A Lightweight Lung Disease Diagnostic System with High Energy Efficiency Based on Knowledge Distillation and Edge-Cloud Coordinated Acceleration

ZHANG Yuhan, WANG Yapeng, ZHANG Xuan, SUN Hao*
(School of Electronic Science and Engineering, Southeast University, Nanjing Jiangsu 211189, China)

Abstract: To address critical challenges regarding real-time performance and portability in clinical lung disease screening, an FPGA-accelerated edge-cloud diagnostic system via knowledge distillation is proposed. Algorithmically, a framework utilizing a ResNet50 teacher to guide a lightweight 4-layer CNN student is constructed to minimize computational complexity while preserving feature fidelity. In terms of hardware, the model is deployed on the PYNQ-Z2 platform via high-level synthesis (HLS), employing optimization approaches such as batch normalization fusion, task-level pipelining, and loop unrolling to ensure efficient edge inference. Furthermore, the system adopts a hierarchical "offline detection-online diagnosis" strategy. While guaranteeing independent edge detection capabilities, it integrates the DeepSeek large language model (LLM) API to generate semantic treatment recommendations, forming a closed-loop diagnostic workflow. Experimental results demonstrate a 94.86% classification accuracy for four lung disease categories, with a single-frame inference latency of 39 ms and power consumption of merely 1.5 W, providing a high-performance solution for rapid diagnosis in resource-constrained scenarios. The source code and hardware design files are available at https://github.com/zhang666max/Efficient-Lung-Diagnosis -via-FPGA-Cloud-and-Knowledge-Distillation.

Key words: pulmonary disease detection; FPGA; knowledge distillation; convolutional neural network; hardware acceleration
EEACC: 6135 **doi**: 10.3969/j.issn.1005-9490.2026.03.034

基于知识蒸馏与 FPGA 边云协同加速的轻量化高能效肺部疾病诊断系统

张宇涵, 王亚鹏, 张 萱, 孙 豪*
(东南大学电子科学与工程学院 江苏 南京 211189)

摘 要: 针对肺部疾病临床筛查对实时性与便携性的严峻挑战, 设计并实现了一种基于知识蒸馏与 FPGA 加速的边云协同辅助诊断系统。算法层面, 构建 ResNet50 教师网络指导轻量化 4 层 CNN 学生网络的知识蒸馏框架, 在保证高维特征提取精度的同时大幅降低模型参数量与计算复杂度。硬件层面, 基于 PYNQ-Z2 平台利用高层次综合技术, 采用批归一化融合、任务级流水线及循环展开等并行优化策略, 实现低延迟、高能效的边缘侧推理检测。系统应用层采用"离线检测-在线诊断"的分级策略, 在保障边缘侧独立检测能力的基础上, 利用云端大模型赋能深层语义诊断。即依托边云协同架构调用 DeepSeek 大语言模型 API, 依据边缘侧推理结果自动生成针对性辅助治疗建议, 构建"影像检测-辅助诊断"闭环。实测表明, 该系统对四类肺部病灶识别准确率达 94.86%, 单张影像推理耗时 39 ms, 功耗仅 1.5 W, 兼具高性能与低功耗特性, 为资源受限场景下的快速辅助诊断提供了有效解决方案。项目源代码及硬件工程文件已开源至: https://github.com/zhang666max/Efficient-Lung-Diagnosis -via-FPGA-Cloud-and-Knowledge-Distillation。

关键词: 肺部疾病检测; FPGA; 知识蒸馏; 卷积神经网络; 硬件加速
中图分类号: TN911.73; TP183 **文献标识码**: A **文章编号**: 1005-9490(2026)03-0712-09

肺部传染性疾病的早期快速筛查是阻断疫情传播的关键防线。尽管胸部 X 光成像因其低辐射与高效率已成为临床初筛的"金标准", 但在医疗资源匮乏的偏远地区, 专业医师短缺严重制约了阅片效率与诊断准确率。深度学习技术虽为自动化辅助诊断提供了新路径, 但高性能卷积神经网络巨大的算力与存储需求, 使其难以直接部署于资源受限的边缘设备。特别是在户外应急筛查与移动巡诊等缺乏稳定网络与电源支持的极端环境下, 诊断终端必须具备离线实时响应能力与高能效比。因此, 如何在算力与功耗双重受限的嵌入式平台上, 构建兼具高精度、低延迟且具备完整诊疗能力的便携式筛查系统, 成为当前医疗

收稿日期: 2026-02-14 **修改日期**: 2026-03-03

领域亟待攻克的难题。

回顾现有研究,尽管基于深度学习的辅助诊断已取得显著进展,但针对边缘侧部署的系统级优化仍面临三个层面的局限。在算法层面,为适应边缘算力,现有工作多直接采用 MobileNet 等通用轻量化网络,降低了运算量,但在处理肺部影像中纹理相似度极高的微小病灶时,往往因浅层特征提取能力不足而牺牲了关键的检测精度,缺乏兼顾轻量化与高精度的专用模型设计[1]。在硬件层面,GPU 的高功耗特性限制了其在移动医疗场景的应用,而现有 FPGA 加速方案多侧重于通用矩阵乘法算子的优化,缺乏针对特定医学影像网络结构的细粒度流水线设计与片上存储适配,导致在低主频下难以突破“访存墙”瓶颈,无法满足毫秒级的实时筛查需求。在系统层面,现有方案多局限于单一的图像分类标签输出,缺乏对病理结果的语义解释与临床建议生成能力,且未建立边缘实时推理与云端大语言模型知识库的有效协同机制,导致系统处于“只检不诊”的开环状态,限制了其在辅助基层医生决策时的临床实用价值[2-3]。

