DOI:10.19651/j.cnki.emt.2518245

频域特征和硬负实例筛选的乳腺癌全切片分类*

鲍刘珍[1] 贾 伟[1,2] 赵雪芬[1,2] 孔德凤[3] 江海峰[4]
(1.宁夏大学信息工程学院 银川 750021;2.宁夏"东数西算"人工智能与信息安全重点实验室 银川 750021;
3.宁夏大学新华学院 银川 750021;4.宁夏医科大学总医院病理科 银川 750021)

摘 要:乳腺癌全切片图像分类对精准诊断至关重要,然而,现有基于伪标签的多实例学习方法存在伪标签质量不高和选取硬负实例比例不合理的问题,为解决上述问题,本文提出一种结合频域特征与动态硬负实例筛选的多实例学习方法。首先,设计多尺度频域特征编码模块,通过频域残差连接与跨层特征融合,增强高频细节与复杂纹理表征;其次,提出双分支包预测模块,基于注意力机制动态调整实例权重,缓解异质性导致的特征稀释,优化伪标签生成质量;最后,提出动态硬负实例伪标签挖掘策略,通过渐进式增加硬负实例比例以提升模型获取区分性特征的能力。实验结果显示,在 Camelyon 和 TCGA-BRCA 数据集上,ACC、AUC、Precision、Recall 分别提升了 3.15%、1.72%、3.06%、2.12%和 2.32%、2.79%、2.22%、2.22%,验证了方法的有效性。

关键词:乳腺癌全切片图像;多尺度频域特征;双分支特征聚合;硬负实例筛选;多实例学习
中图分类号:TP391;TN29 **文献标识码**:A **国家标准学科分类代码**:520.60

Breast cancer whole-slide image classification with frequency domain features and hard negative screening

Bao Liuzhen[1] Jia Wei[1,2] Zhao Xuefen[1,2] Kong Defeng[3] Jiang Haifeng[4]
(1. School of Information Engineering, Ningxia University, Yinchuan 750021, China;
2. Ningxia Key Laboratory of Artificial Intelligence and Information Security for Channeling Computing Resources from the East to the West, Yinchuan 750021, China; 3. Xinhua College, Ningxia University, Yinchuan 750021, China;
4. Department of Pathology, General Hospital of Ningxia Medical University, Yinchuan 750021, China)

Abstract: Breast cancer whole slide image classification is critical for accurate diagnosis. However, existing pseudo-label-based multiple instance learning methods suffer from low-quality pseudo-labels and suboptimal selection of hard negative instance ratios. To address these issues, this paper proposes a multiple instance learning method combining frequency domain features and dynamic hard negative instance screening. First, a multi-scale frequency domain feature encoding module is designed, which enhances high-frequency details and complex texture representations through frequency domain residual connections and cross-layer feature fusion. Second, a dual-branch bag prediction module is proposed to dynamically adjust instance weights via an attention mechanism, mitigating feature dilution caused by heterogeneity and improving pseudo-label generation quality. Finally, a dynamic hard negative instance pseudo-label mining strategy is introduced, progressively increasing the proportion of hard negative instances to enhance the model's ability to capture discriminative features. Experimental results on the Camelyon and TCGA-BRCA datasets demonstrate significant improvements: ACC, AUC, Precision, and Recall increased by 3.15%、1.72%、3.06%、2.12% and 2.32%、2.79%、2.22%、2.22%, respectively. These advancements validate the effectiveness of the proposed method.

Keywords: breast cancer whole slide image; multi-scale frequency domain feature; dual-branch feature aggregation; hard negative instance screening; multiple-instance learning

0 引 言

乳腺癌是全球女性最常见的恶性肿瘤及癌症死亡主因之一[1]。组织病理学诊断作为乳腺癌诊断的金标准,其图像分类准确性直接影响患者生存率[2]。近年来,数字病理技术推动全切片图像(whole slide image, WSI)成为主流数

收稿日期:2025-03-04
* 基金项目:国家自然科学基金(62062057)、宁夏自然科学基金(2024AAC03063,2024AAC03325)项目资助

字化手段[3]。WSI 以金字塔结构存储多级放大信息,但高达 40 000×40 000 像素的尺寸使深度学习模型难以直接处理[4]。基于补丁的多实例学习(multiple-instance learning,MIL)框架通过切片级标注显著降低了标注成本,成为 WSI 分类的主流方法[5-10]。

现有 MIL 方法分为实例级特征编码与切片级特征聚合两个阶段[11-13]。实例级编码通常采用自监督对比学习预训练模型,通过挖掘数据内在结构提升特征表达能力[14-23]。然而,乳腺癌 WSI 复杂的纹理特征、不均匀色彩分布和多尺度细节,导致现有编码器在高频细节捕捉与跨尺度特征融合上存在不足[14]。聚合阶段的最大池化、平均池化和 Top-K 选择等[24]传统方法易导致关键信息丢失,基于注意力机制的改进方法虽通过权重分配增强关键实例关注,但 WSI 实例间的显著异质性仍导致权重偏差与信息损失[11]。

伪标签微调策略通过聚合模型生成实例级伪标签优化特征编码器,在 WSI 分类中展现出潜力[25-31]。Liu 等[30]通过筛选高置信度正负实例集提升判别能力,但简单负实例引入导致训练效率低下。Huang 等[31]采用硬负实例挖掘优化负样本选择,却因静态策略难以适应不同训练阶段的特征学习需求。

为解决上述问题,本研究提出一种结合频域特征与动态硬负实例筛选的 MIL 方法(frequency-domain and dynamic hard negative screening for multi-instance learning,FDHN-MIL),通过频域特征提取增强多尺度信息捕获能力,并引入动态硬负实例挖掘优化负样本选择策略,从而提升关键特征表达能力与模型区分性能。具体贡献如下:

1)提出了多尺度频域特征编码模块(multi-scale frequency domain feature encoding module,MFFE),该模块包括逐层特征融合模块(layer-wise feature fusion module,LFFM)与频域特征增强模块(frequency domain residual enhancement module,FREM)。LFFM 整合浅层与深层特征,实现多尺度信息融合;FREM 通过傅里叶变换(Fourier transform,FT)动态增强高频特征,提升特征编码模块对细节和纹理的敏感度。

2)提出了双分支包预测模块(dual-branch bag prediction module,DBBP),该模块由关键实例选择分支(key instance selection branch,KISB)和伪标签引导的包特征构建分支(pseudo-label guided bag feature construction branch,PBCB)组成。KISB 通过聚类与 Top-K 选择筛选病理显著性实例;PBCB 基于注意力机制生成伪标签,动态调整权重并融合全局特征构建包级表示,缓解特征稀释问题。

3)设计了动态硬负实例伪标签挖掘策略(dynamic hard negative instance pseudo-label mining strategy,DHNIM),动态选择与正实例具有较高相似度的硬负实例伪标签,然后将伪标签对应的硬负实例应用到 MFFE 微调中,使训练过程逐渐聚焦于更具挑战性的硬负实例,增强模型的判别能力。

1 本文模型

本研究提出的 FDHN-MIL 结构如图 1 所示,首先将 WSI 分割为实例,然后将这些实例输入 MFFE 以生成实例特征。接着,这些实例特征被送入 DBBP 以获得切片级特征,并为每个实例生成一个伪标签。随后,在动态硬负实例筛选中,先根据 DHNIM 逐步增加硬负实例伪标签的比例,然后通过硬负实例伪标签选择对应的硬负实例。最后,通过监督对比学习微调 MFFE。这一过程是反复进行的,通过不断迭代优化模型,以提升分类性能。

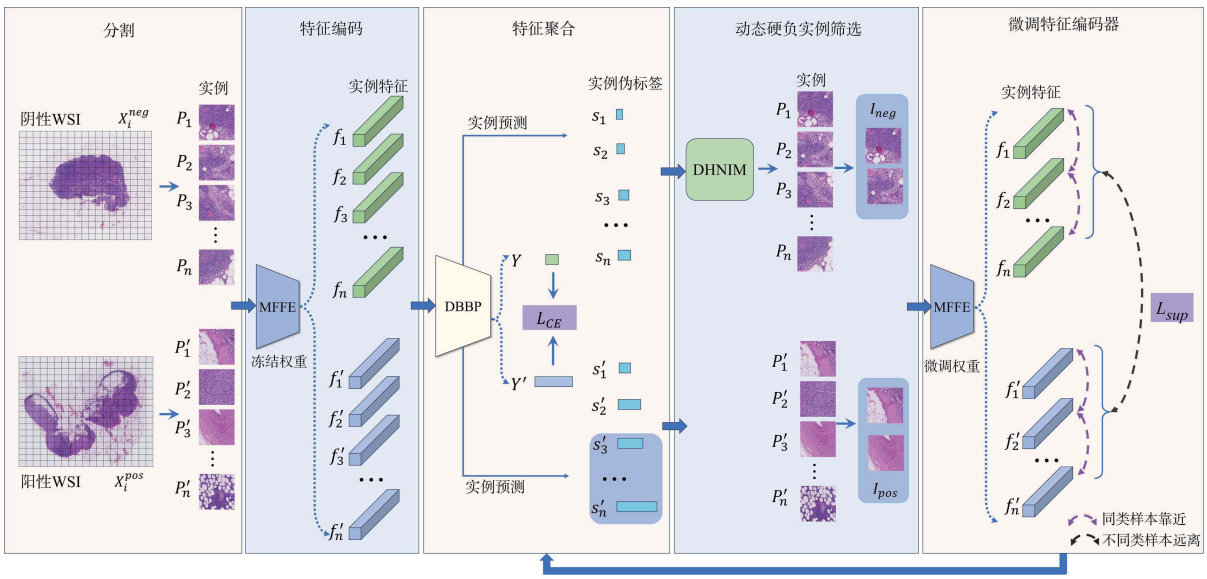


图 1 FDHN-MIL 结构

Fig. 1 FDHN-MIL model architecture

1.1　多尺度频域特征编码模块

乳腺癌 WSI 具有显著的多尺度组织病理特征和高分辨率下的复杂微观结构。现有方法存在跨尺度特征融合不充分导致细节丢失，以及对频域高频成分敏感性不足造成纹理弱化的问题。为了解决这些问题，本文设计了 MFFE。该模块由 LFFM 和 FREM 组成，旨在更有效地融合多尺度信息并增强图像细节，从而提升分类性能。

