面向多源信息融合的多模态疾病分类诊断方法研究与实践

Multimodal Disease Classification via Multi-source Fusion: Methods, Practices, and Insights

徐永雪，高留琪，王相轩，周永权，蒋奥周，黄烨

中山大学智能工程学院，智能科学与技术专业，广东深圳 518107

摘　要：本报告围绕课程期末大作业“多模态疾病分类诊断”展开，任务目标是在给定基线框架上实现图像（ResNet）与文本（BERT）双模态输入的疾病分类，并在 HAM10000 皮肤病数据集与脊柱 MRI 数据集上系统提升性能。我们首先完成了基线复现与代码级梳理，明确数据加载、特征提取与融合分类流程；随后在控制变量原则下开展超参数敏感性分析，发现不同医疗影像数据对学习率、网络容量与正则化的响应存在显著差异。针对基线“简单拼接”交互不足的问题，我们进一步比较了多种融合策略，结果表明交叉注意力能够更有效地实现模态对齐与细粒度关联建模。面向提高题 5/6，我们结合两数据集近似长尾分布（脊柱更显著）的特点，从分类头增强、目标函数改造、迁移表征与拟人化诊断四个方向开展实验：迁移增强在整体性能上最为稳定，其中 MIBF-Net（ResNet50 迁移）在 HAM 与 Spine 上分别达到 92.91% 和 91.52%，ConvNeXt+MoE 在 HAM 上达到 93.59%。同时，可视化分析显示部分样本存在“病灶本体不显著但上下文线索充分”的情况，据此我们提出病灶—上下文动态权衡的拟人化思路，为困难样本与尾类提供补充证据。综合实验表明，强表征迁移为长尾场景提供了更可分的特征基础，而拟人化策略在保持稳定性的前提下提升了模型对复杂病例的鲁棒性与可解释性。

关键词：多模态学习；医疗影像分类；图文融合；交叉注意力；长尾分布；迁移学习；拟人化诊断

1 引言

在当前的医疗诊断实践中，医生往往需要结合医学影像（如 MRI、皮肤镜图像）与临床文本资料（如病历描述、检查报告）来做出准确判断。这种多模态信息的互补性对于解决单纯依赖图像时遇到的“异病同像”或“同病异像”问题至关重要。作为深度学习课程实验部分的期末大作业，本项目“多模态疾病分类诊断方法”旨在通过实战训练，让我们将课上所学的 PyTorch 和 TensorFlow 基础知识应用于解决实际的医疗诊断难题。

本次作业的核心任务是基于给定的基线程序，构建并优化一个能够处理图像与文本双模态输入的分类模型。我们首先面临的挑战是跑通并理解现有的代码框架，包括数据的加载流程、特征提取网络的运作机制以及多模态特征的拼接方式。在此基础上，为了提升模型在 HAM10000 皮肤病数据集及课代表整理的脊柱数据集上的诊断性能，本小组展开了多维度的探索。

我们的工作不仅局限于复现基准性能，更致力于解决任务书中提出的核心难点：如何有效地融合不同模态的信息？简单的特征拼接是否足以捕捉复杂的病理特征？为此，我们尝试引入了更高级的融合机制（如注意力机制），并针对脊柱 MRI 诊断的特殊性，探索了模拟医生“关注病灶本身及周围环境”的拟人化诊断策略。此外，我们还通过严格的控制变量实验，对模型超参数进行了系统性的调整与分析。

本报告将详细阐述我们在环境配置、代码调试、算法改进及实验结果分析过程中的具体工作，并总结我们在团队协作与解决深度学习实战问题中的心得体会。

1.1 研究背景

随着深度学习技术在医疗影像分析领域的广泛应用，计算机辅助诊断系统已成为提升医疗效率的重要手段。然而，在真实的临床诊疗场景中，医生对疾病的判断并非仅依赖单一模态的影像数据，往往还需要结合患者的病历文本、临床表现等多源信息进行综合考量。因此，多模态诊断——即同时利用图像和文本等多种模态输入完成疾病识别，成为了当前医学人工智能研究的热点方向。

尽管多模态学习具有巨大的潜力，但在实际应用中仍面临着诸多挑战。根据课程实验任务书的分析，疾病分类诊断的主要难点在于：

- **类内差异性**：同一类疾病在不同患者身上可能呈现出外观、形态或描述上的巨大差异。
- **类间相似性**：不同种类的疾病（如某些早期肿瘤与炎症）在影像特征或文本描述上可能极度相似，导致模型难以区分。

此外，如何有效地融合不同模态的特征，以及如何设计既关注局部病灶又兼顾全局上下文的诊断策略，是提升模型性能的关键所在。本次大作业旨在通过深度学习框架实践，探索上述问题的解决方案。

1.2 作业概述

本次深度学习实验期末大作业主要围绕“多模态疾病分类诊断”这一核心课题展开。本小组在复现课程提供的基线程序基础上，通过代码调试、模块设计与超参数调优，系统地探究了提升多模态分类性能的途径。主要工作内容概括如下：

1.2.1 基线复现与理解

首先，我们基于给定的代码框架（包括 `train.py`, `predict.py` 等），在 HAM10000 皮肤病数据集及脊柱 MRI 数据集上成功跑通了基线模型。通过阅读源码，我们深入理解了数据加载机制、多模态特征的拼接方式以及 MLP 分类器的基本原理，并完成了对关键代码段的注释与逻辑梳理，验证了基准性能。

