度融合. 接下来将详细描述各组件.

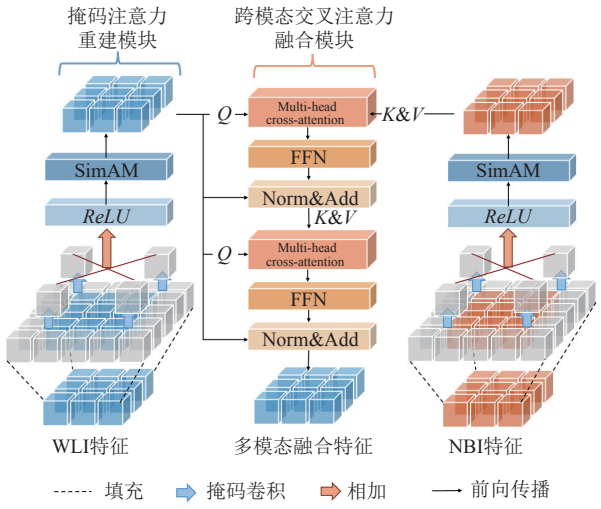


图 2 交叉掩码注意力跨模态融合模块图

2.2.1 掩码注意力重建模块

受扩张卷积[26]的启发, 对给定的单模态特征$x \in \mathbf{R}^{C\times H\times W}$, 选取特征中心点为原点$M \in \mathbf{R}^{1\times 1\times C}$, 在 4 个方向上进行 0 填充, 其填充位置距离原点M特定距离d, 具体实现如下:

$$x_p = pad(x) \in \mathbf{R}^{C\times(H+2d)\times(W+2d)} \tag{1}$$

其中, $pad(\cdot)$表示填充, x_p表示经过填充后的特征.

接着, 在特征的 4 个边角处设置可学习的掩码卷积核$K_i \in \mathbf{R}^{k\times k\times C}, \forall i \in \{1,2,3,4\}$, k代表卷积核大小. 这 4 个可学习的掩码卷积核用于捕捉掩码区域的上下文信息, 并预测掩码中心的特征信息. 卷积核的步长设置为$L = k + 2d + 1$, 将 4 个卷积核的结果融合后, 通过 *ReLU* 函数进行非线性激活. 具体实现如下:

$$x_{\text{mask}} = ReLU\left(\sum_{i=4}^{4} Conv_i(x_p)\right) \tag{2}$$

其中, x_{mask}表示经掩码卷积融合后的结果.

掩码卷积通过对输入特征施加空间约束, 有效增强局部区域的特征表征能力. 然而, 基于固定感受野的卷积操作限制其捕捉远距离特征关系的能力. 为此, 引入注意力机制, 通过动态权重分配策略弥补这一缺陷. 现有的注意力模块仅对特征的空间或通道维度进行细化, 这限制了注意力权重的灵活性, 并且忽略空间和通道之间的可变性. 因此, 引入 SimAM[27]对掩码输出进行自适应重新校准. SimAM 模块无需额外参数即可直接为每个神经元推导出独特的权重, 进而优化神经元功能. 具体实现如下:

$$E = \frac{(x_{\text{mask}} - \mu)^2}{4(\upsilon^2 + \varepsilon)} + \frac{1}{2} \tag{3}$$

其中, E为注意力权重, μ为空间均值, υ^2为空间方差, ε为超参数, 这里ε设置为1×10^{-4}.

最后, 使用 *Sigmoid* 函数对权重进行压缩并得到最终的特征图, 具体实现如下:

$$x_{\text{out}} = Sigmoid\left(\frac{1}{E}\right) \odot x_{\text{mask}} \tag{4}$$

其中, $\odot$表示逐元乘法.

2.2.2 跨模态交叉注意力融合模块

特征在经过掩码注意力重建后, 引入跨模态交叉注意力融合模块实现多模态特征的深度交互与融合. 具体而言, 跨模态交叉注意力融合模块采用双向注意力架构, 包含 2 个注意力层、8 个注意力头, 隐藏层维度为 512, 使用残差连接和 BatchNorm 归一化, 通过查询-键-值映射实现双模态特征的深度交互. 其中, WLI 模态的特征序列作为查询 (Q), 同时将 NBI 模态的特征序列作为键 (K) 和值 (V). WLI 模态具有丰富的纹理信息, 而 NBI 模态在血管结构上具有更高的对比度. 以 WLI 作为查询, NBI 作为键值, 能让模型自适应地融合 WLI 的全局上下文信息, 同时保留 NBI 的血管特征. 具体实现如下:

$$Q_{\text{W}} = F^{Q}(x_{\text{WLI}}), K_{\text{N}} = F^{K}(x_{\text{NBI}}), V_{\text{N}} = F^{V}(x_{\text{NBI}}) \tag{5}$$

$$x_{\text{attn}} = MHCA(Q_{\text{W}}, K_{\text{N}}, V_{\text{N}}) = Softmax\left(\frac{Q_{\text{W}}K_{\text{N}}^{\text{T}}}{\sqrt{C}}\right)V_{\text{N}} \tag{6}$$

$$x_{\text{fusion}} = ReLU(Norm(x_{\text{attn}})) + x_{\text{WLI}} \tag{7}$$

其中, x_{WLI}和x_{NBI}表示 WLI 和 NBI 图像特征通过掩码注意力重建模块重建后的特征, $MHCA(\cdot)$为多头交叉注意力, $F(\cdot)$表示可学习投影矩阵, $Norm(\cdot)$表示标准化.

2.3 多分支跨模态蒸馏架构

传统的多模态蒸馏方法存在严重的模态偏差问题, WLI 因宽光谱成像导致组织边界模糊, 对比度相对较低, NBI 利用窄带光谱成像, 能清晰呈现高对比度的黏膜纹理与微血管结构, 在多模态蒸馏中占主导地位. 同时, 传统多模态方法通常高度依赖像素级对齐的多模态数据, 这一要求严重限制了其在临床实际中的应用.

针对上述问题, 本文设计多分支跨模态知识蒸馏架构. 该架构包含多模态教师网络、WLI 学生分支和 NBI 学生分支. 多模态教师网络以 WtNGAN[11]模型生成的 WLI-NBI 配对图像为输入, 利用交叉掩码注意力跨模态融合模块学习具有丰富语义层次的多模态特征表示, 为后续的跨模态知识迁移奠定基础. 与此同时, 两个独立的单模态学生分支分别处理 WLI 和 NBI, 为使学生分支的输出分布精准对齐多模态教师网络所蕴含的丰富多模态语义, 以 Kullback-Leibler (KL) 散度作为蒸馏损失函数. 针对 WLI 和 NBI 成像复杂度不同的特点, 采用动态蒸馏防止 NBI 特征主导而丢失 WLI 的关键信息. 通过模态感知温度系数τ, 自适应调节知识迁移强度, 抑制 NBI 模态的过度主导倾向. 具体实现如下:

$$L_{\text{KD}} = \sum_{m\in\{\text{WLI, NBI}\}} \tau_m^2 KL\left(\sigma(z_m^{\text{s}}/\tau_m) \| \sigma\left(z^{\text{t}}/\tau_m\right)\right) \tag{8}$$

其中, z_m^s和z^t分别表示学生网络和教师网络在对应模态下输出的预测结果, $\sigma(\cdot)$为 *Softmax* 函数.

2.4 损失函数

在多模态内窥镜图像的异常检测任务中, 重点关注多模态特征融合, 这对于提高内窥镜图像的病灶检出率至关重要. 为提高多模态特征融合质量, 使用全局余弦相似度进行优化, 具体实现如下:

$$L_{\cos}=1-\frac{1}{2}\left(\frac{x_{\mathrm{WLI}}\cdot x_{\mathrm{fusion}}}{\|x_{\mathrm{WLI}}\|\|x_{\mathrm{fusion}}\|}+\frac{x_{\mathrm{NBI}}\cdot x_{\mathrm{fusion}}}{\|x_{\mathrm{NBI}}\|\|x_{\mathrm{fusion}}\|}\right) \tag{9}$$

其中, x_{WLI}、x_{NBI}和x_{fusion}分别表示 WLI、NBI 及融合特征向量, $\|\cdot\|$表示 L2 范数. 模型的总体损失定义如下:

$$L_{\mathrm{total}}=L_{\mathrm{dis}}+\lambda L_{\mathrm{KD}}+\eta L_{\cos} \tag{10}$$

其中, L_{dis}、L_{KD}和$L_{\cos}$分别表示分类器判别损失、蒸馏损失和全局余弦相似度损失, λ和η为超参数.

3 实验分析

实验分析将从数据集、实验设置、对比方法的介绍、实验评估和消融实验这 5 部分展开.