针对上述挑战,本文提出一种基于 FPGA 平台的边云协同肺部疾病辅助诊断系统,旨在通过软硬协同设计解决边缘侧部署的能效与精度矛盾。本文的主要创新工作体现在以下三个方面:首先,在算法层面,提出一种硬件友好的“教师-学生”知识蒸馏架构,利用预训练 ResNet50 教师网络指导轻量化 CNN 学生网络,在大幅压缩模型参数量的同时,通过混合损失函数有效保留了高维病灶特征的提取能力,解决了轻量化模型精度不足的问题。其次,在硬件层面,设计了基于高层次综合(High-Level Synthesis,HLS)的专用 FPGA 硬件加速器,采用批归一化(Batch Normalization,BN)层融合、差异化循环展开及任务级流水线技术,在极低功耗下实现了超低延迟的边缘端并行推理,满足便携式设备的实时性要求。最后,在系统层面,构建了“边缘计算推理+云边认知诊断”的协同架构,创新性地集成了 DeepSeek 大模型 API,依据边缘端的检测结果自动生成包含病理分析与用药指导的结构化报告,实现了从影像感知到辅助决策的完整闭环。

本文后续结构安排如下:第 1 节综述整体系统架构与工作流程;第 2 节详述基于知识蒸馏的高精度轻量化检测模型的构建与训练策略;第 3 节深入分析 FPGA 硬件加速器的并行架构设计与存储优化方法;第 4 节展示系统的实验结果并进行多维度的性能对比分析;第 5 节总结全文。

1　总体系统设计

面向肺部传染性疾病在基层医疗场景中的快速筛查需求,本文构建了一种基于 FPGA 平台的边云协同辅助诊断系统。如图 1 所示,该系统采用软硬协同的设计思想,通过嵌入式边缘计算与云端大模型认知的深度融合,实现了全链路自动化的辅助诊疗流程。具体工作流程如下:

①影像数据采集与传输:系统工作流程起始于 CT 影像获取环节。CT 扫描仪采集的高分辨率肺部影像数据,通过千兆以太网接口加载至 PYNQ-Z2 边缘计算节点的 DDR3 共享内存中,作为后续实时处理的数据源。

②PS 端预处理与任务调度:Zynq 的处理系统(Processing System,PS)端作为主控单元,首先读取内存中的原始影像,执行图片尺寸归一化、去噪等预处理操作,以适配硬件加速器的输入数据规范。随后,PS 端应用层通过高级可扩展接口(Advanced eXtensible Interface,AXI)总线配置可编程逻辑(Programmable Logic,PL)端的控制寄存器,生成启动信号以触发硬件加速任务。

③PL 端硬件加速推理:PL 端部署了基于高层次综合(HLS)优化的轻量化卷积神经网络(Convolutional Neural Network,CNN)加速核。预处理后的图像数据经由直接存储器访问(Direct Memory Ac-

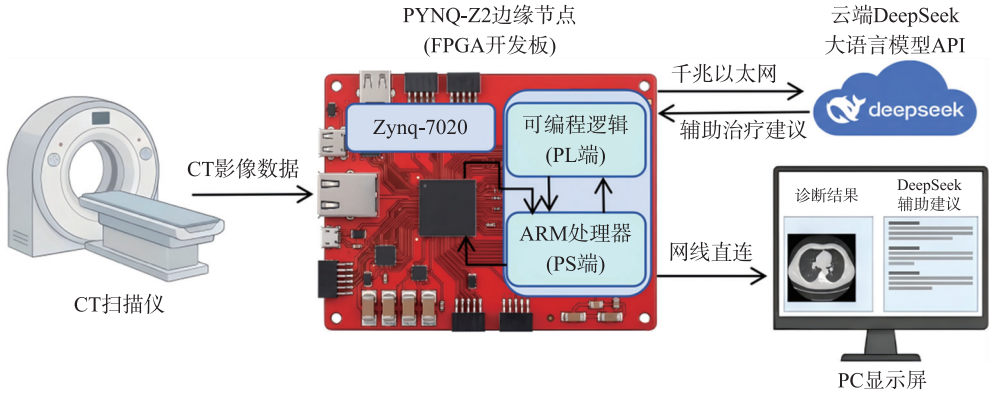

图 1　基于 FPGA 的边云协同辅助诊断系统总体架

cess,DMA)控制器,通过高带宽 AXI 接口流式传输至 PL 端。加速核利用流水线并行、循环展开及批归一化融合策略执行高并发卷积运算,快速输出病灶分类结果,并将推理结果回写至共享内存,完成边缘侧的实时病灶检测。

④边云协同辅助诊断:PS 端捕获边缘推理结果后,利用板载网络接口建立与云端的加密通信链路,调用 DeepSeek 大语言模型 API。系统将病理分类结果封装为结构化提示词发送至云端,由大模型结合医学知识库进行推理,自动生成包含病理特征分析、个性化用药指导及康复端建议的辅助诊疗报告。

⑤PC 端可视化交互:最终的影像分类结果、置信度数据及云端生成的诊疗建议通过网线实时传输至 PC 显示屏。终端软件对多模态数据进行渲染,直观呈现"影像-病灶-建议"的一体化诊断信息,辅助医生进行快速、准确的临床决策。