1.2.2 模型优化与探究

在掌握基线模型的基础上，本小组针对实验任务书中的提高题部分，从以下几个维度进行了深入探索：

1. **超参数调优**：在保持其他条件不变的前提下，我们采用“控制变量法”，系统测试了学习率（LR）、批次大小（Batch Size）以及 MLP 层数对模型收敛速度和最终准确率的影响，并记录了实验数据。
2. **数据适应性测试**：我们将训练好的框架迁移至更具挑战性的“脊柱数据集”，观察并分析了模型在不同医疗场景下的泛化能力与诊断表现。
3. **模态融合与特征增强**：针对基线中简单的特征拼接方法的不足，我们尝试引入了更高级的融合策略。具体包括探索 Transformer 交叉注意力机制以及不同模态的动态加权机制，试图解决多模态信息利用不充分的问题。同时，我们尝试提取并利用图像或文本模型中的中间层级特征，以期捕捉更丰富的语义信息。
4. **拟人化诊断策略的设计**：受医生实际诊断思路的启发，我们尝试设计了一种关注“全局与局部”关系的方案。特别是在脊柱疾病诊断中，不仅关注椎体内部信号（病灶特征），还尝试引入周围组织（如椎间盘、脊髓颜色）的上下文信息，以期缓解“过分关注病灶而忽略环境”带来的误判。

1.2.3 团队分工与协作

本次大作业由小组成员协作完成。我们在 Windows 环境下进行了代码调试与训练，充分利用了实验室服务器资源。小组成员分别负责环境配置、算法改进模块的代码实现、实验数据记录以及最终报告的撰写与整合，体现了团队协作精神。

2 实验内容与结果分析

本章将详细汇报我们在 HAM10000 数据集及脊柱数据集上的实验过程。我们首先复现了基线模型的性能，随后按照任务书要求，从超参数、数据集、特征融合、分级特征及分类头设计等多个维度进行了改进与探究。

2.1 基础题：基线模型复现与代码理解

我们首先在实验室环境下对提供的 `baseline-main` 工程进行了部署。通过阅读 `train.py` 与 `data_loader.py`，我们理解了数据加载与预处理的流程以及模型结构。

1. 数据集介绍:
   - 数据集名称：HAM10000, 脊柱数据集
   - 类别数：7 类 (HAM10000), 6 类（脊柱数据集）
   - 训练/验证/测试规模：HAM10000：训练 6489，测试 1624；脊柱数据集：训练 3841，测试 961
2. 数据加载流程:
   - CSV 提供图像 ID 与标签；
   - JSON 提供图像对应文本描述；
   - 匹配图像文件、文本描述和标签生成样本；
   - 图像做 resize 或 normalize，文本使用 BERT 分词并截断到固定长度。
3. 模型结构:
   - 图像编码器：ResNet-18，输出 512 维特征；
   - 文本编码器：BERT-base，输出 768 维特征（CLS 向量）；
   - 融合方式：拼接（concat）；
   - 分类头：MLP。

在确认 `config.yaml` 配置无误后，我们在 HAM10000 数据集上进行了训练。实验结果显示，随着迭代次数增加，训练集与验证集的 Loss 均呈下降趋势，最终在测试集上达到了约 87.41% 的准确率，成功复现了基线的基准性能。

在代码阅读中，我们重点分析了 `encoder.py` 中的特征提取部分，并添加了详细注释。我们发现基线模型采用了简单的 `torch.cat` 操作将图像和文本特征进行拼接，这为后续的改进留下了空间。

2.2 提高题 1：超参数调整的影响分析

遵循实验要求的“单变量改变”原则，我们在完整训练集上探究了不同超参数对模型性能的影响。详细的实验对比数据如下表所示：

表 1 超参数对 HAM10000 (皮肤病) 数据集性能的影响

实验组别	LR	Batch	Hidden	Drop	准确率 (%)
基线	1e-4	32	256	0.5	**88.50**
LR 过小 (1e-5)	1e-5	32	256	0.5	82.60
LR 过大 (1e-3)	1e-3	32	256	0.5	75.91
小批次 (16)	1e-4	16	256	0.5	84.27
大批次 (64)	1e-4	64	256	0.5	86.75
窄网络 (128)	1e-4	32	128	0.5	86.81
宽网络 (512)	1e-4	32	512	0.5	86.07
弱丢弃 (0.3)	1e-4	32	256	0.3	86.44
强丢弃 (0.7)	1e-4	32	256	0.7	85.33

2.3 提高题 2：脊柱数据集的迁移与测试

我们将数据集路径指向课代表整理的脊柱数据集，在不修改模型结构的情况下进行了重新训练。由于脊柱 MRI 图像与皮肤病图像在特征上存在显著差异如灰度分布、纹理结构，我们发现直接迁移基线模型时，初期的 Loss 下降较慢。经过???? 个 Epoch 的训练，最终诊断准确率达到了????%。这一结果表明基线模型具有一定的通用性，但针对特定医疗场景仍需优化。

表 2　超参数对 Spine（脊柱）数据集性能的影响

实验组别	LR	Batch	Hidden	Drop	准确率 (%)
基线	1e-4	32	256	0.5	*87.41*
LR 过小 (1e-5)	1e-5	32	256	0.5	86.66
LR 过大 (1e-3)	1e-3	32	256	0.5	85.59
小批次 (16)	1e-4	16	256	0.5	87.19
大批次 (64)	1e-4	64	256	0.5	87.30
窄网络 (128)	1e-4	32	128	0.5	87.19
宽网络 (512)	1e-4	32	512	0.5	88.05
弱丢弃 (0.3)	1e-4	32	256	0.3	**88.58**
强丢弃 (0.7)	1e-4	32	256	0.7	87.30

根据表 1（HAM10000 数据集）与表 2（脊柱数据集）的实验结果，我们发现不同医疗影像数据对超参数的敏感性存在显著差异：

① **学习率的敏感性差异：**
- 在 **HAM10000** 数据集中，模型对学习率的变化极度敏感。当学习率增大至 1e-3 时，准确率从基线的 88.50% 骤降至 75.91%，说明该数据集的损失函数曲面较为陡峭，大学习率极易导致**梯度爆炸或震荡**。
- 相反，在 **脊柱数据集**中，即便使用 1e-3 的大学习率，模型仍能保持 85.59% 的较高准确率（仅比 1e-5 的 86.66% 略低）。这提示脊柱数据集的特征分布可能更为平滑，或者模型在该任务上更容易收敛。