3.1 数据集

本研究采用公开的 Kvasir V2[12]数据集, 该数据集专用于计算机辅助胃肠道疾病检测, 包含多类内窥镜图像. Kvasir V2 由挪威 Vestre Viken 健康信托基金通过内窥镜设备采集而成, 其数据根据不同的解剖标志与病理表现进行划分. 其中, 解剖标志包括 Z 线、幽门和盲肠等, 病理表现则涵盖食管炎、息肉和溃疡性结肠炎等多种类型. 数据集划分如表 1 所示. 由于原始 Kvasir V2 数据集仅包含白光成像的 WLI, 为构建配对的多模态内窥镜数据, 利用 WtNGAN 模型, 以该数据集中的 WLI 图像为基准, 生成高质量的配对窄带成像的 NBI.

表 1　数据统计表

数据集划分	总数 (单个模态)
训练集 (正常图像)	1 800
测试集 (正常/异常)	600/600
验证集 (正常/异常)	400/400

WtNGAN 是基于 Vision Mamba[28]构建的生成对抗网络模型, 通过其特有的视觉状态空间模块建立长程依赖关系以增强生成图像的细节表现能力, 并结合结构一致性约束与对比学习机制, 在保持输入图像解剖结构不变的前提下实现 WLI 向 NBI 的高质量非配对转换. 模型训练采用初始学习率为 0.000 2 的 Adam 优化器, 设置对抗损失、对比学习损失和结构一致性损失的权重系数分别为 1.0、1.0 和 10.0, 通过 200 轮训练周期优化模型参数. 生成图像的质量通过 SSIM (structural similarity index measure)、CIEDE (CIE 2000 color-difference formula)、FID (Fréchet inception distance) 和 KID (kernel inception distance) 这 4 个指标进行系统评估. 表 2 展示了 WtNGAN 在 NBI 图像生成任务 (WLI 到 NBI 转换) 中的性能表现. 图 3 展示了 WtNGAN 生成的、涵盖不同解剖标志与病理表现的配对 WLI-NBI 图像示例.

表 2　WtNGAN 在 NBI 图像生成任务中的性能表现

指标	结果
SSIM	98.36%
CIEDE	17.10
FID	111.56
KID	1.72

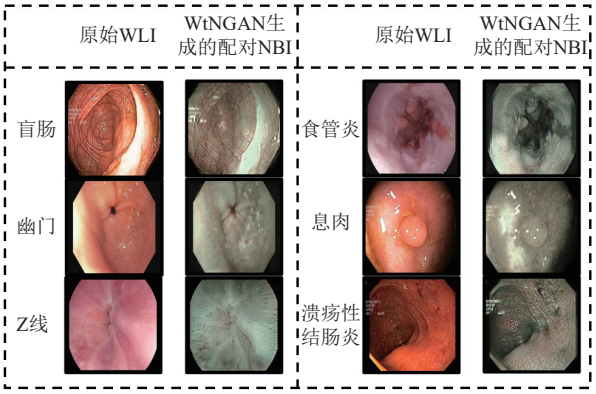


图 3　配对 WLI-NBI 示例图

3.2 实验设置

本文所提出的模型的调试和性能测试的实验条件为 Windows 10 64 位操作系统, NVIDIA GeForce RTX 4060 Ti 显卡, Intel i5-10400 CPU, 深度学习平台为 PyTorch, 采用 3.8 版本. 模型使用 ImageNet 预训练的 ResNet50 作为骨干网络. 模型学习率设置为 0.000 7, 并使用 warmup and decay 策略对学习率进行调整. 此外, 批量大小 (batch size) 和模型迭代次数 (epoch) 分别设置为 24 和 15, 并使用 Adam 优化器优化模型参数. 蒸馏损失中的超参数温度系数设置为 5.

对于模型评估, 本文采用了接收算子曲线下面积 (area under the receiver operator curve, AUROC) 和平均精度 (average precision, AP), 即精确率-召回率曲线下

的面积作为评价指标. 其中, AUROC 反映模型在不同阈值下对正负样本的区分能力, 尤其在类别分布相对均衡时具有较好的参考价值, AP 则反映模型在正样本上的检测精度.

为全面评估 MEMD 模型的多模态性能, 本文所有实验均涵盖了内窥镜下的 WLI 与 NBI 两种模态.