MFFE 结构如图 2 所示。首先对 WSI 分割后的实例进行卷积处理，并依次经过批量归一化层（batch normalization，BN）、ReLU 激活函数和最大池化层（max pooling，MaxPool）进行初步的特征编码。随后，通过 4 个 ResNet 层提取高级特征。提取的高级特征依次通过 LFFM 和 FREM，以实现多尺度信息的融合和图像细节的增强。最终经过平均池化（average pooling）、展平层（flatten layer）、全连接层（fully connected layer，FC），输出分类特征。

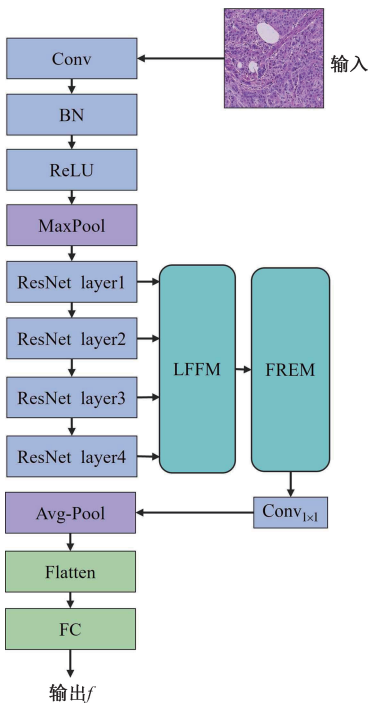


图 2　MFFE 结构

Fig. 2　Structure of the MFFE

1）逐层特征融合模块

在 ResNet 架构中，浅层特征具有较高的空间分辨率，而深层特征则包含更丰富的语义信息。单独依赖浅层或深层特征可能无法充分捕捉图像中的细节与全局信息。为了有效融合来自不同层次的多尺度特征，本文借鉴了 CSFNet 网络[32]设计了 LFFM，以提升模型对细节和全局信息的表达能力。

LFFM 的结构如图 3 所示。首先对深层的低分辨率特征图进行卷积操作来减少通道数，使其与浅层高分辨率特征图通道数一致。然后对处理后的深层特征图上采样，使其空间维度与浅层特征图对齐，最终将二者相加得到融合后的特征图。其操作流程如下：

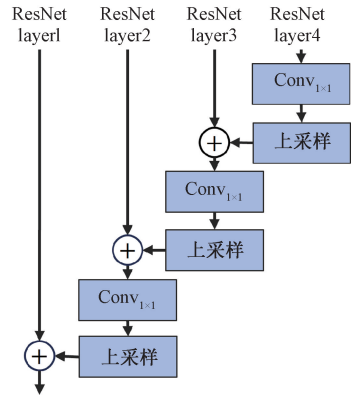


图 3　LFFM 结构

Fig. 3　Structure of the LFFM

假设来自深层的低分辨率特征图为 C_i（i 是残差块编号，$2 \leqslant i \leqslant 4$），来自浅层的高分辨率特征图为 C_{i-1}。逐层特征融合过程如式（1）所示。

$$h_i = C_{i-1} + UpSample(\mathrm{Conv}_{1\times1}(C_i)) \tag{1}$$

式中：$\mathrm{Conv}_{1\times1}$ 表示使用 1×1 卷积操作对 C_i 进行降维，减少通道数，使其与 C_{i-1} 的通道数一致；$UpSample$ 表示对降维后的深层低分辨率特征图进行上采样，使其空间维度维度与 C_{i-1} 对齐。

2）频域残差增强模块

现有的乳腺癌 WSI 分类方法多聚焦于空间域特征编码，却忽略了频域信息。频域可捕捉在空间域易受模糊、噪声干扰的高频细节与微小纹理。引入频域信息能弥补空间域方法不足，增强模型对复杂纹理中细节特征的敏感度，并抑制 WSI 中的噪声。因此，本文设计了 FREM，通过 FT 将空间域特征转换到频域，增强图像的高频部分，以提高对细节和微小纹理的敏感度。

FREM 的结构如图 4 所示。通过 FT 将空间域特征转换到频域，接着应用具有可训练参数的频域滤波器对不同频率成分的强度进行调整和增强，随后再通过逆傅里叶变换（inverse fourier transform，IFT）将增强后的频域特征转换回空间域，并通过残差连接保留原始信息。其操作流程如下：

首先，LFFM 输出的空间域特征 h_2 通过 FT 转换到频域，计算过程如式（2）所示。

$$F(u,w) = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} h_2(\alpha,\beta) \cdot \mathrm{e}^{-\mathrm{j}\cdot 2\pi\left(\frac{u\cdot\alpha}{M}+\frac{w\cdot\beta}{N}\right)} \tag{2}$$

式中：$F(u,v)$ 是经过 FT 得到的特征图的频率成分，u 和 w 是频域中的坐标，代表了不同的频率成分。$h_2(\alpha,\beta)$ 是 LFFM 输出的特征图在空间域的强度分布，α 和 β 分别对应特征图的水平和垂直坐标。M 和 N 分别是特征图在 α 和 β 方向的特征点数。j 是虚数单位，满足 $\mathrm{j}^2 = -1$。

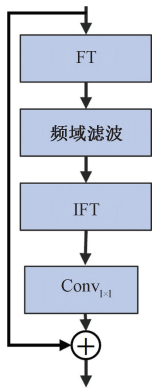


图 4　FREM 结构

Fig. 4　Structure of the FREM

$e^{-j\cdot 2\pi\left(\frac{u\cdot\alpha}{M}+\frac{w\cdot\beta}{N}\right)}$ 是复数权重,用于对特征图中的每个特征点进行加权。通过计算每个特征点的复数权重,并将其与特征点的强度相乘,然后将所有特征点的加权结果进行求和得到 $F(u,w)$。当 u 和 w 值较小时,对应低频成分,反映特征的整体趋势和大致结构;当 u 和 w 值较大时,对应高频成分,体现特征的细节和纹理等精细信息。

接下来,设计了一个具有可训练参数的频域滤波器 $H(u,w)$,用于动态调整不同频率成分的权重,计算过程如式(3)所示。

$$G(u,w)=F(u,w)\odot H(u,w) \tag{3}$$

式中:$G(u,w)$ 是经过滤波后的频域特征图。$H(u,w)$ 是可学习的滤波器参数,通过训练过程自动优化,以适应不同图像的频域特征分布。$\odot$ 表示逐点相乘操作,实现对各频率成分的有针对性增强或抑制。

乳腺癌 WSI 的分类主要依靠空间域特征,频域处理旨在强化其细节纹理信息,所以经滤波器调整增强的频域特征需用 IFT 转换回空间域,计算过程如式(4)所示。

$$h'_2(\alpha,\beta)=\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}G(u,w)\cdot e^{j\cdot 2\pi\left(\frac{u\cdot\alpha}{M}+\frac{w\cdot\beta}{N}\right)} \tag{4}$$

式中:$h'_2(\alpha,\beta)$ 是经过 IFT 恢复的空间域特征图。$G(u,w)$ 是经过频域滤波器调整后的频域特征图。$e^{j\cdot 2\pi\left(\frac{u\cdot\alpha}{M}+\frac{w\cdot\beta}{N}\right)}$ 是逆变换中的复数权重。通过计算每个频率成分的复数权重,并将其与对应的频域特征点相乘,然后将所有点的加权结果进行和,得到恢复的空间域特征图 $h'_2(\alpha,\beta)$。

为了保持信息传递的一致性,频域增强后的特征图与原始输入特征图进行残差连接,计算过程如式(5)所示。

$$f_{out}=h_2(\alpha,\beta)+Conv_{1\times1}(h'_2(\alpha,\beta)) \tag{5}$$

通过 1×1 卷积调整原始输入特征图的通道数后,与频域增强后的特征图相加,确保特征维度匹配。

1.2　双分支包预测模块

WSI 中各实例分布存在显著差异,现有的聚合策略[24]未能充分反映这种异质性,导致关键信息丢失,进而影响伪标签的准确性。为了解决这一问题,本文设计了 DBBP,它由 KISB 和 PBCB 构成。通过精细调整不同实例的重要性权重,DBBP 能够更好地捕捉切片中关键实例的特征,避免现有聚合方法中的关键信息丢失。

DBBP 结构如图 5 所示。KISB 筛选出包含肿瘤细胞分布、背景特征等关键病理信息的关键实例。PBCB 基于 KISB 筛选出的关键实例,先计算所有实例与关键实例的相似度,并经归一化生成实例伪标签;随后,利用这些伪标签对信息向量加权求和,构建反映切片整体病理特征的包级特征表示。最终取 KISB 和 PBCB 两个分支分数的平均值作为包级预测分数。

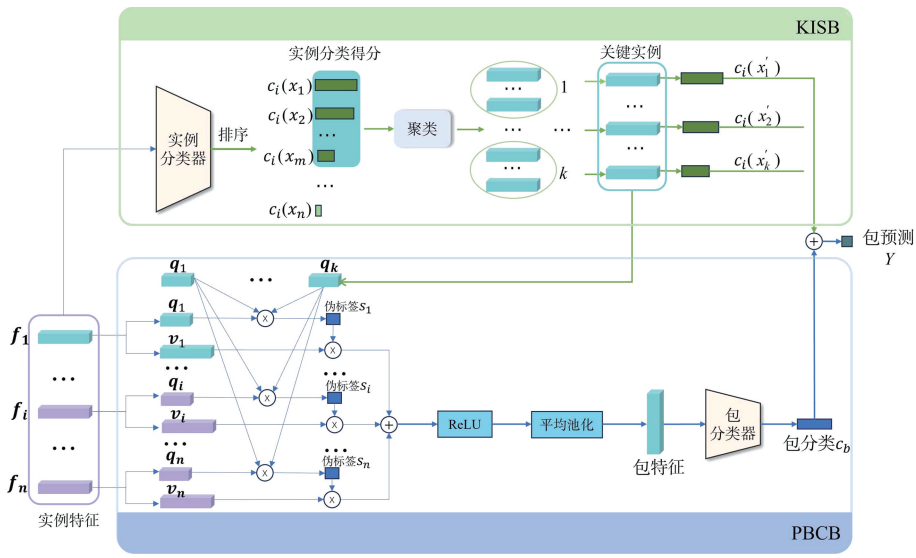


图 5　DBBP 结构

Fig. 5　Structure of the DBBP

通过对乳腺癌的 WSI 分析，发现其呈现出阳性切片仅有少量区域包含肿瘤，且该区域特征对准确分类极为关键的特性。基于此发现，在 KISB 中，所有实例的特征 $\boldsymbol{f}_i$ 经实例分类器计算各实例的分类得分 $\{c_i(x_1),\cdots,c_i(x_n)\}$，接着选择排名前 10% 的实例组成实例集合 $\{x_1,x_2\cdots,x_m\}$。