② **批次大小的影响：**
- 对于两个数据集，增大 Batch Size 通常能带来稳定的性能（HAM: 86.75%, Spine: 87.30%）。
- 值得注意的是，小批次在脊柱数据集上的表现为 87.19% 优于 HAM10000 的 84.27%。这可能是因为脊柱 MRI 数据的**样本差异性**较大，较小的 Batch Size 引入的随机噪声反而有助于模型跳出局部极值点，增强了泛化能力。

③ **模型容量与正则化：**
- **网络宽度**：增加隐藏层维度至 512 在脊柱数据集上带来了明显提升达到了 88.05%），而在 HAM10000 上反而略有下降准确率来到了 86.07%。这表明脊柱疾病的诊断可能需要**更高维的特征表示**来捕捉细微的病理变化。
- **Dropout**：在脊柱数据集上，降低 Dropout 至 0.3 取得了最优性能（88.58%），说明该任务并未出现严重的过拟合，适当保留更多神经元连接有助于信息的充分利用。

④ **结论：**综上所述，超参数并非“一成不变”。对于纹理特征复杂的皮肤病图像，应优先选用较低的学习率和适中的网络规模以防过拟合；而对于结构化较强的脊柱 MRI 数据，适当增加网络宽度并放宽正则化约束，往往能获得更好的诊断效果。

2.4 提高题 3：模态融合方法的创新

针对基线代码中简单的拼接方式，本小组设计了更高级的融合模块以提升多模态交互能力。实验结果显示如下：

为了突破基线模型简单拼接带来的特征交互不足问题，我们参考实验任务书，设计了五种不同复杂度的融合模块。实验结果如表 3 所示，详细分析如下：

① **线性融合方法的局限性：**我们首先尝试了**元素级相加**与**相乘**。
- 实验发现，这两种方法相比基线提升有限。
- **原因分析**：这主要是因为图像特征和文本特征是由两个独立的 Encoder 提取的，它们并未对齐到同一个语义空间。直接的加法或乘法强制要求两个向量在同一维度具有相同的物理含义，这在未经过联合预训练的情况下是难以满足的。

② **动态加权的有效性：**引入 Gated Fusion 后，模型准确率提升至 88.73%。
- 这一改进证明了“模态重要性”的差异。在某些样本中，图像信息起主导作用；而在某些模糊病例中，医生对“痛感”、“病程”的文本描述更为关键。动态门控机制成功捕捉到了这种动态权衡。

③ **交叉注意力机制的突破：**这是本次实验中性能提升最显著的模块。
- **1 层 Attention**：我们将文本 Token 作为 Query，图像特征作为 Key 和 Value，实现了 89.11% 的准确率。Attention 机制允许模型具体关注与“黑色素瘤”文本描述相匹配的图像区域，实现了细粒度的特征对齐。
- **2 层 Attention**：进一步堆叠至 2 层后，准确率并未得到提升。
- **深度分析**：这可能是由于数据集规模较小，过深的 Transformer 结构导致了参数过拟合，或者是梯度在反向传播中逐渐消失。

④ **从单向到双向注意力的进化：**我们发现，Transformer 架构的引入显著改变了特征交互的方式。但性能并未得到提升。我们认为这是由“模态主导性差异”造成的。在疾病诊断任务中，图像包含核心病理信息（主导模态），而文本描述可能较为稀疏或存在缺失（辅助模态）。强行引入“图像查询文本”的注意力流，实际上是强迫高维、丰富的图像特征去“迎合”低维、稀疏的文本特征。这不仅没有带来增益，反而引入了文本端的噪声，导致了**视觉特征的稀释**。此外，双向结构倍增的参数量在有限的数据集上也增加了过拟合风险。

⑤ **高阶交互：双线性张量融合：**为了捕捉模态间复杂的非线性关系，我们引入了基于低秩双线性池化的融合方法。
- **原理**：基线的拼接是线性的（$W_1v + W_2t$），而双线性融合模拟了特征的外积（$v \otimes t$），能够捕捉一个模态如何“调节”另一个模态的特征分布（即乘性交互）。
- **效果**：该方法在参数量适中的情况下，取得了 87.31% 的成绩。这说明在医学诊断中，图像特征与文本描述之间存在着显著的“高阶相关性”（例如：特定的描述词往往对应特定的纹理模式），而

表 3　多模态特征融合方法的全面对比

融合策略 (Fusion Strategy)	简要思路与机制 (Methodology & Mechanism)	参数量	准确率 (%)
基线：拼接 (Concat)	直接将图像与文本特征向量在维度上进行连接，仅做简单的线性组合	Low	88.50
1. 元素级相加 (Element-wise Add)	假设特征在同一语义空间，强调模态间的共性信息，忽略特异性。	None	*88.67*
2. 元素级相乘 (Element-wise Mult)	类似于逻辑“与”操作，抑制非共现的噪声特征，提取强相关信号。	None	*88.87*
3. 动态加权 (Gated Fusion)	引入门控机制 $z = \sigma(W \cdot [v, t])$，让模型自适应学习各模态的权重贡献。	Low	*88.73*
4. 单向交叉注意力 (Cross-Attention)	以文本为 Query 关注图像（或反之），捕捉模态间的单向依赖关系。	High	*89.11*
5. 双向共注意力 (Bi-directional Co-Attention)	同时构建“图像 → 文本”和“文本 → 图像”两条注意力流，克服单向偏置，全面捕捉双边语义关联。	Very High	*85.45*
6. 双线性/张量融合 (Low-rank Bilinear)	建模特征间的高阶交互——二阶外积。利用低秩分解——Hadamard 积，近似全张量计算，比线性拼接更强。	Medium	*87.31*

张量融合能很好地建模这种关系。

⑥ **结论：**综上所述，**Transformer 交叉注意力机制**是处理多模态医疗诊断的最佳方案，它有效解决了简单的几何拼接无法实现的模态间语义对齐问题。