3.3 对比方法

本文将 MEMD 模型与多个先进的多模态和单模态异常检测方法进行对比以证明 MEMD 模型的有效性. 其中, STPM[29]、RD4AD[22]、CMP[30]、SimpleNet[31]、D2UE[32]、EDC[33]、SCRD4AD[34]为单模态异常检测方法, MMRAD[18]为多模态异常检测方法, 对比方法的具体介绍如下.

STPM: 该方法是一种基于师生网络的多尺度层级特征对齐方法. 该方法通过引入增强型监督机制, 实现多尺度特征的同步学习, 从而在不同空间尺度上精准感知并定位异常区域.

RD4AD: 该方法是一种基于师生异构架构的反向蒸馏异常检测方法. 该方法通过单类瓶颈嵌入将高层语义反向还原为多尺度正常特征.

CMP: 该方法是一种自监督异常检测方法. 该方法通过合成异常样本, 在肺部 CT 影像中高效学习样本表征并完成异常定位.

SimpleNet: 该方法是一种基于特征深度嵌入的异常检测方法. 该方法通过异常特征生成器向正常特征注入高斯噪声以模拟异常分布, 并借助二元异常鉴别器进行真伪判别, 实现端到端的异常检测.

D2UE: 该方法是一种基于多元双空间不确定性估计的医学无监督异常检测方法. 该方法通过在潜在空间和图像空间分别建模特征分布的不确定性, 实现对异常区域的检测.

EDC: 该方法是一种基于编码器-解码器对比的医学无监督异常检测方法. 该方法通过构建编码特征与解码重建之间的多尺度对比学习任务, 增强模型对正常模式的结构一致性建模.

SCRD4AD: 该方法是一种基于结构一致性正则化的噪声鲁棒异常检测方法. 该方法通过向训练图像注入可控的 Simplex 噪声构建扰动样本, 并利用多尺度特征间的余弦相似度约束, 在潜在空间强制模型学习噪声不变的结构化表征, 从而增强模型在复杂医学影像中的异常识别与定位能力.

MMRAD: 该方法是一种基于多模态融合与 SAM (segment anything model)[35]模型适配的视网膜动脉阻塞异常检测方法. 该方法通过引入模态共享解码器与任务特定令牌, 使 SAM 支持多模态图像设置, 并结合异常模拟与提示调优策略, 实现无需真实病变样本的异常检测与定位.

3.4 实验评估

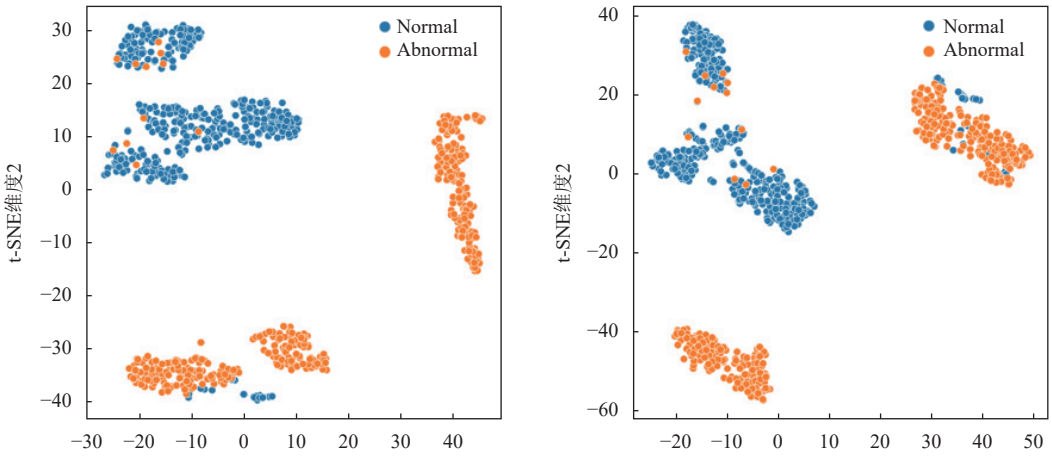
表 3 展示了 MEMD 模型与其他方法的定量实验结果对比. 分别用粗体、红色和绿色标明性能前 3 名. 结果表明, MEMD 模型方法无论在 WLI 模态还是 NBI 模态均取得最优异的性能指标, 该模型不仅通过有效融合多模态特征显著提升不同内窥镜模态下的异常检测性能, 还有效缓解多模态学习中的模态偏差问题, 增强模型在不同成像条件下的泛化能力与稳定性. 这一优势使得 MEMD 在多种内镜影像分析任务中表现出更高的实用性与鲁棒性. 同时, 图 4 中可视化了不同模态下的特征分布. 结果显示, 正常样本与异常样本的特征分布呈现出显著差异, 该结果进一步验证 MEMD 模型能够有效学习具有判别性的特征表示, 从而精准识别异常样本. 此外, 图 5 展示了不同病例与不同模态的胃镜图像异常检测结果可视化对比图, MEMD 模型能够精准地定位异常区域, 为临床诊断提供了有力的支持. 然而, 在溃疡性结肠炎的检测中, MEMD 模型的性能表现欠佳. 这主要是因为溃疡性结肠炎的病变特征较为复杂且多样, 其在胃镜图像中表现为广泛的黏膜充血、水肿、糜烂和溃疡, 这些特征在不同患者之间存在显著差异, 导致模型难以准确识别和区分正常与异常区域.