为了进一步筛选出具有代表性的实例，运用 Mini-Batch K-Means[33]算法，将已选出的实例划分为 k 个聚类，随后把每个聚类的中心视作关键实例 $\{x'_1,x'_2\cdots,x'_k\}$。这些关键实例能够代表图像中不同病理区域所蕴含的关键信息，有效覆盖了肿瘤细胞分布以及背景特征等异质性较强的情况，为后续构建更具代表性的包级特征奠定良好基础。PBCB 旨在将所有实例的特征进行聚合，生成包的特征表示，从而实现包级分类。具体过程如下：

首先，将所有实例的特征 $\boldsymbol{f}_i$ 转换为查询向量 $\boldsymbol{q}_i$ 和信息向量 $\boldsymbol{v}_i$，计算过程如式(6)所示。

$$\begin{cases}\boldsymbol{q}_i=\boldsymbol{W}_q\boldsymbol{f}_i,\ i=0,1,\cdots,n-1\\ \boldsymbol{v}_i=\boldsymbol{W}_v\boldsymbol{f}_i,\ i=0,1,\cdots,n-1\end{cases}\tag{6}$$

式中：$\boldsymbol{W}_q$ 和 $\boldsymbol{W}_v$ 分别是权重矩阵，用于生成查询向量和信息向量。

接着，通过内积计算每个实例与关键实例之间的相似度，并对其进行归一化处理以得到注意力权重，进而作为实例的伪标签。伪标签的计算过程如式(7)所示。

$$s_i=\sum_{j=1}^{k}\frac{\exp(\boldsymbol{q}_i^{\mathrm{T}}\cdot\boldsymbol{q}_j)}{\sum_{i=1}^{n}\exp(\boldsymbol{q}_i^{\mathrm{T}}\cdot\boldsymbol{q}_j)}\tag{7}$$

式中：$\boldsymbol{q}_j$ 是第 j 个关键实例的查询向量，k 是关键实例的数量。

然后，利用注意力权重对信息向量 $\boldsymbol{v}_i$ 进行加权求和，得到包的初步特征表示。随后，通过非线性激活函数 ReLU 引入非线性变换，并使用平均操作对多个关键实例的特征进行聚合。最终通过包级分类器得到包的分类得分。计算过程如式(8)所示。

$$c_b=BClassifier\left(\frac{1}{k}\cdot\left(ReLU\left(\sum_{i=1}^{n_i}s_i\cdot\boldsymbol{v}_i\right)\right)\right)\tag{8}$$

最终包级预测分数取 KISB 和 PBCB 两个分支的分数的平均值。计算过程如式(9)所示。

$$Y=\frac{1}{2}\cdot\left(\frac{1}{k}\cdot(c_i(x'_1)+\cdots+c_i(x'_k))+c_b\right)\tag{9}$$

式中：$\frac{1}{k}\cdot(c_i(x'_1)+\cdots+c_i(x'_k))$ 和 c_b 分别是 KISB 和 PBCB 两个分支的得分。

1.3　动态硬负实例伪标签挖掘策略

为提升模型对关键差异的识别能力，现有方法通常侧重于筛选与正实例相近的硬负实例[31]，然而，在现有方法在硬负实例策略中，硬负实例的比例选取不合理，导致无法获取具有区分性的特征信息，影响了分类性能。为解决该问题，本文提出了 DHNIM，通过动态调整负实例伪标签，实现动态筛选不同分类难度的负实例，从而使微调阶段的 MFFE 逐步聚焦于更具挑战性的硬负实例。

设 I_{pos} 表示正实例集合，I_{neg} 表示硬负实例集合。对于 I_{pos}，采用与现有研究[30-31]相同的方法，将其定义为阈值 η 之上排名靠前的正实例的集合。对于 I_{neg}，先设计 DHNIM 构建硬负实例伪标签集 I'_{neg}，然后通过 I'_{neg} 从阴性实例中挑选出硬负实例伪标签对应的实例，最终形成硬负实例集合 I_{neg}。

在 DHNIM 中，定义了一个单调递减函数 $r_{neg}(t)$，用于训练轮次 $t(t=1,2,\cdots,T$，其中 T 为总的训练次数)中选择硬负实例。初期 $r_{neg}(t)$ 设置较高，使模型从易区分的硬负实例中学习。随着训练的进行，$r_{neg}(t)$ 逐渐降低，促使模型关注更难以区分的硬负实例，逐步提高其判别能力。

DHNIM 如图 6 所示，首先将阴性 WSI 中所有实例的伪标签值 $s(x_i)$ 按从低到高排序，形成有序集合 $\tilde{I}_{neg}=\{s(x''_1),s(x''_2),\cdots,s(x''_{N_{neg}})\}$，其中 N_{neg} 表示阴性实例的总数，并满足 $s(x''_1)\leqslant s(x''_2)\leqslant\cdots\leqslant s(x''_{N_{neg}})$。

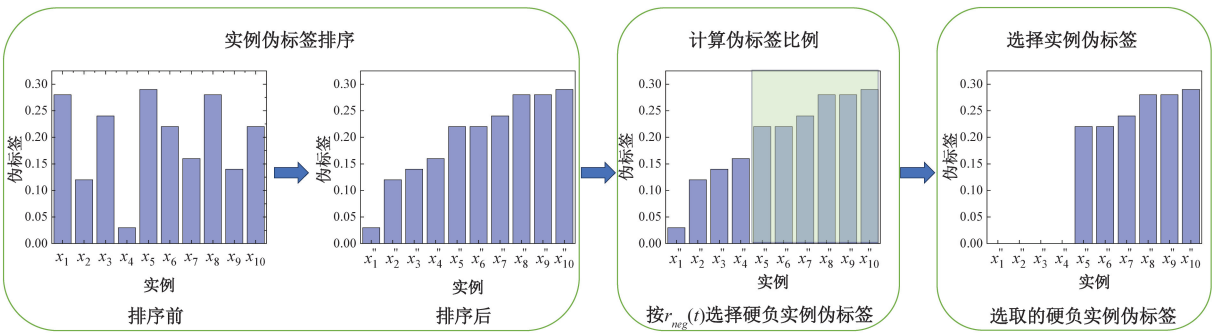


图 6　DHNIM 示意图

Fig. 6　Schematic diagram of the DHNIM

接着，利用硬负实例伪标签选择函数 $r_{neg}(t)$ 作为当前迭代步 t 下的比例参数来选择硬负实例伪标签。

最后，在 $\tilde{I}_{neg}$ 中选取伪标签值最高的 $r_{neg}(t)$ 实例构成硬负实例伪标签集合 I'_{neg}。I'_{neg} 的构建过程如式(10)所示。

$$I'_{neg}=\{s(x''_j)\mid j=N_{neg}-\lfloor r_{neg}(t)\cdot N_{neg}\rfloor+1,\cdots,N_{neg}\} \tag{10}$$

函数 $r_{neg}(t)$ 的形式可以有多种选择,例如凸函数 $r_{neg}(t)=r_e+(r_b-r_e)\cdot\sqrt{1-\frac{t}{T}}$。其中 r_e 是最终比例,r_b 是初始比例,不同的函数形式会导致在不同训练阶段选择不同难度的负实例,进而可能影响模型的训练效果,本文将在实验部分详细探讨这些函数形式在数据集上的表现。

1.4 损失函数

每轮训练 MIL 聚合器后,使用其生成各实例的伪标签,通过缩小同标签实例的表示距离、拉远异标签实例的表示距离来微调 MFFE。

设 γ 为锚点实例,γ_s 是从集合 S_γ 中选取的与 γ 具有相同伪标签的实例,γ_d 是从集合 D_γ 中选取的与 γ 伪标签不同的实例。S_γ 和 D_γ 的构建规则分别如式(11)、(12)所示。

若 $\gamma\in I_{pos}$:

$$\begin{cases}S_\gamma\leftarrow I_{pos}\\ D_\gamma\leftarrow I_{neg}\end{cases} \tag{11}$$

若 $\gamma\in I_{neg}$:

$$\begin{cases}S_\gamma\leftarrow I_{neg}\\ D_\gamma\leftarrow I_{pos}\end{cases} \tag{12}$$

式中:"$\leftarrow$"表示从实例集合中随机采样。如果 γ 是从正实例集合 I_{pos} 中采样得到的,那么 S_γ 和 D_γ 分别从 I_{pos} 和 I_{neg} 中进行采样构建。如果 γ 是从 I_{neg} 中采样得到的,S_γ 和 D_γ 分别从 I_{neg} 和 I_{pos} 中进行采样构建。

损失函数定义如式(13)所示。

$$L_{\sup}(\gamma)=\frac{1}{|S_\gamma|}\sum_{\gamma_s\in S_\gamma}-\log\left(\frac{sim(\gamma,\gamma_s)}{\sum_{\gamma_s\in S_\gamma}sim(\gamma,\gamma_s)+\sum_{\gamma_d\in D_\gamma}sim(\gamma,\gamma_d)}\right) \tag{13}$$