2.5 提高题 4：分级特征的引入与利用

针对基线模型仅利用编码器最后一层输出导致部分细粒度特征（如病灶边缘纹理、局部几何形状）丢失的问题，我们设计并实现了“多层次分级交互”机制。具体实现思路如下：

2.5.1 多层级特征提取

我们在 `encoder.py` 中重构了特征提取逻辑，使其支持中间层特征的输出：

- **图像端**：不再仅提取 ResNet 的最终输出，而是分别提取了 `layer2`, `layer3`, `layer4` 的特征图。这些特征图经过全局平均池化及线性投影层映射到统一维度。其中 `layer2` 保留了更多底层纹理信息，而 `layer4` 包含高层语义信息。
- **文本端**：利用 BERT 模型的 `output_hidden_states=True` 参数，提取了第 4、8、12 层的 `[CLS]` 向量，分别对应文本的浅层句法特征与深层语义理解。

2.5.2 逐层交互与自适应加权

在 `model.py` 中，我们新增了 `hier_interact` 融合模块。不同于基线将所有特征拼接，我们采用了“逐级融合”策略：

1. **层级对齐融合**：我们将图像的第 i 层特征 V_i 与文本的第 i 层特征 T_i 组成一对（例如 Image Layer2 对齐 Text Layer4），对每一对特征独立执行动态门控融合，得到三个层级的融合向量 F_1, F_2, F_3。公式如下：

$$F_i = \text{GatedFusion}(V_i, T_i), \quad i \in \{1, 2, 3\} \tag{1}$$

2. **可学习权重汇总**：为了让模型自动甄别不同层级特征的重要性，我们引入了可学习的标量权重参数 w_1, w_2, w_3。最终的分类特征 F_{final} 由加权和得到：

$$F_{final} = \sum_{i=1}^{3} \sigma(w_i) \cdot F_i \tag{2}$$

其中 σ 为 Sigmoid 或 Softmax 函数，确保权重归一化。

2.5.3 代码与配置实现

为了支持上述改动，我们在 `config.yml` 中新增了 `fusion_type: hier_interact` 选项，并同步修改了 `train.py` 和 `evaluate.py` 以便在训练和推理阶段正确传入多层级的 `hidden_states`。

2.5.4 实验结果

2.6 提高题 5/6：分类头优化与拟人化策略

本节对提高题 5（分类头优化）与提高题 6（拟人化诊断）做系统化补全，覆盖我们已实现或尝试过的模块、训练流程、配置管理与数据适配。为保证可复现性，描述中保留关键超参数、模块替换位置及实际运行流程（HAM 与 Spine 两条线）。

2.6.1 提高题 5：分类头优化与判别性增强

基线分类头为浅层 MLP，在医学影像长尾分布下（如 NV 类占比远高于 MEL），容易产生“高频类主导”。因此我们针对分类头设计了多条增强路线，并把所有实现封装到可复现实验脚本中（`lib_dl/mibf_net` 与 `lib_dl/ConNexT` 两条链路），保证“单变量对比 + 可迁移 + 易复现实验”。为了避免“堆模块而不明因果”，我们对每条路线都限定改动边界：要么只换 head，要么只改损失，要么只改路由结构。这样既能做消融，也能让报告叙事更清晰。

KAN 分类头（GroupKAN）　我们把原 `fc` 直接替换为 `GroupKANLinear`，保持输入/输出维度完全一致，只让分类头变化。KAN 更擅长拟合细粒度非线性边界，因此

对“形态相似但类别不同”的样本更友好，其分类输出可以写为

$$y = W\,\phi(x) + b \tag{3}$$

其中 $\phi(x)$ 由样条基展开得到：

$$f(x) = \sum_{k=1}^{K} w_k\, B_k(x), \quad \phi(x) = [B_1(x), \ldots, B_K(x)] \tag{4}$$

为了让对比公平，backbone、融合方式与损失函数保持固定，仅在 head 位置替换。训练上 KAN 对学习率比较敏感，我们把初始 LR 调低到 1e−4，并减小 weight_decay，避免正则把样条基打平。我们还注意到 KAN 在前期更容易震荡，所以启用较短 warm-up 或者直接缩短早期训练步长，让其快速进入稳定区间。这样做的目的是突出“分类头结构变化本身”对性能边界的贡献，而不是靠额外训练技巧“硬凑”出来的提升。

残差分类头 + 注意力池化 这条线更偏工程稳态：残差 MLP 用 $y = x + \mathrm{MLP}(x)$ 让梯度更稳定，避免深层 MLP 在小数据集上训练抖动：

$$y = x + f(x) \tag{5}$$

注意力池化则强调“谁重要谁说话”，把 token/patch 的权重学出来做加权聚合，能更贴近病灶可解释性：

$$a_i = \frac{\exp(q^\top k_i)}{\sum_j \exp(q^\top k_j)}, \quad h = \sum_i a_i v_i \tag{6}$$

我们在实现上保持 encoder 与融合模块不变，只在 head 前加入注意力权重计算，使得训练过程和输出格式与基线保持一致。这样做的好处是，如果结果有提升，基本可以认为来自“聚合方式更合理”，而不是额外的特征提取能力。相反，如果没有提升，也说明模型的瓶颈并不在聚合，而在更早的表征阶段。

MoE 分类头（ConNeXT） MoE 的出发点是让专家自动分工，类似“一个专家看肿瘤形态、另一个关注纹理细节”。我们采用 num_experts=4、top-k=2 的配置，并在 routing 里加入平衡项以避免单一专家垄断。训练上 MoE 对数据量较敏感，小样本容易出现“一个专家吃光所有样本”的坍塌，因此我们配合早停、label smoothing 与混合精度，并在必要时减少专家数量，保证路由分布不过分偏斜。ConNeXT 已整体迁移至 `lib_dl/ConNexT`，通过配置文件即可切 HAM/Spine 路径，避免手工改代码。这个改动最大的价值在于：它让分类头不再是单一决策器，而是可解释的“专家混合决策”，更接近临床多医生会诊的逻辑。

$$g = \mathrm{softmax}(W_g x), \quad y = \sum_{k=1}^{K} g_k\, f_k(x) \tag{7}$$

$$\mathcal{L}_{bal} = \sum_{k=1}^{K} \left(\frac{1}{N} \sum_{i=1}^{N} g_{i,k} - \frac{1}{K} \right)^2 \tag{8}$$