该架构充分利用了 FPGA 内部的高速互联特性,有效规避了海量影像数据在云端传输造成的延迟瓶颈,同时通过边云协同机制弥补了边缘设备认知能力不足的缺陷,显著提升了筛查系统的时效性与临床实用价值。

2 轻量化知识蒸馏网络模型构建

本章构建了一种基于知识蒸馏的轻量化卷积神经网络算法框架,旨在通过模型压缩技术实现计算效率与特征保真度的协同优化。文中将从网络架构与训练策略两个维度展开论述:2.1 节定义了"教师-学生"协同蒸馏网络的拓扑结构及跨维度特征适配机制;2.2 节推导了基于温度系数的混合损失函数优化策略,阐明利用软标签监督信号增强边缘模型泛化性能的数学原理。

2.1 教师-学生蒸馏网络架构设计

基于边缘设备存储带宽受限与高维医学影像处理的矛盾,本文提出一种面向 FPGA 硬件加速的"教师-学生"协同蒸馏架构(如图 2 所示)。该架构采用算法-硬件协同设计思想,在压缩计算开销的同时维持特征保真度。具体设计流程包含以下四个关键环节:

①图像预处理与维度适配:为适配 FPGA 片上块随机存取存储器(Block Random Access Memory,BRAM)存储深度并降低乘累加负载,采用双线性插值算法将原始 X 光影像重采样至 128×128。该操作在保留肺部磨玻璃影等关键病理结构的同时,有效滤除高频伪影噪声。针对教师网络 ResNet50 的输入规范,利用通道复制技术将单通道灰度图扩展为 RGB 格式,并通过像素归一化约束数值分布,既加速算法收敛,也为后续 FPGA 端的高精度定点化量化奠定数据基础。

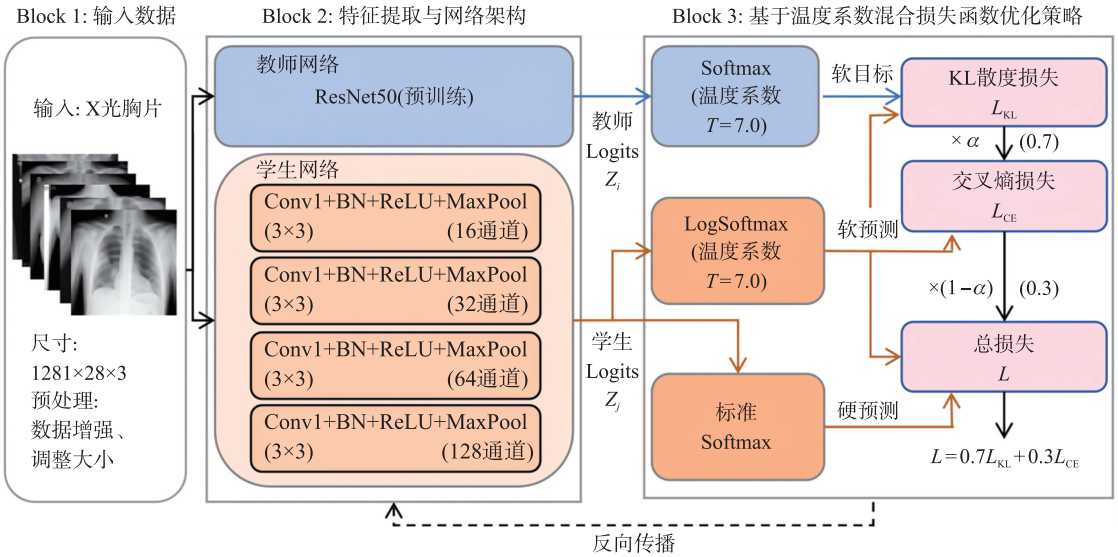

图 2 基于知识蒸馏的教师-学生网络架构示意图

②教师网络构建:选取在 ImageNet 大规模数据集预训练的 ResNet50 作为教师模型,利用其深层残差单元捕获病灶的高级语义特征及类别间的隐含关联。在协同训练阶段,冻结其权重参数,将输出的非归一化对数概率经温度缩放处理为"软标签"。与硬标签相比,软标签携带了丰富的类间相似度信息,为学生网络提供包含概率密度分布的高阶监督信号。

③学生网络设计:针对 PYNQ-Z2 平台的架构特性,定制设计 4 层轻量化卷积神经网络。摒弃不利于硬件并行的复杂残差跳连,采用流式计算友好的线性堆叠拓扑:每层依次级联卷积、批归一化

(BN)及最大池化算子。该精简结构极大地降低了逻辑资源占用,并便于在 HLS 处理中应用流水线与循环展开指令,从而最大化数据吞吐率。

④混合优化策略:引入温度缩放机制调控概率分布熵值,通过联合优化交叉熵损失与 KL 散度,迫使学生网络在拟合真实标签的同时,继承教师网络的泛化表征能力。

综上所述,该架构通过“预处理降维—深层特征提取—浅层架构模仿”的闭环设计,有效突破了边缘侧的算力瓶颈。实验表明,该方案在显著降低模型参数量与推理延迟的同时,保持了对微小病灶的鲁棒特征提取能力,实现了临床筛查精度与硬件能效的最优平衡[4]。

2.2　基于温度系数混合损失函数优化策略

针对传统轻量化模型在复杂病理特征识别中普遍存在的精度瓶颈,以及学生网络因层数缩减而导致的特征提取能力退化问题,本研究引入了基于温度缩放的混合损失函数优化策略。该策略旨在将预训练教师网络在特征空间的概率分布作为“软标签”,引导学生网络挖掘不同病灶类别间的隐含关联,从而增强其泛化表征能力。