式中:$sim(x_1,x_2)$ 表示实例 x_1 和 x_2 之间的相似度。

总损失函数如式(14)所示。

$$L=L_{CE}+L_{\sup} \tag{14}$$

式中:L_{CE} 是用于分类的交叉熵损失,其定义为:$L_{CE}=-y\cdot\log(p)-(1-y)\cdot\log(1-p)$,其中 y 是真实标签(0 或 1),p 是预测为阳性的概率。

2 实验结果与分析

2.1 实验环境与配置

所有实验均在 NVIDIA GeForce RTX3090 24 GB、Ubuntu 20.04、CUDA 12.2 的服务器上实现。深度学习框架为 PyTorch 1.8.1。

在对特征编码器进行微调时,使用 SGD 优化器,对 MFFE 进行 50 个周期的训练,批次大小设置为 512,学习率设置为 10^{-2}。在训练 MIL 聚合器时,使用 Adam 优化器,批次大小设置为每次一张 WSI 图像,最多训练 350 个周期。初始学习率为 2×10^{-4},并使用 StepLR 调度器每隔 75 个周期将学习率减半。

2.2 数据集

本文使用了 Camelyon 和 TCGA-BRCA 两个公开数据集。Camelyon 数据集整合了 Camelyon16 的 399 张 WSI 和 Camelyon17 的 500 张 WSI,共计 899 张 WSI,覆盖 0.2 mm 至>2 mm 的多尺度转移病理特征,包含无转移、宏观、微观转移及孤立肿瘤细胞 4 种类别。

TCGA-BRCA 包含 1 041 张 WSI,标注含浸润性导管癌和浸润性小叶癌两类乳腺癌亚型。

本文按照 64:16:20 的比例对数据集进行了划分,分别构建训练集、验证集和测试集。具体划分结果如表 1 所示。

表 1 数据集划分情况

Table 1 Dataset division situation

数据集	训练集	验证集	测试集
Camelyon	576	143	180
TCGA-BRCA	666	166	209

2.3 评价指标

在分类任务中,准确率(accuracy,ACC)、曲线下面积(area under curve,AUC)、精确率(precision,P)和召回率(recall,R)是常用的性能评价指标。从不同的角度衡量模型性能,全面评估分类模型的表现。

1)准确率

准确率表示模型在测试集上正确分类的实例占总实例的比例,是最直观的分类性能指标。其定义如式(15)所示。

$$ACC=\frac{TP+TN}{TP+TN+FP+FN} \tag{15}$$

式中:TP 为真阳性实例数,TN 为真阴性实例数,FP 为假阳性实例数,FN 为假阴性实例数。

2)曲线下面积

AUC 通常指 ROC 曲线(receiver operating characteristic curve)下面积,它是模型在不同阈值下性能的综合度量。ROC 曲线以假阳性率(false positive rate, FPR)为横轴,真正率(true positive rate, TPR) 为纵轴,其定义如式(16)所示。

$$\begin{cases}FPR=\frac{FP}{FP+TN}\\ TPR=\frac{TP}{TP+FN}\end{cases} \tag{16}$$

3)精确率

精确率表示被模型预测为正类的实例中,实际为正类的比例,用来衡量模型预测正类的准确性。其定义如式(17)所示。

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

4)召回率

召回率表示实际为正类的实例中,被模型正确预测为正类的比例,用来衡量模型对正类实例的覆盖能力。其定义如式(18)所示。

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

2.4　参数实验

为了分析超参数对模型分类性能的影响,本文设计并进行了多项实验,主要探讨了固定硬负实例采样比例 ψ_{neg}、硬负实例选择函数 $r_{neg}(t)$、阈值 η、关键实例个数 k 以及超参数对分类性能的影响。

1)固定硬负实例采样比例 ψ_{neg} 对分类性能影响

本实验旨在探究 ψ_{neg} 对模型分类性能的影响,以找出最优的 ψ_{neg} 区间。实验结果如表 2 和 3 所示。

表 2　ψ_{neg} 对 Camelyon 分类性能的影响

Table 2　Impact of ψ_{neg} on Camelyon classification performance

采样比例	ACC	AUC	Precision	Recall
100%	0.883 3	0.926 7	0.899 1	0.907 4
80%	0.888 9	0.928 1	0.916 7	0.900 0
50%	0.894 4	0.930 7	0.925 9	0.900 9
20%	0.905 6	0.938 9	0.935 2	0.909 9
5%	**0.911 1**	**0.942 6**	**0.935 2**	**0.918 2**
2%	0.900 0	0.934 8	0.891 9	0.891 9

表 3　ψ_{neg} 对 TCGA-BRCA 分类性能的影响

Table 3　Impact of ψ_{neg} on TCGA-BRCA classification performance

采样比例	ACC	AUC	Precision	Recall
100%	0.823 0	0.883 9	0.882 4	0.782 6
80%	0.827 8	0.890 9	0.882 4	0.789 5
50%	0.837 3	0.898 7	0.902 0	0.793 1
20%	**0.842 1**	**0.908 6**	**0.902 0**	**0.800 0**
5%	0.837 3	0.900 1	0.892 2	0.798 2
2%	0.832 5	0.892 4	0.892 2	0.791 3

实验结果显示,随着 ψ_{neg} 降低,,AUC 值逐渐提高。在 Camelyon 数据集中,5% 的采样比例表现最佳,而在 TCGA-BRCA 中,20%的采样比例效果更优。过高或过低的比例均会降低分类性能。

当 ψ_{neg} 过高时,模型会倾向于学习大量负样本特征,导致对正样本的区分能力下降,最终影响分类性能。适当降低 ψ_{neg} 可以减少简单负样本的干扰,使模型更聚焦于硬负实例,从而提升判别能力。然而,ψ_{neg} 过低则可能导致负样本信息不足,限制模型对数据特征的全面学习。

通过上述分析可知,ψ_{neg} 显著影响模型的分类性能。综合考虑,本文认为 ψ_{neg} 在[0.05,0.5]较为合理。

2)硬负实例选择函数 $r_{neg}(t)$ 对分类性能的影响

本实验围绕动态硬负采样策略的动态调整机制与收敛性分析展开深入剖析,验证了如式(19)所示的凹函数、式(20)所示的线性函数和式(21)所示的凸函数 3 种动态调整策略对对模型训练的差异化影响。实验基于[0.05,0.5]区间,通过 5 次独立实验取均值,定量分析了不同函数形式对分类性能与训练稳定性的综合影响。3 种函数均通过渐进式调整硬负实例的采样比例,但增速模式显著不同:凹函数在训练初期以快速增速引入高难度样本,随后增速放缓;线性函数保持恒定速率平稳增加;凸函数则呈现低-高增速转变特性,初始阶段增速平缓,中后期加速引入高难度样本。实验结果表明,凸函数的动态调整机制在分类性能与收敛稳定性上表现最优,其作用机制与训练阶段的特征学习需求高度契合。

$$r_{neg}(t) = r_e + (r_b - r_e) \cdot (1 - \frac{t}{T})^2 \tag{19}$$

$$r_{neg}(t) = r_e + (r_b - r_e) \cdot (1 - \frac{t}{T}) \tag{20}$$

$$r_{neg}(t) = r_e + (r_b - r_e) \cdot \sqrt{1 - \frac{t}{T}} \tag{21}$$

如表 4 所示,在 Camelyon 数据集中,凸函数的 AUC 为 0.942 6±0.002 0,显著优于凹函数的 0.920 1±0.003 0 和线性函数的 0.931 7±0.002 5。其 ACC 值达到 0.911 1±0.001 2,相较凹函数提升 0.59%,且 Precision 与 Recall 指标稳定在 0.935 2±0.001 8 和 0.918 2±0.001 5。在 TCGA-BRCA 数据集上,凸函数的 AUC 为 0.908 6±0.002 5,较凹函数的 0.875 7±0.003 8 提升 3.76%,ACC 达 0.842 1±0.002 0,Recall 值稳定在 0.800 0±0.001 2。这种性能优势源于凸函数的动态调整机制:初期通过低比例硬负实例建立基础特征表达,避免过早陷入局部最优;中后期加速引入高难度样本,推动模型突破性能瓶颈。

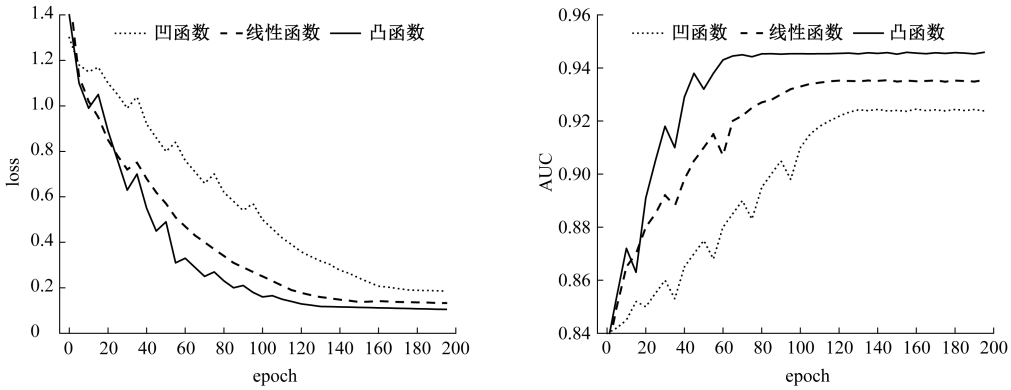
图 7(a)的训练曲线进一步验证了凸函数的收敛优势。在 Camelyon 数据集中,凸函数的 loss 于前 50 轮快速下降,并在 100 轮时趋于稳定,此时测试集 AUC 收敛至 0.942 6±0.002 0;而凹函数和线性函数分别需要 150 轮和 120 轮才达到收敛。在图 7(b)的 TCGA-BRCA 数据集中,凸函数的 AUC 在 120 轮时达到稳定状态,相比之下,凹函数和线性函数的收敛周期延长至 180 轮。这一差异表明凸函数的非线性增速模式有效匹配了模型的优化阶段,既保障了训练初期的稳定性,又通过持续增强的区分性特征提升了最终分类能力。

进一步分析训练稳定性发现,凸函数的 AUC 标准差在 Camelyon 和 TCGA-BRCA 数据集上分别比凹函数降低 33.3%和 34.2%。这一结果表明,凸函数的动态调整机制通过分阶段的样本难度控制,显著抑制了梯度震荡。凹