表 3 定量实验结果 (%)

方法	WLI模态		NBI模态	
	AUROC	AP	AUROC	AP
STPM	70.71	73.52	69.39	70.98
RD4AD	71.80	74.83	73.20	78.26
CMP	77.97	80.09	76.27	73.68
SimpleNet	72.97	72.29	69.23	63.86
D2UE	78.80	78.60	64.36	72.85
EDC	72.29	71.02	73.12	70.01
SCRD4AD	73.88	60.65	73.02	68.33
MMRAD	71.73	68.91	71.73	68.91
MEMD	**80.84**	**83.46**	**81.16**	**82.74**

RD4AD 与 CMP 方法在不同模态的实验中均取得良好的性能表现, 这主要得益于二者在特征提取方面的增强设计. RD4AD 通过反向蒸馏结构实现多尺度正

常特征的重建, CMP 则通过自监督策略强化对判别性特征的学习. 这一结果进一步验证了 MEMD 模型所采用的特征增强策略在提升多模态特征融合表示一致性方面的有效性.



(a) WLI模态的特征分布 (b) NBI模态的特征分布

图 4 MEMD 对不同模态特征分布的可视化

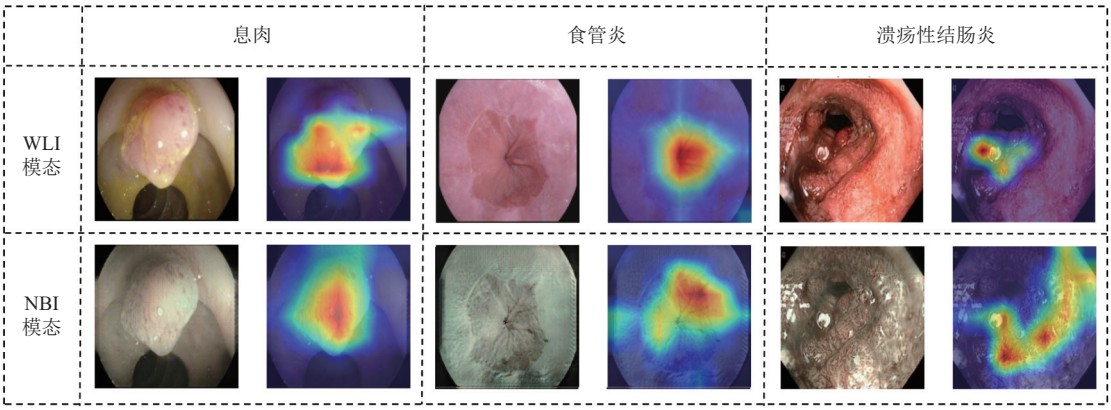


图 5 异常检测结果可视化对比图

MMRAD 方法在实验中的性能表现不佳, 这主要是由于其基础架构依赖于 SAM 模型, 而 SAM 在面对复杂多变的内窥镜图像时表现出一定的泛化能力不足. 值得注意的是, MMRAD 是一种多模态融合方法, 其训练与测试过程均需依赖多模态数据协同进行, 因此该方法在 WLI 与 NBI 两种模态下所得结果保持一致.

3.5 消融实验

为评估本文模型 MEMD 的合理性, 设置多组消融实验验证不同模块的有效性, 最优结果通过粗体标明.

首先, 为验证交叉掩码注意力跨模态融合模块对模型的影响, 分别对其中的掩码注意力重建模块 (MAR) 和跨模态交叉注意力融合模块 (CMCA) 进行消融实验, 并以 ResNet50 基础网络为评价基线, 逐步纳入上述模块以评估贡献, 结果如表 4 所示. 结果表明, MAR 与 CMCA 模块对性能指标均有重要贡献. 其中, MAR 模块通过掩码注意力机制有效提取多模态中的关键特征, 该优化显著增强了 CMCA 模块通过交叉注意力实现的跨模态特征融合效果.