在常规卷积神经网络训练中,标准 Softmax 激活函数倾向于将网络末端的非归一化对数概率映射为接近独热编码形式的陡峭分布。这种硬分布虽然明确指示了最大概率类别,却显著抑制了非正确类别中携带的细微概率结构(如不同肺炎亚型间的视觉相似度)。为此,算法引入温度系数 T 对输出分布进行平滑处理。对于输入样本 x,第 i 类别的软化概率 $p_i(x;T)$ 定义如下:

$$p_i(x;T)=\frac{\exp(z_i/T)}{\sum_{j=1}^{C}\exp(z_j/T)}$$

式中:$C=4$ 为肺部疾病类别总数。当设定 $T>1$ 时,指数函数的输出分布趋于平缓,使得负标签中蕴含的“暗知识”得以显现。为了在“模仿教师分布”与“拟合真实标签”之间取得平衡,总损失函数 L 构建为 KL 散度与标准交叉熵损失的加权组合:

$$L=\alpha\cdot T^2\cdot L_{\mathrm{KL}}(p^S,p^{\mathrm{T}})+(1-\alpha)\cdot L_{\mathrm{CE}}(p^S,y)$$

式中:第一项 L_{KL} 衡量学生网络软预测 p^S 与教师网络软目标 p^{T} 之间的分布差异,系数 T^2 用于补偿因温度缩放导致的梯度幅值衰减;第二项 L_{CE} 计算学生网络硬预测与真实标签 y 的交叉熵,保障分类的绝对准确性。

知识蒸馏的特征迁移效能高度依赖于温度(T)与软标签损失权重(α)的联合配置。由于这两个参数在寻优空间中存在深度的耦合关系,单一变量的控制法往往容易陷入局部最优。因此,本文采用网格搜索策略,结合医学影像特征分布规律与端侧轻量化网络的容量限制,全面论证了二者的最佳参数组合。

关于参数的搜索边界,根据经典知识蒸馏理论[5],一般 T 取 1 到 10 之间;而损失权重 α 通常限定在(0,1) 区间内。图 3 所示的二维超参数热力图直观展示了不同组合下学生模型的测试集准确率分布。从图中的准确率梯度演变可以发现:在 α 的维度上,由于复杂的 ResNet50 教师网络与极度精简的 4 层定制化 CNN 之间存在巨大的“模型容量鸿沟”,若 α 取值过小,总损失过度依赖真实硬标签,极易导致表征能力受限的浅层网络在训练后期出现严重的过拟合现象;反之,若设定过大,网络则会过度迎合教师的分布而丧失真实病理标签的基准约束力。在 T 的确定上,较高的温度能够使教师网络输出的 Softmax 概率分布更加平滑,从而释放出非目标类别中的“暗知识”。与常规自然图像分类任务惯用的中低温度不同,医学影像中的不同肺部病理特征往往呈现极高的类间相似性。采用偏高的温度有助于引导学生模型深度捕获不同病灶之间微弱却关键的关联特征,而非单纯拟合独热编码的绝对决策边界。

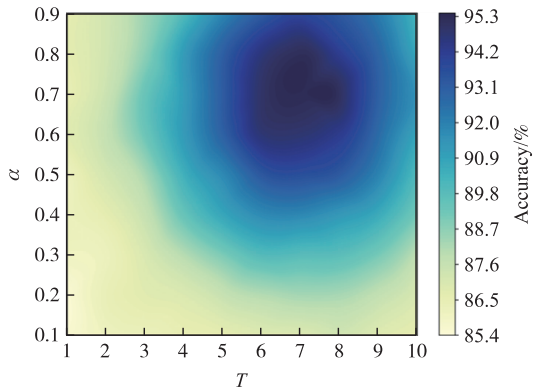

图 3　不同温度(T)与权重(α)组合下学生网络准确率的网格搜索热力图

热力图中颜色最深的峰值区域精准印证了上述理论研判:当在 $T=7.0$ 且 $\alpha=0.7$ 附近时,硬标签的分类指导与软标签的正则化平滑效应达到了全局最优平衡,模型识别准确率随之达到峰值(94.86%)。综上所述,本文最终确立 $T=7.0$ 与 $\alpha=0.7$ 作为蒸馏框架的超参数,以最大化医学特征的迁移效能。

3　边云协同低延迟图像识别诊断

本章致力于在 Zynq 平台上实现轻量化网络的硬件映射与系统集成,构建具备高能效比的边云协

同诊疗框架。文中将从架构设计等维度展开:3.1 节阐述流式计算的边云协同架构与数据机制;3.2 节详述面向 FPGA 的细粒度流水线并行与逻辑压缩策略;3.3 节介绍基于 DeepSeek 大模型的边云协同接口设计,实现从影像特征到临床语义的跨域融合。

3.1 基于流式计算的边云协同架构与数据机制

针对嵌入式端部署 CNN 面临的算力与带宽瓶颈,本文构建了基于流水线的软硬协同加速架构(如图 4 所示)。该系统依托 FPGA 平台,利用 AXI 互联矩阵实现处理系统(Processing System,PS)逻辑控制与可编程逻辑(Programmable Logic,PL)密集计算的深度解耦。架构遵循"控制流主导、数据流驱动"范式,通过 DDR3 共享存储体系结合硬件流水线级联技术,有效规避了板级总线通信延迟,实现了计算任务的高效运行与并行执行。