表 4 不同采样策略在 Camelyon 和 TCGA-BRCA 数据集上的分类性能对比

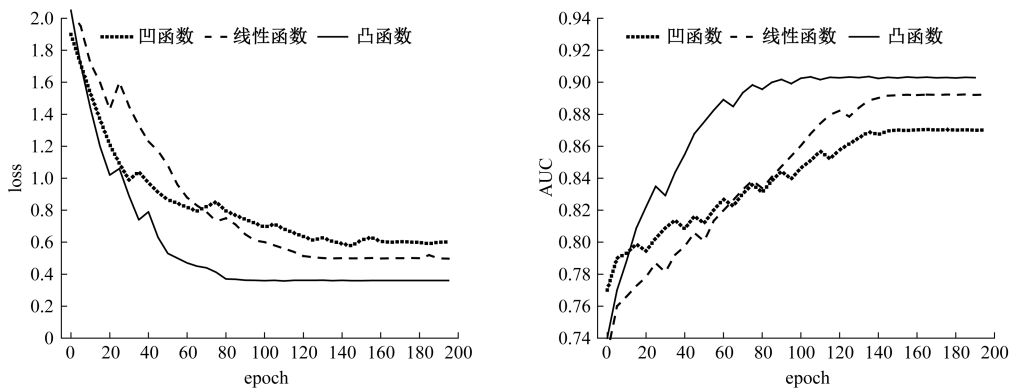
Table 4 Comparative classification performance of sampling strategies on Camelyon and TCGA-BRCA datasets

函数类型	Camelyon				TCGA-BRCA			
	ACC	AUC	Precision	Recall	ACC	AUC	Precision	Recall
凹函数	0.905 6± 0.002 0	0.920 1± 0.003 0	0.935 2± 0.002 2	0.909 9± 0.002 5	0.837 3± 0.003 5	0.875 7± 0.003 8	0.892 2± 0.002 8	0.798 2± 0.002 0
线性函数	0.905 6± 0.001 5	0.931 7± 0.002 5	0.925 9± 0.002 0	0.917 4± 0.002 0	0.837 3± 0.002 5	0.889 8± 0.003 2	0.892 2± 0.002 0	0.798 2± 0.001 5
凸函数	**0.911 1± 0.001 2**	**0.942 6± 0.002 0**	**0.935 2± 0.001 8**	**0.918 2± 0.001 5**	**0.842 1± 0.002 0**	**0.908 6± 0.002 5**	**0.902 0± 0.001 8**	**0.800 0± 0.001 2**



(a) 不同采样策略在Camelyon数据集的训练曲线

(a) Training curves of different sampling strategies on Camelyon dataset



(b) 不同采样策略在TCGA-BRCA数据集的训练曲线

(b)Training curves of different sampling strategies on TCGA - BRCA dataset

图 7 不同采样策略的训练曲线

Fig. 7 Training curves of different sampling strategies

函数因初期增速过快导致训练波动加剧,而线性函数因缺乏动态调节能力未能有效优化收敛曲线。实验结果充分证明,凸函数的自适应样本难度调节机制在动态调整机制与收敛性分析上均表现最优,为模型提供了兼具稳定性和高效性的训练路径。

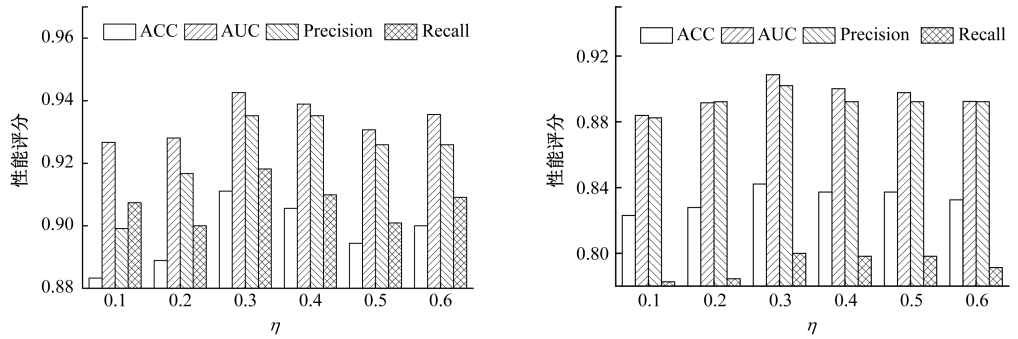
3)阈值 η 对分类性能的影响

本实验研究了阈值 η 对模型分类性能的影响,以找出最优的 η。实验结果如图 8 所示。

实验结果显示,当 η 设置过小或过大时,Precision 和 Recall 指标均受到影响,导致分类结果的可靠性下降。η 处于中间值时,AUC 和 ACC 表现相对较好,同时 Precision 和 Recall 之间取得了较好的平衡。

η 的变化影响了 Precision 和 Recall 之间的权衡关系。较低的 η 倾向于提高 Recall,但假阳性率增加;较高的 η 则对假阳性有严格过滤,但漏掉了部分阳性实例。

考虑到本研究中阴性实例的数量多于阳性实例,且阈值$<$0.1 或$>$0.6 时性能明显下降,本实验将范围限定为 0.1～0.6,并最终选择为 0.3,以兼顾分类性能和阳性实例



(a) 阈值η对Camelyon分类性能的影响
(a) Threshold impact on Camelyon performance

(b) 阈值η对TCGA-BRCA分类性能的影响
(b) Threshold impact on TCGA-BRCA performance

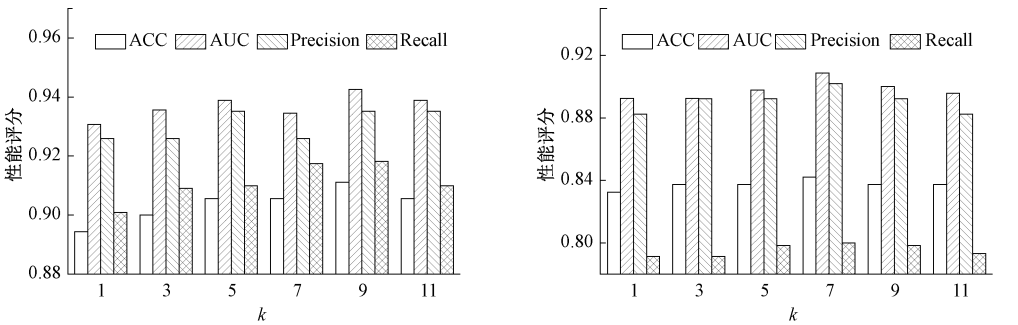
图 8　阈值 η 对分类性能的影响

Fig. 8　Impact of threshold values η on classification performance

识别率。

4)关键实例个数 k 对分类性能的影响

本实验探究了不同关键实例个数 k 对模型分类性能的影响,以找出最优的关键实例数量。实验结果如图 9 所示。为平衡计算复杂度与分类性能,同时避免关键实例过多引入冗余或噪声,本实验将关键实例数量范围限制在 1～11。



(a) 关键实例k对Camelyon分类性能的影响
(a)Impact of key instance k on Camelyon performance

(b) 关键实例k对TCGA-BRCA分类性能的影响
(b)Impact of key instance k on TCGA - BRCA performance

图 9　关键实例 k 对分类性能影响

Fig. 9　Impact of key instance k on classification performance

实验结果表明,对于 Camelyon 数据集,当 k 约为 9 时能获得最佳的 AUC 值;而对于 TCGA-BRCA 数据集,当 k 约为 7 时能获得较好的 AUC 值。整体来看,关键实例数量对 Camelyon 数据集的分类性能影响更大。

这一现象可能与数据集的实例复杂性和类别分布有关。关键实例数量的选择对于模型捕捉关键信息至关重要,过多或过少都可能影响模型的分类性能。

通过上述分析可知,Camelyon 数据集的最优 k 值为 9,而 TCGA-BRCA 数据集的最优 k 值为 7。

5)超参数对分类性能的影响

为探究关键超参数对模型性能的影响,在 Camelyon 和 TCGA-BRCA 数据集上进行了网格搜索实验,结果如表 5 所示。

实验表明,超参数配置对模型性能具有显著影响。在特征编码器微调阶段,批次大小与学习率的协同作用尤为关键。当批次大小设置为 512 时,Camelyon 数据集的 ACC 达到 0.911 1,AUC 为 0.942 6,这验证了该批次在梯度估计稳定性与计算资源间取得的平衡。批次缩减至 128 时,模型 ACC 下降至 0.884 5,AUC 降低至 0.929 7,表明小批次导致的梯度噪声增加削弱了特征学习的稳定性。TCGA-BRCA 数据集在相同批次下表现最佳,其对小批次的较高容忍度可能源于样本异质性要求更大批次以维持特征分布的统计代表性。当尝试 1 024 批次时,显存限制成为实验瓶颈,这提示未来需优化内存管理策略以探索更大批次的可能性。

Camelyon 数据集在 SGD 优化器中采用 1×10^{-2} 学习率时性能最优,过高的 1×10^{-1} 学习率导致参数更新幅度过大,ACC 下降 1.1%至 0.900 0;过低的 1×10^{-3} 学习率则因收敛速度不足使 ACC 损失 1.7%至 0.894 4。TCGA-BRCA 数据集在 Adam 优化器中使用 2×10^{-4} 学习率时 AUC 达到 0.908 6,而学习率提升至 3×10^{-4} 时 AUC 下降 0.85%至 0.900 1,表明该数据集对学习率变化更为敏感,可能与其样本分布稀疏性相关。优化器与数据集特性的适配性差异进一步验证了超参数调优需结合具体任务特征的必要性。