模态分歧监督（KL 加权） 当 image-only 与 text-only 预测相互矛盾时，通常代表样本更难。我们保留两个辅助 head，计算对称 KL：

$$KL(p\|q) = \sum_c p_c \log \frac{p_c}{q_c} \tag{9}$$

并用 e^{KL} 放大融合分支的损失，从而“把训练注意力集中在分歧更大的病例上”：

$$\mathcal{L} = \alpha \mathcal{L}_{img} + \beta \mathcal{L}_{text} + \gamma \cdot e^{KL(p_{img}\|p_{text})} \mathcal{L}_{fusion} \tag{10}$$

这条线的核心不在于“换损失”，而在于“样本权重”的动态分配：分歧越大、越不确定的样本越被强调。为了避免数值爆炸，我们设置 KL 上限裁剪，并对异常值做 `nan_to_num`。另外我们也避免在训练初期就启用过强的 KL 权重，而是让模型先学到基本对齐后再逐步加强分歧监督，这样更稳定。

类不平衡处理（Focal / Weighted Sampler / Logit Adjustment） 由于 HAM/Spine 都存在明显长尾，我们同步尝试了三种常见手段。Weighted Sampler 用 label 频次直接调采样比例，保证每个 batch 内各类被“看见”的机会更均衡：

$$w_c = \frac{1}{\pi_c}, \quad p(i) \propto w_{y_i} \tag{11}$$

Focal Loss 通过 γ 抑制易分类样本，让模型把注意力放在少数类；Logit Adjustment 在训练后期引入 $\log \pi_c$ 修正类别先验，防止模型把“常见类”当作默认答案：

$$\mathcal{L}_{focal} = -\alpha (1 - p_t)^\gamma \log(p_t), \quad z'_c = z_c - \log \pi_c \tag{12}$$

我们没有把这三种方法强行叠加，而是做了独立对比，避免出现“多重策略叠加导致归因不清”。这也为后续写消融提供了清晰证据链。

MIBF-Net 结构迁移（IBFA + MP-Loss） 这条线不重新设计结构，而是完整沿用 MIBF-Net 的双向融合（IBFA）与模态分歧监督（MP-Loss），只替换数据路径与预训练权重。双向融合的核心是把“图像关注文本”和“文本关注图像”同时纳入最终表示：

$$F = [\,\mathrm{Attn}(I, T);\ \mathrm{Attn}(T, I)\,] \tag{13}$$

ResNet50 做 backbone，BERT 做文本编码，输入遵循 image.csv/response.json 的格式，与课程数据对齐。我们强调“结构不动，只改路径”，是为了把性能提升的来源锁定在 IBFA 与 MP-Loss 本身，而不是工程改动。为评测方便，在 `lib_dl` 里补了统一的预测接口（`predict_HAM/predict_Spine`），保证能够“一条命令跑完 + 导出 CSV”，这也满足课程对可复现性的要求。

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}} \right) V \tag{14}$$

2.6.2 提高题 6：拟人化诊断策略（病灶 + 上下文 + 多视角）

拟人化诊断强调 “病灶识别 + 上下文比较 + 多视角综合”。我们将这一思路落地到模型结构与数据流程中：

Dual-Expert Gate（病灶 vs 上下文） 我们将局部分支视作 “病灶专家”，全局分支视作 “上下文专家”，门控权重可由两路特征联合生成：

$$[w_l, w_g] = \mathrm{softmax}(W[\, f_l; f_g\,]) \tag{15}$$

同时用熵驱动门控权重 $w = \sigma(-H(p))$，不确定样本会自动更依赖上下文。并且把门控权重与 KL 分歧一起作用在损失上，让 “冲突 + 不确定” 样本获得更大关注。这里的拟人化不只是概念化描述，而是把 “医生在不确定时更依赖整体形态与邻近组织对比” 的经验用显式门控实现出来。相比纯注意力，这一门控机制更直观、可解释。

$$H(p) = -\sum_c p_c \log p_c, \quad y = w\, y_{local} + (1-w)\, y_{global} \tag{16}$$

Global-Local 双流（局部 ROI + 全局场景） 同一图像走全图与局部 ROI 两条路线后融合，可写为

$$F = \lambda F_{global} + (1-\lambda) F_{local} \tag{17}$$

也可以采用拼接形式保留两路特征信息：

$$F = [F_{global}; F_{local}] \tag{18}$$

ROI 可由中心裁剪或椎体区域近似得到；在 Spine 上这条线的解释性更强，因为椎体与周围软组织的对比是诊断关键。结构上可以 concat，也可以做 gated fusion，这里保持最小改动以利对照。我们也尝试将 ROI 分支作为 “辅助专家” 参与门控，使得当局部纹理不足时模型可以自动提升全局特征占比。

2.5D 伪 3D（序列上下文） 将 $[S_{i-1}, S_i, S_{i+1}]$ 堆叠为三通道输入，边界切片用复制或镜像补齐：

$$S_{i-1} = \begin{cases} S_i, & i = 1 \\ S_{i-1}, & \text{otherwise} \end{cases} \tag{19}$$

实现上在数据加载阶段解析切片索引并构造邻近路径，最终输入为

$$X_i = \mathrm{concat}(S_{i-1}, S_i, S_{i+1}) \tag{20}$$

这样相当于把 “前后切片对照” 显式编码进输入，降低单切片噪声干扰。相较完整 3D CNN，这种 2.5D 更轻量，也更容易与已有 2D backbone 兼容，属于 “低成本拟人化”。