表 4 交叉掩码注意力跨模态融合模块消融实验结果 (%)

模块		WLI模态		NBI模态	
MAR	CMCA	AUROC	AP	AUROC	AP
—	—	66.92	74.52	69.59	73.86
√	—	74.59	80.22	76.48	79.28
—	√	70.14	75.25	74.34	75.49
√	√	**80.84**	**83.46**	**81.16**	**82.74**

其次, 为验证本文提出的多分支跨模态蒸馏架构的有效性, 对其进行消融实验, 结果如表 5 所示.

表 5 多分支跨模态蒸馏架构消融实验结果 (%)

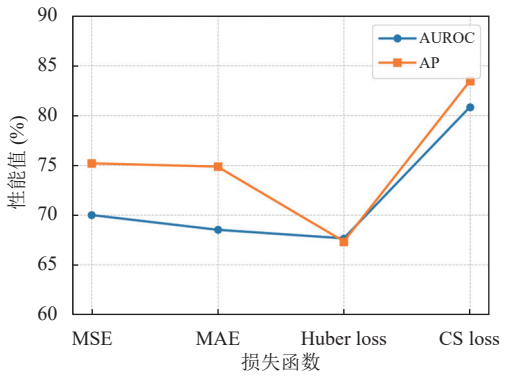
方法	WLI模态		NBI模态	
	AUROC	AP	AUROC	AP
无蒸馏	73.52	78.26	73.52	78.26
单分支蒸馏	74.28	79.45	79.05	80.19
多分支跨模态蒸馏	**80.84**	**83.46**	**81.16**	**82.74**

表 5 中, 无蒸馏方法表示仅使用教师网络进行训练和测试, 由于教师网络使用多模态数据进行训练和测试, 因此, 无蒸馏方法在 WLI 与 NBI 两种模态下所得结果保持一致. 单分支蒸馏指仅使用一个学生分支进行蒸馏. 结果表明, 尽管单分支蒸馏方法能够提升检测精度, 但其存在明显的模态偏差问题. 由于 NBI 模态成像较为简单、图像复杂度较低, 学生模型往往倾向于从该模态中学习, 导致 WLI 与 NBI 模态之间的性能差距显著. 而多分支跨模态蒸馏架构通过引入不同模态的学生分支, 在蒸馏过程中与多模态教师的丰富语义进行精准对齐, 提升指标的同时缓解模态偏差问题. 同时, 表 6 展示了不同模态贡献度的定量实验结果. 在单分支蒸馏中, NBI 模态的分类正确数量显著高于 WLI 模态, 这表明单分支蒸馏存在严重的模态偏差现象. 相比之下, 在多分支跨模态蒸馏中, 两种模态的分类正确数量达到了平衡. 这一结果进一步验证了多分支跨模态蒸馏能够有效缓解模态偏差现象, 提升模型在不同模态下的鲁棒性和准确性.

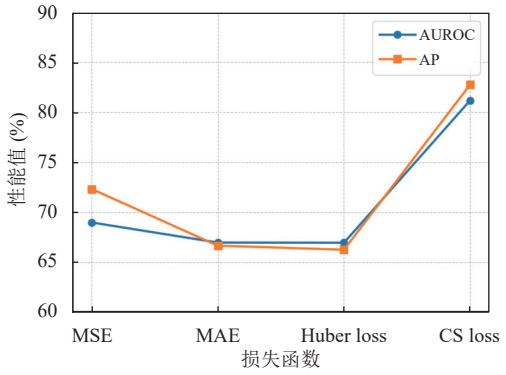
表 6 模态贡献度定量实验结果

分类结果		单分支蒸馏		多分支跨模态蒸馏	
WLI模态	NBI模态	分类正确数量	模态贡献度 (%)	分类正确数量	模态贡献度 (%)
√	—	19	90.92	26	97.21
—	√	56	94.01	18	95.98
√	√	542	—	602	—

最后, 针对本文所使用的全局余弦相似度损失函数 (cosine similarity loss, CS loss), 通过替换不同的损失函数进一步验证其有效性, 选取均方误差 (mean squared error, MSE), 平均绝对误差 (mean absolute error, MAE), Huber 损失 (Huber loss), 结果如图 6 所示. 结果表明, 在 WLI 与 NBI 两种模态下, 全局余弦相似度损失均显著优于传统回归损失 (MSE、MAE、Huber loss), 表明余弦相似度损失更能抑制模态差异, 进一步提高多模态特征融合质量与一致性表示.