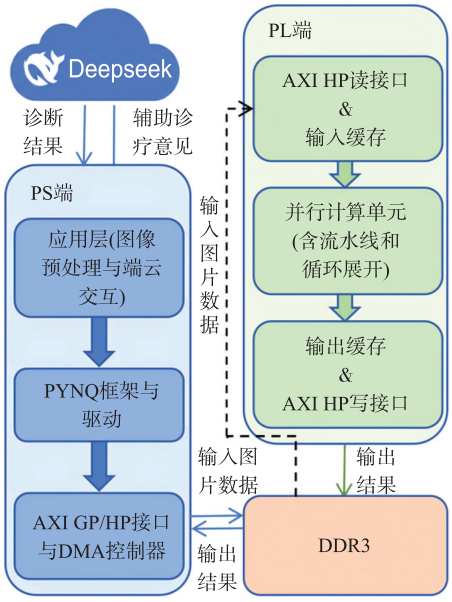

图 4 基于 Zynq 的软硬协同加速架构与数据交互机制

在软件控制域(Processing System,PS 端),双核 ARM 处理器承担任务编排与数据预处理。原始影像首先在应用层执行尺寸归一化与定点量化适配,以消除噪声并符合硬件规范。随后,主控程序通过 AXI-GP 接口配置直接存储器访问(Direct Memory Access,DMA)控制器的传输描述符表,并利用 AXI 总线向加速器下发网络权重与配置参数。系统就绪后,PS 端通过置位硬件使能信号精确触发加速逻辑,完成软硬件协同调度。

在硬件计算域(Programmable Logic,PL 端),集成了基于 HLS 定制的 CNN 加速 IP 核。该核心通过独立 AXI-HP 端口构建全双工高通量数据通路。任务执行时,图像数据经 DMA 引擎以突发传输模式加载至片上缓存,随即送入基于数据流优化的卷积阵列。各层运算利用循环展开与流水线技术级联处理,最终分类概率向量直接回写至共享内存。

边缘侧推理完成后,系统进入边云协同阶段。PS 端通过以太网接口读取诊断结果,将其封装为请求发送至云端 DeepSeek 大语言模型 API,以获取深层次病理分析与治疗建议,该过程实现了从底层特征计算到上层语义认知的跨域融合。

3.2 面向硬件加速的并行优化与逻辑压缩

针对卷积神经网络在通用处理器上逐层串行执行导致的高推理延迟,以及直接硬件映射面临的计算流水线停顿与资源分配不均问题,本研究在 IP 核设计阶段实施了数据流与计算流的深度重构。通过引入细粒度并行策略,在时空维度上最大化 FPGA 片上数字信号处理单元(Digital Signal Processor,DSP)与逻辑资源的利用率。具体优化措施包含以下三个维度:

①任务级流水线架构构建

如图 5 所示,传统卷积神经网络在通用处理器上的执行遵循严格的层级串行机制,即第 $L+1$ 层必须等待第 L 层特征图完全计算完毕并写回主存后方

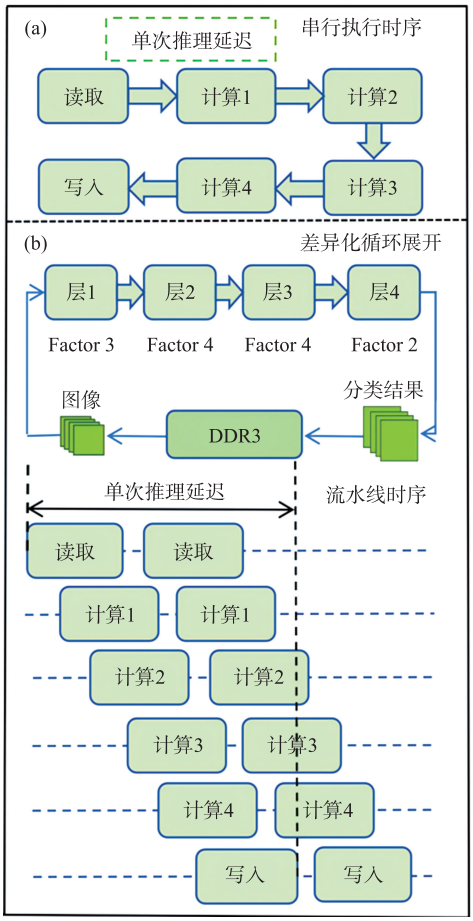

图 5 基于任务级流水线的并行架构与传统串行架构的时序对比

可启动,导致了严重的访存延迟与计算单元闲置。为打破此强时序依赖,本研究设计了基于片上级联通道的任务级流水线架构。该架构将粗粒度的层级同步解耦为细粒度的像素流驱动,在各计算层间建立深度的先进先出数据通路。一旦前置层输出首个有效特征像素,后继层即刻触发运算,从而实现了全网络算子的时间重叠并发执行,彻底消除了中间特征图的片外存储开销[6-7]。

②差异化循环展开策略

如图 5 所示,为解决不同网络层计算密度差异导致的流水线“气泡”问题,基于层级特征维度与计算负载的异构性分布,实施了适配层级特性的差异化循环展开策略,针对首层 RGB 输入,设定展开因子为 3 以精确匹配通道维度;针对计算负载最重的中间特征提取层,配置最大并行因子为 4,利用更多的 DSP 单元并行处理多通道数据以削减核心路径延迟;而针对末端层,为避免布线拥塞将因子动态调整为 2。这种混合粒度并行设计在有限的 FPGA 逻辑资源约束下,实现了计算吞吐率与硬件开销的最优平衡[8]。