StepLR 调度器以 75 周期为衰减间隔时,在 Camelyon

表 5 超参数对分类性能的影响

Table 5 Effects of hyperparameters on classification performance

设置	值	Camelyon				TCGA-BRCA			
		ACC	AUC	Precision	Recall	ACC	AUC	Precision	Recall
批次大小	128	0.884 5	0.929 7	0.899 1	0.907 4	0.832 5	0.893 8	0.882 4	0.789 5
	256	0.902 4	0.938 0	0.916 7	0.900 0	0.837 3	0.897 7	0.892 2	0.798 2
	512	**0.911 1**	**0.942 6**	**0.935 2**	**0.918 2**	**0.842 1**	**0.908 6**	**0.902 0**	**0.793 1**
	1 024	—	—	—	—	—	—	—	—
SGD 学习率	1×10^{-3}	0.894 4	0.931 2	0.925 9	0.900 9	0.832 5	0.892 4	0.892 2	0.791 3
	1×10^{-2}	**0.911 1**	**0.942 6**	**0.935 2**	**0.918 2**	**0.842 1**	**0.908 6**	**0.902 0**	**0.793 1**
	1×10^{-1}	0.900 0	0.935 6	0.925 9	0.909 1	0.837 3	0.895 8	0.902 0	0.793 1
Adam 学习率	1×10^{-4}	0.894 4	0.929 8	0.916 7	0.908 3	0.827 8	0.895 8	0.902 0	0.793 1
	2×10^{-4}	**0.911 1**	**0.942 6**	**0.935 2**	**0.918 2**	**0.842 1**	**0.908 6**	**0.902 0**	**0.793 1**
	3×10^{-4}	0.905 6	0.934 5	0.925 9	0.917 4	0.837 3	0.900 1	0.892 2	0.798 2
StepLR 间隔	50	0.905 6	0.938 9	0.935 2	0.918 2	0.832 5	0.893 8	0.882 4	0.789 5
	75	**0.911 1**	**0.942 6**	**0.935 2**	**0.918 2**	**0.842 1**	**0.908 6**	**0.902 0**	**0.793 1**
	100	0.900 0	0.934 8	0.916 7	0.916 7	0.837 3	0.904 7	0.892 2	0.798 2

和 TCGA-BRCA 数据集上分别取得最优 AUC。缩短衰减间隔至 50 周期,过早衰减破坏了收敛过程,导致 TCGA-BRCA 的 ACC 下降 0.96%至 0.832 5;延长至 100 周期则使 Camelyon 的 AUC 下降 0.78%至 0.934 8,TCGA-BRCA 的 AUC 下降 0.39%至 0.904 7,表明过长间隔可能加剧过拟合风险。这种差异源于 Camelyon 数据集高度集中的分布特性,其模型需要更精确的学习率控制以避免过早收敛于局部最优。

2.5 消融实验

为了验证各模块对模型性能的贡献,本文在 Camelyon 和 TCGA-BRCA 数据集上进行了消融实验,实验结果分别如表 6、7 所示。

表 6 在 Camelyon 数据集上的消融实验

Table 6 Ablation experiments on the Camelyon datasets

实验	基线	LFFM	FREM	DBBP	DHNIM	ACC	AUC	Precision	Recall
1	√					0.883 3	0.926 7	0.899 1	0.907 4
2	√	√				0.888 9	0.928 1	0.916 7	0.900 0
3	√		√			0.894 4	0.929 8	0.916 7	0.908 3
4	√			√		0.894 4	0.931 2	0.925 9	0.900 9
5	√				√	0.900 0	0.934 8	0.916 7	0.916 7
6	√	√	√			0.894 4	0.930 7	0.925 9	0.900 9
7	√	√	√	√		0.900 0	0.935 6	0.925 9	0.909 1
8	√	√	√		√	0.905 6	0.938 9	0.935 2	0.909 9
9	√			√	√	0.905 6	0.934 5	0.925 9	0.917 4
10	√	√	√	√	√	**0.911 1**	**0.942 6**	**0.935 2**	**0.918 2**

在表 6 的 Camelyon 数据集上,模块间的协同与冗余效应呈现显著的数据依赖性。MFFE 中的 LFFM 与 FREM 单独启用时存在显著冗余效应:实验 2 中 LFFM 仅使 AUC 从 0.926 7 微增至 0.928 1,而实验 3 的 FREM 将 Recall 从 0.907 4 提升至 0.908 3。这种局部优化揭示单一模块难以突破特征表征的维度限制。当二者通过 MFFE 协同作用时,实验 6 的 AUC 提升至 0.930 7 且 Precision 达 0.925 9,表明高频细节增强与多尺度融合形成了特征互补。DBBP 模块的独立应用在实验 4 中使 Precision 提升 2.68%至 0.925 9,但其与 MFFE 的协同在实验 7 中进一步将 AUC 提高 0.49%至 0.935 6,说明 DBBP 需依赖优化的特征空间。DHNIM 的独立增益效应显著,实验 5 使 Recall 提升 0.93%至 0.916 7,而在全模块协同下实验 10 的 Recall 稳定于 0.918 2,表明硬负实例筛选需前置特征优化。值得注意的是,实验 8 中 MFFE 与 DHNIM 组合使 ACC 达 0.905 6,较实验 5 提升 0.56%,揭示特征质量与动

表 7　在 TCGA-BRCA 数据集上的消融实验

Table 7　Ablation experiments on the TCGA-BRCA datasets

实验	基线	LFFM	FREM	DBBP	DHNIM	ACC	AUC	Precision	Recall
1	√					0.823 0	0.883 9	0.882 4	0.782 6
2	√	√				0.827 8	0.891 5	0.892 2	0.784 5
3	√		√			0.827 8	0.890 9	0.882 4	0.789 5
4	√			√		0.832 5	0.893 8	0.882 4	0.789 5
5	√				√	0.837 3	0.898 7	0.902 0	0.793 1
6	√	√	√			0.832 5	0.892 4	0.892 2	0.791 3
7	√	√	√	√		0.837 3	0.895 8	0.902 0	0.793 1
8	√	√	√		√	0.837 3	0.897 7	0.892 2	0.798 2
9	√			√	√	0.837 3	0.900 1	0.892 2	0.798 2
10	√	√	√	√	√	**0.842 1**	**0.908 6**	**0.902 0**	**0.800 0**

态硬负实例筛选的协同效应。

在表 7 的 TCGA-BRCA 数据集上，模块间的冗余效应呈现差异化特征。LFFM 单独启用时实验 2 的 AUC 仅增长 0.76%，与 FREM 单独作用时的 0.890 9 形成性能冗余，表明在组织异质性更强的数据场景下，单一特征增强策略易受噪声干扰。DBBP 与 DHNIM 的协同路径在该数据集表现显著，实验 9 中二者组合使 AUC 达到 0.900 1，较单独 DBBP 的 0.893 8 提升 0.63%，说明动态硬负实例筛选需要伪标签优化策略的配合。模块间的协同效应在实验 10 达到峰值，MFFE 通过 0.908 3 的 Recall 为 DBBP 提供可靠的特征支撑，而 DHNIM 的 0.800 0 Recall 则通过渐进式硬样本挖掘强化了模型鲁棒性，三重协同使 AUC 达到 0.908 6。值得注意的是，实验 7 中 MFFE 与 DBBP 的协同使 Precision 达到 0.902 0，但 Recall 仅为 0.793 1，说明出现了特征增强与实例筛选的局部冗余；而当引入 DHNIM 形成完整协同链后，实验 10 在保持高 Precision 的同时将 Recall 提升至 0.800 0，证明动态样本挖掘能有效平衡精度与召回的矛盾。冗余效应在跨模块组合中同样存在，实验 6 中 MFFE 全组件启用时的 AUC 为 0.892 4，反而低于单独 DBBP 的 0.893 8，提示在特定数据分布下多尺度特征可能引入干扰信息。

跨数据集分析表明，模块协同通过三级优化路径实现性能突破：MFFE 的多尺度频域特征为 DBBP 提供判别依据，DBBP 的聚合优化减少特征稀释干扰，DHNIM 通过动态调节硬负实例筛选比例使模型持续聚焦最具挑战性样本。这种闭环协同机制在 Camelyon 数据集使 ACC 从 0.883 3 提升至 0.911 1，增幅达 2.78%；在 TCGA-BRCA 数据集推动 AUC 从 0.883 9 增长至 0.908 6，增幅 2.47%。冗余效应主要源于模块功能的局部重叠，如 MFFE 的特征增强与 DBBP 的实例筛选在信息过滤维度存在交叉，但当通过 DHNIM 建立动态调节机制后，这种冗余可转化为互补优势，体现在实验 10 的两个数据集的 Recall 分别达到 0.918 2 和 0.800 0，较基线提升 1.08% 和 1.74%。

2.6　对比实验的定量分析

为了验证 FDHN-MIL 在乳腺癌 WSI 分类任务中的有效性，在 Camelyon 和 TCGA-BRCA 数据集上与基于注意力的 ABMIL[11]、双流架构的 DSMIL[14]、引入聚类约束的注意力方法 CLAM[13]、采用迭代自我调节的监督对比学习方法 ItS2CLR[30]，以及基于硬负样本挖掘策略的 HNM[31] 这 5 种具有代表性的 MIL 方法进行了对比实验。其中 ABMIL、DSMIL 和 CLAM 未微调特征编码器，ItS2CLR 和 HNM 则通过微调编码器显著提升了性能。

从表 8 的实验结果可以看出，FDHN-MIL 在 Camelyon 数据集上的 ACC 分别比 ABMIL、CLAM、DSMIL、ItS2CLR 和 HNM 提高了 7.19%、6.49%、7.19%、3.15% 和 1.23%；在 AUC 上也有显著优势，分别提高了 14.13%、13.43%、12.49%、1.71% 和 0.99%。在 TCGA-BRCA 数据集上，ACC 提升了 9.99%、2.32%、9.99%、2.32% 和 1.15%，AUC 提升了 10.76%、2.75%、9.27%、2.79% 和 0.69%。此外，在 Precision 和 Recall 指标上，本文的方法也表现优异，如图 10 的混淆矩阵所示，大部分阳性样本和阴性样本被成功识别，表明模型的区分能力强且准确性高。在 Camelyon 数据集中，真正例和真负例数量明显高于误分类样本，例如阳性样本的分类 Recall 达到了 0.935 2，与 HNM 的 0.925 9 相比提升了约 1.02%，同时 Precision 也有所提升，达到了 0.918 2，表现更加均衡。在 TCGA-BRCA 数据集上，同样表现出较低的假阳性和假阴性数量。