(a) WLI模态下不同损失函数性能对比



(b) NBI模态下不同损失函数性能对比

图 6 不同模态下损失函数性能对比图

4 结论与展望

本文针对内窥镜影像异常检测中多模态融合所面临的模态偏差与配对数据稀缺问题, 提出一种基于多模态增强融合与多分支蒸馏的无监督异常检测模型 (MEMD). 通过设计交叉掩码注意力跨模态融合模块, 有效挖掘 WLI 与 NBI 模态间的深层特征关联, 增强多模态特征的一致性表示. 同时, 引入多分支蒸馏架构, 缓解模态偏差并降低模型对配对多模态数据的依赖, 显著提升了模型在临床实际中的适用性. 在公开数据集 Kvasir V2 上的实验结果表明, MEMD 在 WLI 与 NBI 模态下均取得领先的异常检测性能, AUROC 与 AP 指标优于现有主流方法, 验证了模型在特征融合与偏差抑制方面的有效性. 然而, 由于不同品牌内窥镜设备的 WLI 与 NBI 成像存在显著的光谱响应差异, MEMD 在跨设备泛化能力方面仍有待进一步提升.

参考文献

1 陈东. 基于深度学习的胃镜图像分类算法研究 [硕士学位论文]. 南昌: 南昌大学, 2023.

2 温庭栋, 宋文爱, 赵莉, 等. 胃镜下早期胃癌计算机辅助分析研究综述. 计算机工程与应用, 2021, 57(10): 39–47.
3 Choi J, Kim SG, Im JP, *et al*. Comparison of endoscopic ultrasonography and conventional endoscopy for prediction of depth of tumor invasion in early gastric cancer. Endoscopy, 2010, 42(9): 705–713. [doi: 10.1055/s-0030-1255617]
4 孟佳娜, 马腾飞, 赵迪, 等. 基于多模态自适应特征融合的谣言检测. 数据分析与知识发现. https://link.cnki.net/urlid/10.1478.g2.20250911.1356.012, (2025-09-12).
5 杨宇飞, 钱育蓉, 公维军, 等. 大模型优化的 BERT 图文多模态情感分析. 计算机系统应用, 2025, 34(8): 62–69. [doi: 10.15888/j.cnki.csa.009923]
6 段高乐, 王长元, 吴恭朴, 等. 基于多模态数据融合的飞行员注视区域分类. 计算机系统应用, 2024, 33(11): 1–14. [doi: 10.15888/j.cnki.csa.009677]
7 Lu MC, Chai Y, Xu KX, *et al*. Multimodal fusion and knowledge distillation for improved anomaly detection. The Visual Computer, 2025, 41(8): 5311–5322. [doi: 10.1007/s00371-024-03723-6]
8 Sun ZB, Li XL, Li YR, *et al*. Memoryless multimodal anomaly detection via student-teacher network and signed distance learning. Electronics, 2024, 13(19): 3914. [doi: 10.3390/electronics13193914]
9 Gu ZH, Zhang JN, Liu L, *et al*. Rethinking reverse distillation for multi-modal anomaly detection. Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver: AAAI Press, 2024. 8445–8453.
10 Zheng X, Liao CF, Fu YQ, *et al*. MLLMs are deeply affected by modality bias. arXiv:2505.18657, 2025.
11 Lin QH, Li ZY, Zeng K, *et al*. WtNGAN: Unpaired image translation from white light images to narrow-band images. Pattern Recognition, 2025, 162: 111431. [doi: 10.1016/j.patcog.2025.111431]
12 Pogorelov K, Randel K R, Griwodz C, *et al*. KVASIR: A multi-class image dataset for computer aided gastrointestinal disease detection. Proceedings of the 8th ACM on Multimedia Systems Conference. New York: ACM, 2017. 164–169.
13 Lu S, Zhang WH, Guo J, *et al*. PatchCL-AE: Anomaly detection for medical images using patch-wise contrastive learning-based auto-encoder. Computerized Medical Imaging and Graphics, 2024, 114: 102366. [doi: 10.1016/j.compmedimag.2024.102366]
14 Lu S, Zhang WH, Zhao H, *et al*. Anomaly detection for medical images using heterogeneous auto-encoder. IEEE Transactions on Image Processing, 2024, 33: 2770–2782. [doi: 10.1109/TIP.2024.3381435]
15 Cai Y, Chen H, Yang X, *et al*. Dual-distribution discrepancy with self-supervised refinement for anomaly detection in medical images. Medical Image Analysis, 2023, 86: 102794. [doi: 10.1016/j.media.2023.102794]
16 Zhang H, Zuo XH, Zhou HB, *et al*. A robust mutual-reinforcing framework for 3D multi-modal medical image fusion based on visual-semantic consistency. Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver: AAAI Press, 2024. 7087–7095.
17 Zhu PF, Sun Y, Cao B, *et al*. Task-customized mixture of adapters for general image fusion. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 7099–7108.
18 Li JT, Chen T, Wang XY, *et al*. Adapting the segment anything model for multi-modal retinal anomaly detection and localization. Information Fusion, 2025, 113: 102631. [doi: 10.1016/j.inffus.2024.102631]
19 Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science, 2006, 313(5786): 504–507. [doi: 10.1126/science.1127647]
20 Romero A, Ballas N, Kahou S E, *et al*. FitNets: Hints for thin deep nets. Proceedings of the 3rd International Conference on Learning Representations. San Diego: ICLR, 2015. 1–13.
21 Hsieh CY, Li CL, Yeh CK, *et al*. Distilling step-by-step! Outperforming larger language models with less training data and smaller model sizes. Findings of the Association for Computational Linguistics: ACL 2023. Toronto: ACL, 2023. 8003–8017.
22 Deng HQ, Li XY. Anomaly detection via reverse distillation from one-class embedding. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 9727–9736.
23 Liu MX, Jiao YR, Chen H. Skip-ST: Anomaly detection for medical images using student-teacher network with skip connections. Proceedings of the 2023 IEEE International Symposium on Circuits and Systems. Monterey: IEEE, 2023. 1–5.
24 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.
25 Chen CFR, Fan QF, Panda R. CrossViT: Cross-attention multi-scale vision Transformer for image classification. Proceedings of the 2021 IEEE/CVF International Conference