③批归一化层融合与逻辑压缩

鉴于推理阶段模型参数的静态特性,采用离线重参数化技术将批归一化(BN)层的统计特性与仿射变换参数融入前置卷积层的权重与偏置中。该策略在硬件实现上消除了独立的 BN 模块,规避了 FPGA 不擅长的高昂浮点除法与开方运算逻辑。这不仅减少了 BRAM 与 DSP 的占用,更缩短了关键路径长度,显著提升了系统的主频稳定性。

综上所述,通过上述多维度的硬件重构,加速器成功实现了从“计算受限”向“访存带宽受限”的架构优化。实验表明,该设计在保证算法精度的前提下,将单帧推理延迟压缩至毫秒级,有效支撑了边缘侧的高并发实时诊断需求。

3.3　边云协同诊断与 DeepSeek API 集成

基于 Zynq 的异构计算优势,本研究构建了云边协同的辅助诊疗架构。当 FPGA 硬件加速器完成影像的高速推理后,PS 端主控程序立即启动数据聚合流程,将边缘侧输出的定量预测指标(病灶类别)与患者电子病历中的临床元数据(年龄、性别、基础病史)进行多模态融合。该机制有效解决了传统计算机辅助诊断系统仅输出离散分类标签、缺乏个性化背景支持的痛点,为上层认知模型的推理提供了完备的上下文输入。

为实现符合临床逻辑的辅助决策,系统集成了 DeepSeek 大语言模型 API,并设计了专用的结构化提示词工程。主控端基于结构化提示词模板,构建查询指令“作为呼吸科专家,针对[年龄]岁[性别]确诊为[类别]且伴有[基础疾病]的患者,请依据最新指南生成诊疗简报。”模型据此进行深度语义推理,反馈包含病理机制分析、个性化药物治疗方案及康复护理建议的文本报告。最终,云端语义结果与边缘侧视觉热力图实现对齐展示,构建了从“像素级计算”到“语义级决策”的完整闭环[9]。

4　实验结果与分析

4.1　实验设置

实验选用具有临床代表性的由卡塔尔大学和达卡大学的研究团队整理发布的 COVID-19 Radiography Database 作为数据基准[10]。该库包含 21 165 张高分辨率 X 光影像,涵盖正常、新冠肺炎、病毒性肺炎及细菌性肺炎四类典型样本。其中,新冠病例呈现显著的磨玻璃影特征,肺炎样本则表现为不同程度的肺实质病变。为确保评估客观性,数据集经标准化预处理后,严格遵循 8∶1∶1 比例随机划分为训练、验证及测试子集(分布见表 1)。

表 1　肺部疾病 4 分类数据集数据分布

类别	训练集(80%)	验证集(10%)	测试集(10%)	总计
正常	8 154	1 019	1 019	10 192
细菌性肺炎	4 810	601	601	6 012
新冠肺炎	2 892	362	362	3 616
病毒性肺炎	1 076	134	135	1 345
总计	16 932	2 116	2 117	21 165

模型训练基于 PyTorch 深度学习框架,输入图像归一化为 128×128,批次大小设为 32,总迭代轮次为 120。蒸馏温度系数设定为 $T=7.0$,软目标权重 α 取 0.7。量化评估指标选取精确率,并结合 FPGA 端的推理延迟、功耗等数据,综合衡量系统的边缘计算效能。

4.2　训练结果与实验分析

本实验基于轻量化 CNN 展开知识蒸馏训练,设总迭代轮次为 120,采用最优超参数组合($T=7.0$,$\alpha=0.7$)。学生模型的动态收敛特性如图 6、图 7 所示。图 6 显示,模型损失(Loss)在初期快速下降后,于第 100 轮左右平滑收敛至稳定区间(Loss<0.2),深刻验证了软标签对优化空间的正则化与平滑作用。图 7 表明,验证集准确度随训练稳健攀升,最终稳定于 94.86%。贯穿整个训练周期,训练集与验证集曲线始终保持高度一致,全程未现发散迹象。这表明在蒸馏机制的“暗知识”约束下,该极简网络有

效规避了针对复杂肺部病理特征极易发生的过拟合风险,展现出卓越的泛化表征与临床推理能力。

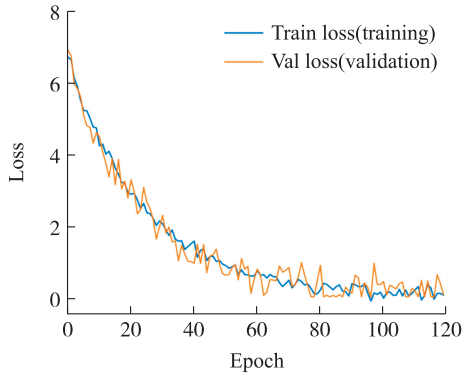

图 6 训练集与验证集的损失函数(Loss)变化趋势

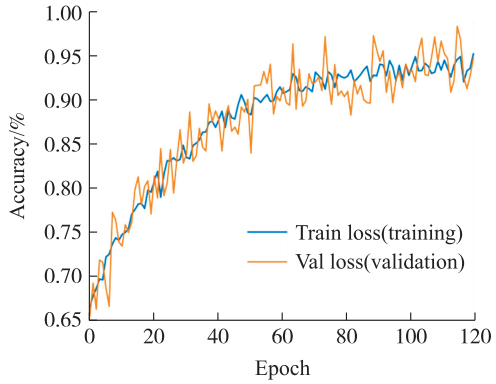

图 7 训练集与验证集的准确度(Accuracy)变化趋势

为量化知识蒸馏策略的有效性,本研究进行了蒸馏前后实验对比,结果如图 8 所示。在缺乏教师监督的基准条件下,仅依靠硬标签训练的轻量化模型受限于浅层结构,测试集准确率仅为 85.68%。