ABMIL、DSMIL 和 CLAM 依赖于预训练的编码器，其特征编码能力受到限制，无法有效捕捉全切片中的细节信息和多尺度特征，导致在复杂分类任务中的表现不佳。虽然 ItS2CLR 和 HNM 通过微调编码器提升了特征编码能力，但伪标签仍含有噪声，未能彻底消除对模型训练的负面影响，同时在硬负实例的筛选上仍存在不足，影响了模型的分类能力。

表 8　在 Camelyon 和 TCGA-BRCA 数据集上的对比实验

Table 8　Comparative experiments on the Camelyon and TCGA-BRCA datasets

方法	Camelyon				TCGA-BRCA			
	ACC	AUC	Precision	Recall	ACC	AUC	Precision	Recall
ABMIL	0.850 0	0.825 9	0.888 9	0.864 9	0.765 6	0.820 3	0.823 5	0.730 4
CLAM	0.855 6	0.831 0	0.898 1	0.966 1	0.823 0	0.884 3	0.882 4	0.782 6
DSMIL	0.850 0	0.837 8	0.870 4	0.878 5	0.765 6	0.831 3	0.823 5	0.730 4
Its2CLR	0.883 3	0.926 7	0.907 4	0.899 1	0.823 0	0.883 9	0.882 4	0.782 6
HNM	0.900 0	0.933 3	0.925 9	0.909 1	0.832 5	0.902 3	0.892 2	0.791 3
FDHN-MIL	**0.911 1**	**0.942 6**	**0.935 2**	**0.918 2**	**0.842 1**	**0.908 6**	**0.902 0**	**0.800 0**

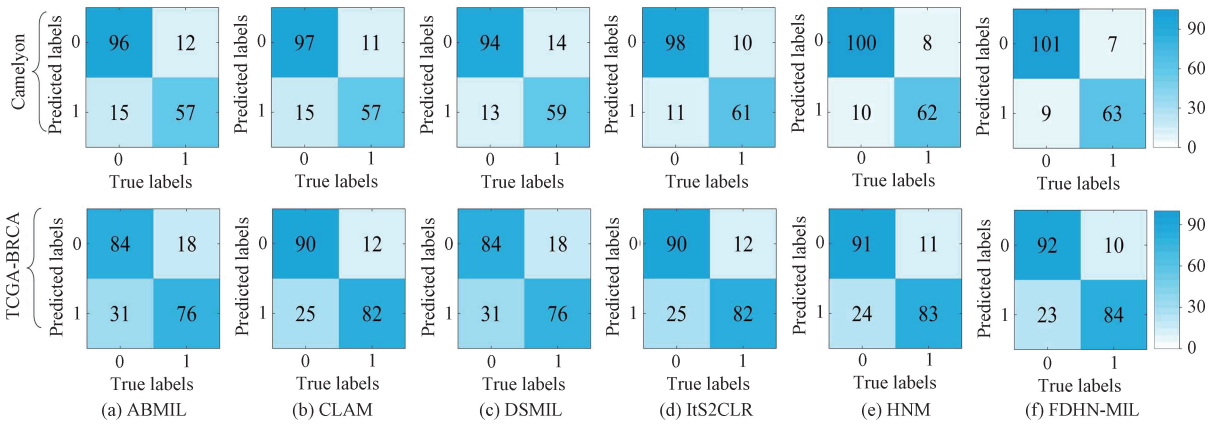


图 10　Camlyon 和 TCGA-BRCA 数据集混淆矩阵

Fig. 10　Confusion matrices for Camlyon and TCGA-BRCA dataset

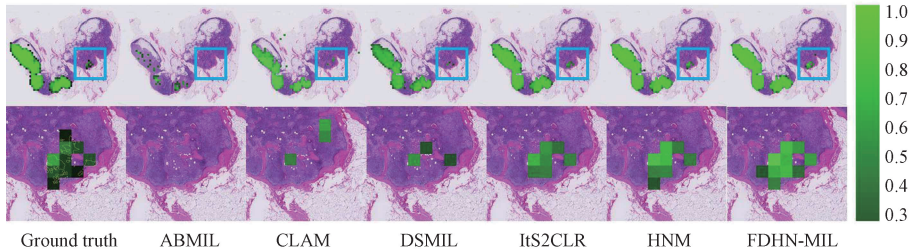
FDHN-MIL 通过 MFFE、DBBP 和 DHNIM 的协同作用,解决了现有研究在特征编码、特征聚合以及负实例选择方面的不足,提升了分类性能。MFFE 通过多尺度信息融合和频域特征增强,更好地捕捉图像细节,提升病灶区域的识别能力。DBBP 针对乳腺癌 WSI 阳性切片肿瘤区域少且关键的特性,更有效筛选代表性区域与聚合特征,生成更高质量伪标签,增强对重要病理信息的捕捉。DHNIM 将阴性 WSI 实例按伪标签值从低到高排序,以硬负实例伪标签选择函数作比例参数,筛选出更具挑战性的硬负实例,减少简单负实例干扰。通过这些优化,FDHN-MIL 在乳腺癌的 WSI 分类任务中展现出明显优势,超越了其他对比方法。

2.7　对比实验的定性分析

为了更加直观地对比不同方法生成的伪标签质量差异,给出不同方法在 Camelyon 和 TCGA-BRCA 数据集上的伪标签可视化对比结果。如图 11 所示。绿色区域的显著程度与对应方法生成的伪标签数值正相关,伪标签值越高,绿色区域越突出,表示该区域被判定为肿瘤病灶的概率越大。

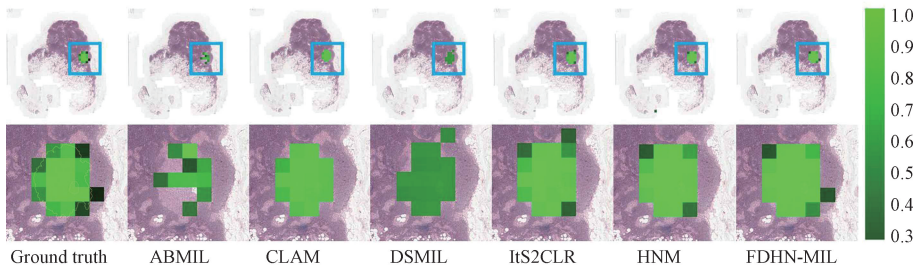
在 Camelyon 数据集上,真实标签(Ground truth)中肿瘤区域被清晰标记为绿色,且病变区域明确且范围较大,蓝色框内的局部病灶区域边界清晰,为伪标签质量评估提供了标准。从图 11 可以看出,ABMIL 方法生成的伪标签较为分散,局部病灶区域的绿色小点与真实肿瘤区域差异较大,准确性较低;CLAM 方法生成的伪标签有所改善,绿色区域的分布有所增加,但仍存在肿瘤区域预测不完整和边界偏差问题;DSMIL 方法生成的伪标签在整体面积和分布上与真实标签差距较大,对肿瘤细胞聚集区域的识别存在遗漏和假阳性问题;ItS2CLR 方法生成的伪标签在分布密度上有所进步,但肿瘤边缘的伪标签值仍存在波动;HNM 方法生成的伪标签能够基本定位肿瘤范围,但在细节区域仍存在小范围伪标签缺失或误判。相比之下,FDHN-MIL 生成的伪标签与真实标签高度吻合,解决了现有方法因关键信息丢失或特征稀释导致伪标签质量不佳的问题,其绿色区域在空间分布、面积覆盖和边界清晰度方面均表现出最优的伪标签质量,特别是在局部病灶区域展现出精准的伪标签预测能力。

接下来,对比不同方法在 TCGA-BRCA 数据集上的表现,该数据集相较于 Camelyon 病变区域相对较小且位置较为集中,对各方法的检测精度要求更高。同样地,蓝色框内标注了局部病灶区域,为评估各方法的预测能力提供了参考标准。从图 11 可以看出,ABMIL 方法表现较差,整张 WSI 的绿色区域较少,且与真实肿瘤区域的对应位置和范围差异较大,局部病灶假阴性率极高,参考价值较低。



(a) Camlyon 数据集可视化结果

(a) Visualization results for Camlyon dataset



(b) TCGA-BRCA 数据集可视化结果

(b) Visualization results for TCGA-BRCA dataset

图 11　Camlyon 和 TCGA-BRCA 数据集可视化结果

Fig. 11　Visualization results for Camlyon and TCGA-BRCA datasets

CLAM 方法虽然绿色区域有所增加，但肿瘤位置和范围的预测偏差较大，局部病灶无法准确识别。DSMIL 方法与 Camelyon 数据集类似，整体和局部病灶的肿瘤区域预测不准确，存在遗漏和假阳性问题。ItS2CLR 方法的绿色区域有所改善，但肿瘤区域的完整性和准确性仍存在差距，局病灶肿瘤边缘预测不精确。HNM 方法能够识别部分肿瘤区域，但与真实标签相比，仍存在差距，局部病灶的细节识别仍存在不足。本文的方法所生成的伪标签在该数据集上表现优越，绿色区域能够准确覆盖肿瘤区域，局部病灶肿瘤细胞聚集区域识别精准，边界清晰，虽然仍存在细微不足，但在检测准确性和完整性上的优势明显，生成的伪标签更接近真实标签，为病理诊断提供了可靠支持和准确信息。从两个数据集的可视化结果可以看出，ABMIL、CLAM、DSMIL、ItS2CLR 和 HNM 作为代表性算法，表现出各自的优缺点。ABMIL 方法依赖于基于注意力的深度学习模型，但未对特征编码器进行微调，这导致其在肿瘤区域准确识别上存在较大缺陷，尤其在复杂图像细节捕捉和多尺度特征融合方面表现不足。CLAM 通过引入聚类约束进行注意力学习，但由于模型对肿瘤区域细节的处理能力有限，依然存在预测不完整和边界偏差等问题。DSMIL 方法虽然采用了双流网络结构处理不同特征，但未对特征编码器进行微调，其特征提取能力受到限制，导致肿瘤区域的预测分布不均，存在大量假阳性和遗漏。ItS2CLR 通过自我调节监督对比学习优化了特征表达，尽管绿色区域有所改善，但伪标签中的噪声问题仍未完全消除，导致局部病灶边缘预测不精确。HNM 方法引入硬负实例挖掘，尽管能够改善一些性能，但在细节处理上仍然无法完全避免遗漏和误判，尤其是在肿瘤内部小区域的识别上。