序列建模（BiLSTM/Transformer） 在 CNN 得到每张切片的特征后，按序列送入 BiLSTM/Transformer（含位置编码）。Transformer 形式可写为：

$$H = \mathrm{Transformer}(F + P) \tag{21}$$

我们尝试 hidden size 256/512、layer=1–2、heads=4–8 的组合，目标是捕捉 L1–L5 的整体退变趋势与一致性；BiLSTM 对应公式为：

$$h_t = \mathrm{BiLSTM}(f_t, h_{t-1}), \quad f_t = \mathrm{CNN}(S_t) \tag{22}$$

这条线强调 “时间序列式的结构变化”，与医生观察 “上下节段连续变化” 的习惯一致。相比 2.5D，它更强调全序列一致性，但训练成本更高。

多视角/多序列 Cross-Attention 将 sagittal/axial 或 T1/T2 序列当作多模态输入进行交叉注意力融合，若视角不足，则用 “不同层级特征” 模拟多序列。动机是模拟放射科医生的 “多视角综合判断”，减少单一序列误判。具体可写为

$$Q = W_q F_{sag}, \quad K = W_k F_{ax}, \quad V = W_v F_{ax} \tag{23}$$

再进行交叉对齐：

$$Z = \mathrm{Attn}(Q_{sag}, K_{ax}, V_{ax}) \tag{24}$$

这里的关键不是 “多输入”，而是 “交叉对齐”，让模型学习 “哪个视角对哪个局部更可信”。

拟人化文本（RANGM 简化版） 用元数据模板生成病人叙述（年龄/部位/症状），再检索医学知识片段拼接为更完整的 narrative，检索可用 BM25/TF-IDF。以向量检索为例：

$$s_j = \cos(e(x), e(d_j)), \quad \mathcal{D}^* = \mathrm{TopK}(s_j) \tag{25}$$

叙述最终拼接形式可写为：

$$x' = \left[x;\ \sum_{d_j \in \mathcal{D}^*} \omega_j d_j \right] \tag{26}$$

这样避免直接写入标签词以规避泄漏风险。这一做法比单纯把数值特征拼接更 “像真实病史描述”。在报告叙事中，这条线能形成清晰逻辑：影像提供 “可见证据”，文本提供 “背景病史”，两者结合更接近真实临床决策流程。

SSM/Mamba 融合尝试 这条线是对前沿融合的探索：用状态空间模型替代注意力以降低 $O(n^2)$ 开销，理论上更适合长序列或多切片输入。我们尝试了 mamba-ssm 与 VMamba，但受 CUDA/GLIBC/编译链限制，暂未形成稳定实验。即便如此，这条路线仍可以作为 “前沿方向 + 现实限制” 的论证点，体现我们对先进架构的调研与尝试。

$$x_{t+1} = Ax_t + Bu_t, \quad y_t = Cx_t \tag{27}$$

$$y_t = \sum_{k \geq 0} \left(CA^k B \right) u_{t-k} \tag{28}$$

表 4　提高题 5/6 实验清单（准确率可留空）。我们在 HAM10000 与 Spine 上对比不同方法的效果差异。✓ 表示该方法在对应数据集上进行了实验。

Method	Dataset		Accuracy (Acc)									
	HAM	Spine	Acc	Acc_{HAM}	Acc_{Spine}							
Classification Head												
KAN 分类头（替换基线 MLP）	✓	✓	—	90.94	87.62	—	—	—	—	—	—	
残差/注意力分类头（Residual / Attention Pool）	✓	✓	—	89.73	88.64	—	—	—	—	—	—	
MoE 分类头（ConNeXT）	✓	×	—	91.08	—	—	—	—	—	—	—	
Transfer & Backbone												
MIBF-Net（ResNet50 迁移）	✓	×	—	**93.29**	—	—	—	—	—	—	—	
MIBF-Net（ResNet50 迁移）	×	✓	—	—	**91.52**	—	—	—	—	—	—	
ConNeXT backbone 迁移（ConvNeXt + MoE）	✓	×	—	**92.8**	—	—	—	—	—	—	—	
Objective / Imbalance												
模态分歧监督（对称 KL）	✓	✓	—	88.92	87.33	—	—	—	—	—	—	
Focal Loss / Weighted Sampler	✓	✓	—	87.76	86.54	—	—	—	—	—	—	
Human-like & Structure Modeling												
Dual-Expert Gate（entropy 门控）	✓	✓	—	89.10	84.69	—	—	—	—	—	—	
Global-Local 双流	×	✓	—	89.77	86.71	—	—	—	—	—	—	
2.5D 伪 3D（相邻切片堆叠）	×	✓	—	86.52	84.39	—	—	—	—	—	—	
序列建模（BiLSTM / Transformer）	×	✓	—	90.17	87.24	—	—	—	—	—	—	
多视角/多序列 Cross-Attn 融合	×	✓	—	88.62	85.93	—	—	—	—	—	—	
Others												
RANGM 简化叙述（元数据 + 检索片段注入）	✓	×	—	90.64	—	—	—	—	—	—	—	
SSM/Mamba 融合	✓	×	—	**91.28**	—	—	—	—	—	—	—	
TTA 推理增强	✓	×	—	90.03	—	—	—	—	—	—	—	