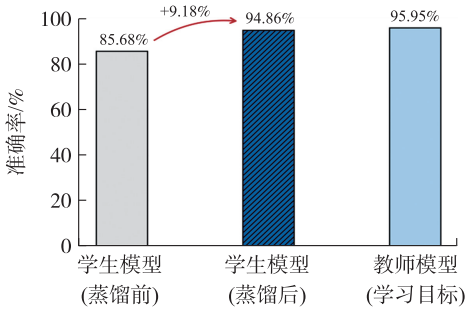

图 8 知识蒸馏策略对模型准确率影响的消融实验

在引入预训练 ResNet50(准确率为 95.95%)进行蒸馏后,学生模型的准确率显著跃升至 94.86%。这 9.18%的性能增益表明,学生模型成功“继承”了教师模型对微小病灶特征的判别能力,以极小的参数代价实现了逼近深层大模型的性能。

此外,测试集混淆矩阵如表 2 所示。模型在“正常”与“新冠肺炎”类别上表现出极高的敏感性。尽管病毒性与细菌性肺炎因纹理相似存在少量混淆,但整体分类性能稳健,满足临床快速筛查的需求。

表 2 测试集上的分类混淆矩阵

实际类别	预测类别			
	正常	新冠肺炎	病毒性肺炎	细菌性肺炎
正常	987	5	8	19
新冠肺炎	3	354	2	3
病毒性肺炎	4	1	115	15
细菌性肺炎	18	3	28	552

4.3 系统性能评估

为验证微架构优化策略的有效性,在综合环境下执行了硬件消融实验,结果见表 3。未优化的方案由于遵循严格的指令串行逻辑,单帧耗时高达 220.5 ms,无法满足实时诊查需求。通过引入任务级流水线与差异化循环展开,系统有效解耦了层间时序依赖并极大释放了片上并行算力,使延迟压缩至 52.1 ms。最终方案在融入 BN 层融合策略后,消除了逻辑冗余,在资源占用下降 7%的同时,实现了 5.6 倍的总加速增益,体现了资源与性能的最优均衡。

表 3 硬件加速策略消融实验结果对比

实验方案	耗时/ms	性能增益
方案 1(无优化)	220.5	1.0
方案 2(+流水线)	65.2	3.4
方案 3(+循环展开)	52.1	4.2
方案 4(完整方案)	39.0	5.6

在系统性能综合评估维度,表 4 详尽对比了本研究方案与现有主流诊断平台的关键量化指标。实验数据表明,系统单帧端到端推理延迟仅为 39 ms,动态功耗稳定在 1.5 W。与 Jetson Nano 等嵌入式 GPU 平台相比,本系统在功耗大幅降低 70%的前提下,推理速度提升了近 4 倍,其能效比增益更是达到了 10 倍以上。这一结果有力证明了专用硬件流水线架构在突破通用指令集架构的“访存墙”瓶颈方面具有显著优势,能够以极低能耗实现极高的计算密度。

横向对比其他前沿 FPGA 加速研究,本系统在执行复杂的病灶四分类任务时,依然维持了 94.86%的高分类精度,充分验证了知识蒸馏对学生网络表征能力的提升效果。在算力调度方面,本研究通过流式并行架构实现了 22.38 Mpixels/s 的高吞吐量。更为关键的是,系统集成了 DeepSeek 云端语义认知模块,弥补了现有方案仅输出离散标签的不足,构建了闭环的辅助诊疗体系。综上所述,该系统在检测精度、实时响应与低功耗维度上实现了最佳平衡,契合基层医疗设备的便携化应用。

表 4　本研究与现有肺部疾病检测系统的详细性能指标对比

性能指标	Bae et al.[11]	Swaminathan et al.[12]	Wang et al.[13]	Guddati et al.[14]	本研究
硬件平台	Jetson Nano (Embedded GPU)	Jetson Nano (Embedded GPU)	Zedboard (Zynq-7020)	PYNQ-Z2 (Zynq-7020)	PYNQ-Z2 (Zynq-7020)
任务类型	3 分类	3 分类	2 分类	3 分类	4 分类
硬件频率/MHz	600	1000	100	100	100
LUT	N/A	N/A	5,266	28,154	27,059
FF	N/A	N/A	7,128	36,214	35,289
DSP	N/A	N/A	82	112	139
BRAM	N/A	N/A	69	62.5	84
推理延迟/ms	130	180	48	41	39
功耗/W	5	10	1.72	1.5	1.5
帧率/FPS	7.7	5.5	20.8	24.4	25.6
吞吐量/(Mpixels/s)	–	–	10.04	15.7	22.38
能效比/(FPS/W)	1.54	0.55	12.09	16.3	17.0
检测准确率/%	91.74	91.60	88.43	93.16	94.86