与上述方法不同，FDHN-MIL 通过 MFFE、DBBP 和 DHNIM 等模块的协同作用，解决了传统方法在多尺度特征融合和细节处理上的不足，实现了病理区域检测准确性与伪标签可靠性的双重提升。MFFE 模块通过多尺度信息融合和频域特征增强，克服了现有方法在细节捕捉和噪声抑制方面的不足，有效提升了对肿瘤区域的识别能力。DBBP 模块通过关键实例选择和伪标签生成，使得模型能够更加准确地聚焦于具有代表性的病理区域，减少不相关区域的干扰，生成的伪标签在空间分布和区域覆盖度上更接近真实标签。DHNIM 模块通过动态选择更具挑战性的硬负实例伪标签，进一步增强了模型对复杂区域的辨识能力，减少伪标签噪声的负面影响。通过这些优化，FDHN-MIL 能够更精准地识别肿瘤区域，尤其在病灶定位定位和边界预测方面表现出明显的优势。

3　结　　论

本文提出了一种结合频域特征提取与硬负实例挖掘的乳腺癌全切片图像分类方法。通过 MFFE、DBBP 以及 DHNIM，FDHN-MIL 有效解决了当前研究中特征编码模块在多尺度信息捕捉、复杂纹理处理及应对不均匀色彩分布方面的不足，改善了特征聚合过程中关键信息丢失或特征稀释的问题，同时优化了负实例的选择策略，提高了伪标签生成质量，并更有效地筛选负实例。Camelyon 和 TCGA-BRCA 数据集上的实验结果表明，本方法在分类 Precision、AUC 等指标上均显著优于现有方法，尤其是在

关键实例选择和硬负实例利用上展现出优势。

尽管取得了较好的分类性能,本研究仍有一些值得进一步探索的方向。例如,不同全切片放大倍率特征间关系的深入研究仍具有潜力,这将有助于进一步提升模型的跨尺度特征提取能力。此外,在多样化实例分布的情况下,本方法的泛化能力仍需进一步验证。未来的研究可以重点关注不同放大倍率特征的跨尺度交互,并在更大规模、更复杂的数据集上验证方法的适用性。此外,将本文的分类框架与临床诊断需求相结合,开发更具实际应用价值的辅助诊断工具,将是重要的研究方向。

参考文献

[1] ZHUANG J J, WU X, MENG D, et al. A swin transformer and residualnetwork combined model for breast cancer disease multi-classification using histopathological images[J]. Instrumentation, 2024, 11(1): 112-120.

[2] ELLIS I, WEBSTER F, ALLISON K H, et al. Dataset for reporting of the invasive carcinoma of the breast: Recommendations from the International collaboration on cancer reporting[J]. Histopathology, 2024, 85(3): 418-436.

[3] EVANS A J, BROWN R W, BUI M M, et al. Validating whole slide imaging systems for diagnostic purposes in pathology[J]. Archives of Pathology and Laboratory Medicine, 2022, 146(4): 440-450.

[4] ZARELLA M D, BOWMAN D, AEFFNER F, et al. A practical guide to whole slide imaging: a white paper from the digital pathology association[J]. Archives of Pathology and Laboratory Medicine, 2019, 143(2): 222-234.

[5] DAVIS J E, SALTZ J H. Patch-based convolutional neural network for whole slide tissue image classification[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2424-2433.

[6] XU Y, JIA ZH P, AI Y Q, et al. Deep convolutional activation features for large scale brain tumor histopathology image classification and segmentation[C]. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, 2015: 947-951.

[7] CARBONNEAU M A, CHEPLYGINA V, GRANGER E, et al. Multiple instance learning: A survey of problem characteristics and applications[J]. Pattern Recognition, 2018, 77: 329-353.

[8] QUELLEC G, CAZUGUEL G, COCHENER B, et al. Multiple-instance learning for medical image and video analysis [J]. IEEE Reviews in Biomedical Engineering, 2017, 10: 213-234.

[9] CHEPLYGINA V, DE BRUIJNE M, PLUIM J P W, et al. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis[J]. Medical Image Analysis, 2019, 54: 280-296.

[10] CAMPANELLA G, HANNA M G, GENESLAW L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images[J]. Nature Medicine, 2019, 25(8): 1301-1309.

[11] ILSE M, TOMCZAK J, WELLING M. Attention-based deep multiple instance learning[C]. 2018 35th International Conference on Machine Learning, 2018: 2127-2136.

[12] OQUAB M, BOTTOU L, LAPTEV I, et al. Is object localization for free? Weakly-supervised learning with convolutional neural networks[C]. 2015 IEEE Conference on Computer Vision and Pattern Recognition, 2015: 685-694.

[13] LU M Y, WILLIAMSON D F K, CHEN T Y, et al. Data-efficient and weakly supervised computational pathology on whole-slide images [J]. Nature Biomedical Engineering, 2021, 5(6): 555-570.

[14] LI B, LI Y, ELICEIRI K. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning [C]. 2021 IEEE Conference on Computer Vision and Pattern Recognition, 2021: 14313-14323.

[15] AZIZI S, MUSTAFA B, RYAN F, et al. Big self-supervised models advance medical image classification[C]. 2021 IEEE International Conference on Computer Vision, 2021: 3458-3468.

[16] CARON M, TOUVRON H, MISRA I, et al. Emerging properties in self-supervised vision transformers[C]. 2021 IEEE International Conference on Computer Vision, 2021: 9630-9640.

[17] GRILL J B, STRUB F, ALTCHÉ F, et al. Bootstrap your own latent: a new approach to self-supervised learning[C]. 2020 34th International Conference on Neural Information Processing Systems, 2020: 21271-21284.

[18] HE K M, FAN H Q, WU Y X, et al. Momentum contrast for unsupervised visual representation learning[C]. 2020 IEEE Conference on Computer Vision and Pattern Recognition, 2020: 9726-9735.

[19] ZBONTAR J, JING L, MISRA I, et al. Barlow twins: Self-supervised learning via redundancy reduction[C]. 2021 38th International Conference on Machine Learning, 2021: 12310-12320.

[20] ZHANG H R, MENG Y D, ZHAO Y T, et al.

DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification[C]. 2022 IEEE Conference on Computer Vision and Pattern Recognition, 2022: 18780-18790.

[21] CIGA O, XU T, MARTEL A L. Self-supervised contrastive learning for digital histopathology [J]. Machine Learning with Applications, 2022, 7: 100198.

[22] KAKU A, UPADHYA S, RAZAVIAN N. Intermediate layers matter in momentum contrastive self-supervised learning[C]. 2021 Advances in Neural Information Processing Systems, 2021: 24063-24074.

[23] ZHU W, FERNANDEZ-GRANDA C, RAZAVIAN N. Interpretable prediction of lung squamous cell carcinoma recurrence with self-supervised learning [C]. 2022 5th International Conference on Medical Imaging with Deep Learning, 2022: 1504-1522.

[24] ZHANG J N, HAO F, LIU X Y, et al. Multi-scale multi-instance contrastive learning for whole slide image classification [J]. Engineering Applications of Artificial Intelligence, 2024, 138: 109300.

[25] HU Z J, YANG ZH Y, HU X F, et al. SimPLE: Similar pseudo label exploitation for semi-supervised classification[C]. 2021 IEEE Conference on Computer Vision and Pattern Recognition, 2021: 15099-15108.

[26] ZHANG B, WANG Y D, HOU W X, et al. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling [C]. 2021 Advances in Neural Information Processing Systems, 2021: 18408-18419.

[27] SHIN I, WOO S, PAN F, et al. Two-phase pseudo label densification for self-training based domain adaptation [C]. 2020 European Conference on Computer Vision, 2020: 532-548.

[28] WU H, PRASAD S. Semi-supervised deep learning using pseudo labels for hyperspectral image classification [J]. IEEE Transactions on Image Processing, 2018, 27(3): 1259-1270.

[29] CASCANTE-BONILLA P, TAN F, QI Y, et al. Curriculum labeling: revisiting pseudo-labeling for semi-supervised learning [C]. 2021 35th AAAI Conference on Artificial Intelligence, 2021: 6912-6920.

[30] LIU K N, ZHU W CH, SHEN Y Q, et al. Multiple instance learning via iterative self-paced supervised contrastive learning [C]. 2023 IEEE Conference on Computer Vision and Pattern Recognition, 2023: 3355-3365.

[31] HUANG W T, HU X L, ABOUSAMRA S, et al. Hard negative sample mining for whole slide image classification[C]. 2024 27th International Conference on Medical Image Computing and Computer Assisted Intervention, 2024: 144-154.

[32] 杨昆,王尉丞,秦赓,等.肾透明细胞癌数字病理图像细胞核 ISUP 分级预测[J].电子测量技术,2023,46(4): 121-128.
YANG K, WANG Y CH, QIN G, et al. Prediction of nuclear ISUP grading in digital histopathological images of renal clear cell carcinoma [J]. Electronic Measurement Technology, 2023, 46(4): 121-128.

[33] NEWLING J, FLEURET F. Nested mini-batch K-means [C]. 2016 Advances in Neural Information Processing Systems, 2016: 1360-1368.

作者简介

鲍刘珍,硕士研究生,主要研究方向为计算病理学。
E-mail:b18234111126@163.com

贾伟(通信作者),博士,副教授,CCF 会员,主要研究方向为医学图像处理与分析、计算病理学。
E-mail:jiawnx@163.com

赵雪芬,博士,副教授,主要研究方向为图像处理与分析。
E-mail:snownfen@163.com

孔德凤,硕士,副教授,主要研究方向为图像处理与分析。
E-mail:fengkongde2012@163.com

江海峰,博士,主治医师,主要研究方向为肿瘤病理诊断。
E-mail:jhf0347@163.com