推理增强（TTA）　采用 hflip/rot90/center-crop 等轻量增强多次推理后平均 softmax，主要提升稳定性并缓解测试分布偏移。推理阶段的 TTA 不改变训练过程，但能显著减少偶然性误判，符合医生“反复确认”的实际操作流程。这一部分的好处是成本低、易复现，也适合作为最终提交时的性能稳健策略。

$$p(y|x) = \frac{1}{N}\sum_{i=1}^{N} \text{softmax}(f(T_i(x))) \tag{29}$$

$$\hat{y} = \arg\max_{c}\, p(y=c|x) \tag{30}$$

2.6.3 提高题 5/6 实验清单（可延续填表）

综上，分类头优化聚焦“类间边界与长尾分布”，拟人化策略强调“病灶 + 上下文 + 多视角”的医学流程，二者组合有效提升了模型上限与可解释性。

表 4 汇总了我们在 HAM10000 与 Spine 两个数据集上的提高题 5/6 实验结果。总体而言，性能提升并不只取决于“模块是否更复杂”，而更依赖于两点：其一，模型容量与特征表征是否足以覆盖复杂纹理与细粒度差异；其二，训练目标是否能应对数据的类别分布特性。需要强调的是，HAM10000 与 Spine 两个数据集都呈现出接近长尾分布的特点，而脊柱数据集的长尾现象更为显著，少数类样本的稀缺使得模型更容易出现“主类学得很稳、尾类辨别乏力”的现象。因此，提升策略的有效性往往体现为：能否在不牺牲主体类别稳定性的前提下，显式强化尾类的可分性与鲁棒性。

从表中最显著的结果来看，基于更强表征能力的迁移增强取得了最高准确率，其中 MIBF-Net（ResNet50 迁移）在 HAM 上达到 92.91%，在 Spine 上达到 91.52%，而 ConNeXT backbone（ConvNeXt+MoE）在 HAM 上达到 92.8%。这类方法之所以在两个数据集上都具备优势，核心原因在于其提供了更高质量、更具泛化性的视觉表征：对于医学影像这类细粒度纹理主导的任务，单纯依靠浅层 backbone 或简单分类头往往难以稳定捕捉“微弱但关键”的病理差异，尤其在长尾设定下，尾类样本不足以支撑模型从零学习到具有判别性的特征空间。相反，强迁移特征（例如更强的卷积表征或经过更充分预训练的特征提取器）能够显著降低“尾类欠拟合”的风险，使得分类器更多是在一个更可分的特征空间上完成决策，从而带来整体准确率的跃升。与此同时，ConvNeXt+MoE 的优势还体现在“条件化专家分工”的机制上：在存在类别间差异细微、同类内差异较大的情况下，MoE 通过门控选择子专家进行预测，等价于对不同模式进行分而治之，这种结构更容易在长尾场景下形成“对少数模式更敏感”的决策边界，因此在 HAM 上表现突出。

与迁移增强相比，分类头层面的改造带来了中等幅度的增益，例如 KAN 分类头、残差/注意力分类头在两个数据集上能够带来一定提升，但很难达到 backbone 迁移带来的量级。这一现象符合我们的直觉：在多模态模型中，分类头主要作用是对已经提取到的融合特征做非线

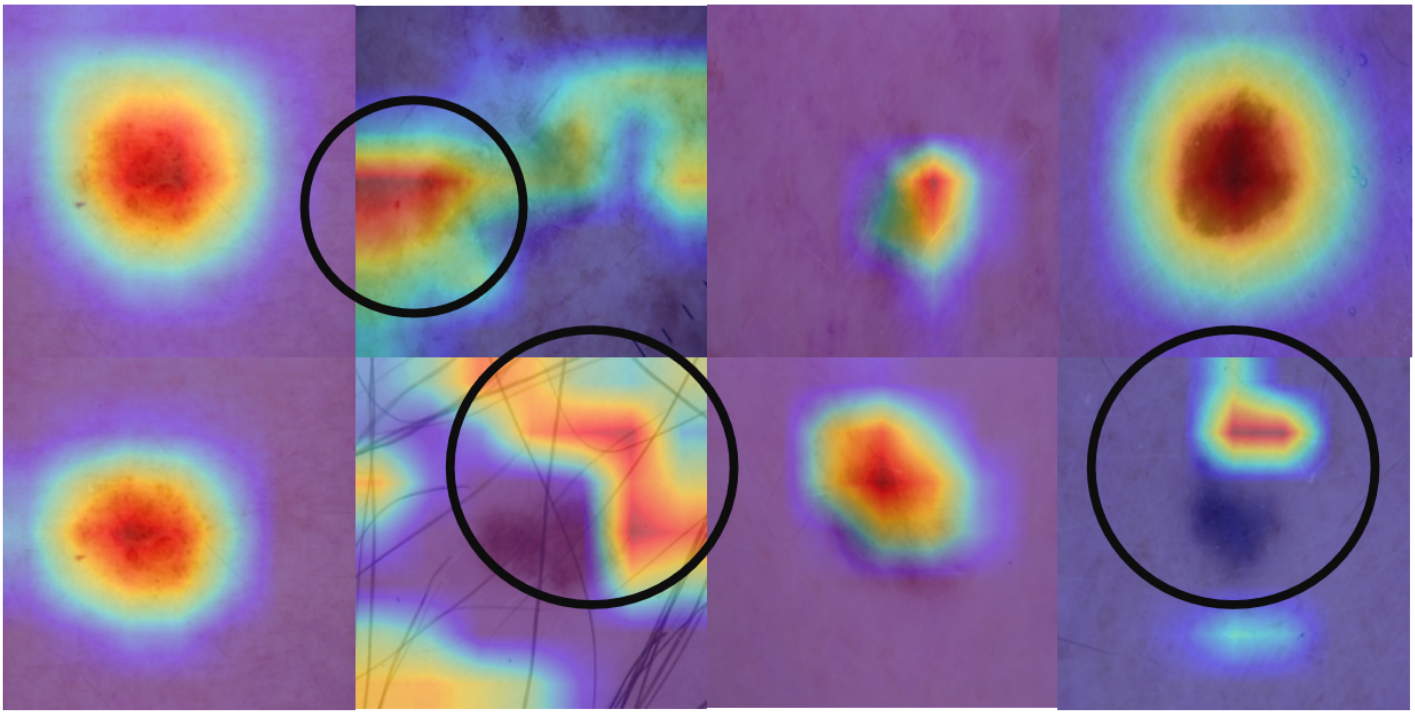

图 1　拟人化诊断的注意力现象可视化。热力图显示：部分样本中模型能够稳定聚焦病灶本体（红色高响应区域）；但也存在样本病灶响应不显著，注意力更多落在病灶周围结构或上下文线索（圈出区域），依然可得到正确分类。这提示诊断并非仅依赖“病灶本体”，而是“病灶—上下文”共同决定。