on Computer Vision. Montreal: IEEE, 2021. 347–356.

26 Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. Proceedings of the 4th International Conference on Learning Representations. San Juan: ICLR, 2016. 1–13.

27 Yang LX, Zhang RY, Li LD, *et al*. SimAM: A simple, parameter-free attention module for convolutional neural networks. Proceedings of the 38th International Conference on Machine Learning. Vienna: PMLR, 2021. 11863–11874.

28 Zhu LH, Liao BC, Zhang Q, *et al*. Vision Mamba: Efficient visual representation learning with bidirectional state space model. Proceedings of the 41st International Conference on Machine Learning. Vienna: ICML, 2024. 1–14.

29 Wang GD, Han SM, Ding ER, *et al*. Student-teacher feature pyramid matching for anomaly detection. Proceedings of the 32nd British Machine Vision Conference. BMVC, 2021. 1–14.

30 Li W, Liu GH, Fan HY, *et al*. Self-supervised multi-scale cropping and simple masked attentive predicting for lung CT-scan anomaly detection. IEEE Transactions on Medical Imaging, 2024, 43(1): 594–607. [doi: 10.1109/TMI.2023.3313778]

31 Liu ZK, Zhou YM, Xu YS, *et al*. SimpleNet: A simple network for image anomaly detection and localization. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 20402–20411.

32 Gu Y, Lin Y, Cheng KT, *et al*. Revisiting deep ensemble uncertainty for enhanced medical anomaly detection. Proceedings of the 27th International Conference on Medical Image Computing and Computer Assisted Intervention. Marrakesh: Springer, 2024. 520–530.

33 Li CL, Shi YL, Hu JL, *et al*. Scale-aware contrastive reverse distillation for unsupervised medical anomaly detection. Proceedings of the 13th International Conference on Learning Representations. Singapore: OpenReview.net, 2025. 1–15.

34 Guo J, Lu S, Jia LZ, *et al*. Encoder-decoder contrast for unsupervised anomaly detection in medical images. IEEE Transactions on Medical Imaging, 2024, 43(3): 1102–1112. [doi: 10.1109/TMI.2023.3327720]

35 Kirillov A, Mintun E, Ravi N, *et al*. Segment anything. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023. 3992–4003.

(校对责编: 李慧鑫)