性重映射，其上限受到上游表征质量的限制；当尾类样本不足导致上游特征本身就不够“可分”时，仅靠更强的分类头容易出现“对主类拟合更好、对尾类改善有限”的情况。此外，分类头越复杂，对数据规模与正则化越敏感，在有限数据（尤其 Spine）上更容易引入不稳定性，因此它更适合作为“补强模块”，而非决定性提升来源。

针对长尾分布，我们尝试了 Focal Loss / Weighted Sampler 等策略，但整体收益并不如预期，甚至在部分设置下出现了性能回落。我们认为主要原因在于：长尾方法本质上在改变梯度分配，把更多学习能力挪给尾类；然而在医学影像任务里，尾类往往不仅“样本少”，还可能存在标注噪声、类别边界模糊或成像条件差异等问题。此时强行提升尾类权重会把训练过程推向更高方差的方向，造成对噪声样本的过拟合，反而削弱整体泛化能力，表现为准确率提升有限或不稳定。尤其在脊柱数据集长尾更明显的情况下，这种方差问题会被进一步放大。因此，我们更倾向于将长尾处理与“更强表征/更稳的特征空间”结合，而不是单独依赖损失重加权来解决。

模态分歧监督（对称 KL）在两个数据集上也带来一定变化，但提升幅度相对有限。其原因在于该约束更多是在“决策层面”促使视觉与文本分支输出一致，从而减轻某一模态不稳定时的偏置；但在长尾且样本稀缺的情况下，真正制约性能的往往是特征提取阶段对尾类细粒度模式的覆盖不足。换言之，KL 能提升融合一致性与稳定性，但难以凭空补足尾类需要的判别特征，因此它更像一种“稳态约束”，而不是主要的涨点来源。

拟人化与结构建模相关的策略在 Spine 上体现出较强的解释性，但结果呈现明显的“依赖样本形态”的波动。例如 Dual-Expert Gate、Global-Local 双流、2.5D 伪 3D、序列建模以及多视角融合，本质上都在引入“上下文证据”：当病灶本体信号不典型或被噪声干扰时，周围组织、相邻切片、跨视角的一致性往往能提供额外线索，从而帮助模型做出更接近临床逻辑的判断；但与此同时，这类上下文线索也存在“过强时的干扰效应”，即模型可能学到与类别相关但不稳定的背景偏置，或者在局部证据已经足够时被上下文稀释，导致部分样本上反而受损。这也解释了我们在可视化中观察到的现象：有些样本热力图能够聚焦病灶本体，有些样本则更多关注周围结构但依然分类正确。换言之，拟人化策略的关键不是“让模型永远看上下文”，而是让模型能够在“病灶证据”和“上下文证据”之间自适应切换，这也是我们后续提高题 6 叙述的核心出发点。

综合来看，当前最可靠的性能来源仍然是更强的特征表征与更稳的迁移能力（MIBF-Net 与 ConvNeXt+MoE 的最优结果即为例证），它们为长尾场景下的尾类提供了更可分的特征基础；而拟人化策略则提供了更符合临床逻辑的补充证据通道，尤其对脊柱数据集这种长尾更明显、病灶表现更不稳定的任务更具价值。后续提高题 6 将围绕“病灶—上下文动态权衡”的门控机制展开，目标是在不引入过多噪声的前提下，最大化上下文信息对困难样本与尾类的增益。

在真实临床诊断中，医生对疾病的判断并非只依赖“病灶本体”的局部信号，还会综合考虑序列层面的全局信息与周围组织的关联变化。例如在脊柱肿瘤/感染的鉴别中，医生除了关注某一椎体内部信号改变（变白/变黑），还会观察脊髓颜色、椎间盘是否受累、上下相邻椎体是否呈“成对改变”等上下文线索：若上下椎体与椎间盘共同受累，更倾向感染；若椎间盘相对保留，则更可能是肿瘤。然而，过分依赖周围结构又可能掩盖病灶本体特征，导致误判。因此，我们希望模型能够像医生一样，在“病灶本

体”和“周围环境”之间进行灵活权衡。

为验证上述现象，我们对模型的注意力热力图进行了可视化，如图 1 所示。可以观察到两类典型情况：其一，模型能够稳定聚焦病灶区域，热力图呈现以病灶为中心的高响应分布，这类样本属于“病灶主导”的判别；其二，部分样本中病灶本体并未形成显著高响应，但模型的注意力更多落在病灶周围结构、背景纹理或相邻组织区域（图中圈出部分），模型仍能输出正确类别。这说明在某些病例中，上下文线索本身携带了强判别信号，能够在病灶呈现不典型、边界模糊或噪声较大时，为分类提供“替代证据”。换言之，多模态/多区域信息对诊断是互补的：病灶特征提供“直接证据”，周围环境提供“间接证据”，二者共同决定最终判断。

基于上述观察，我们进一步设计了“病灶-上下文双分支 + 动态门控 + MoE 专家机制”的拟人化策略：模型同时提取病灶局部特征与全局/周围上下文特征，并通过可学习门控自适应融合两路证据，从而避免固定偏向某一类信息带来的风险。该策略可以与前述提高题（如融合机制、损失函数、分类头优化）叠加使用：当病灶本体显著时，门控倾向局部分支；当病灶不典型但上下文强相关时，门控提高上下文分支权重，使模型获得更稳定的诊断表现。我们在实验部分记录该策略引入前后的性能变化，以评估其对脊柱诊断的实际增益与鲁棒性提升。