4.4　诊断与辅助治疗效果展示

为验证所提系统在实际医疗场景中的工程效能,本研究搭建了边云协同的硬件实验平台进行集成测试。如图 9 所示,系统以 PYNQ-Z2 开发板为边缘计算核心,通过千兆以太网接口与上位机终端建立全链路通信。实物图中清晰展示了硬件加速器在板载运行时的状态,上位机实时呈现可视化交互界面。界面左侧同步显示待测 X 光影像及其病灶定位框,右侧区域则动态刷新由 DeepSeek 大模型生成的结构化诊疗报告,涵盖病理分析、用药建议及康复指导。该实验结果证实,系统具备稳定的端到端推理能力与毫秒级响应速度,成功实现了从影像采集、边缘计算到云端认知服务的完整闭环。

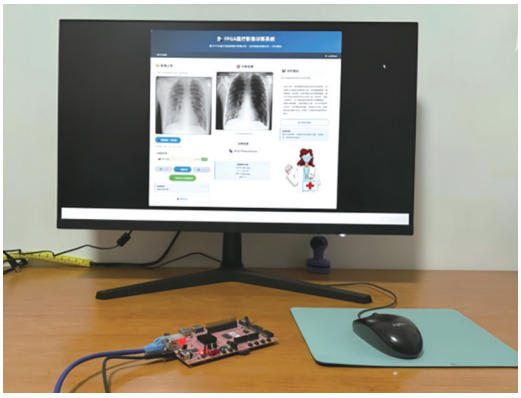

图 9　实物图

5　结论

本研究针对肺部疾病筛查在边缘计算场景中面临的算力受限与实时性不足问题,提出了一种基于 Zynq 的软硬协同智能辅助诊断系统。算法层面,构建“教师-学生”知识蒸馏架构,将 ResNet50 的高维语义特征迁移至轻量化 CNN 模型,在大幅压缩参数量的同时对四类病灶实现了 94.86% 的高精度识别[15];硬件架构上,基于 HLS 技术设计了任务级流水线与批归一化(BN)融合加速器,通过循环展开与数据流优化显著降低了资源开销。实测表明,该系统在 PYNQ-Z2 平台上单帧推理仅 39 ms,动态功耗低至 1.5 W,能效比优于同类纯软件方案。此外,通过调用 DeepSeek API 实现了从影像检测到诊疗建议的闭环。该系统有效突破了边缘侧性能瓶颈,为基层医疗提供了高可靠、低延迟的工程方案。

面向复杂的实际临床部署,本系统仍存在优化空间,未来研究将聚焦以下三个方向:一是引入量化感知训练探索 INT8 或混合精度量化,进一步降低边缘端对片上 BRAM 的存储依赖;二是针对当前循环展开受限于 DSP 物理资源的问题,拟引入 Winograd 快速卷积算法,并深度优化 AXI 总线的突发传输机制,以在有限资源下最大化计算并行度与访存效率[8];三是结合联邦学习重构边云协同机制,在避免原始敏感影像上传的前提下实现模型的联合演进[16]。本研究不仅提供了一种轻量化医学影像加速基准,也为下一代高隐私、高能效的医疗物联网系统奠定了工程基础。

参考文献:

[1]　李洲诚,张毅,孙晋. 面向边缘智能应用的多出口深度神经网络随机优化方法[J]. 计算机科学,2025,52(4):85-93.

[2]　包振山,郭俊南,张文博,等. UltraAcc:基于 FPGA 流水架构的低功耗高性能 CNN 加速器定制设计[J]. 计算机学报,2023,46(6):1139-1155.

[3]　唐智贤,王一淼,周靓怡,等. 人工智能技术在肺部影

像辅助诊断中的应用进展[J]. 中国医学物理学杂志,2022,39(5):655-660.

[4] 司兆峰,齐洪钢. 知识蒸馏方法研究与应用综述[J]. 中国图像图形学报,2023,28(9):2817-2832.

[5] Gou J,Yu B,Maybank S J,et al. Knowledge Distillation: A Survey[J]. International Journal of Computer Vision,2021,129(6):1789-1819.

[6] 秦华标,曹钦平. 基于 FPGA 的卷积神经网络硬件加速器设计[J]. 电子与信息学报,2019,41(11):2599-2605.

[7] 季鹏辉,沈佳波,吴俊. 基于 FPGA 的 DDR3 读写操作实现与研究[J]. 电子器件,2025,48(5):999-1007.

[8] 牛朝旭,孙海江. 基于 FPGA 的 Winograd 算法卷积神经网络加速器设计与实现[J]. 液晶与显示,2023,38(11):1521-1530.

[9] 张文柱,余静华. 移动边缘计算中基于云边端协同的任务卸载策略[J]. 计算机研究与发展,2023,60(2):371-385.

[10] Chowdhury M E H,Rahman T,Khandakar A,et al. Can AI Help in Screening Viral and COVID-19 Pneumonia?[J]. IEEE Access,2020,8:132665-132676.

[11] Bae H J,Kim J,Jeong D. Real-Time Fluorescence-Based COVID-19 Diagnosis Using a Lightweight Deep Learning System[J]. Sensors,2026,26(1):339.

[12] Swaminathan T P,Silver C,Akilan T,et al. Benchmarking Deep Learning Models on NVIDIA Jetson Nano for Real-Time Systems: An Empirical Investigation[J]. Procedia Computer Science,2025,260:906-913.

[13] Wang J,Ikechukwu O M,Dang K N,et al. Spike-Event X-Ray Image Classification for 3D-NoC-Based Neuromorphic Pneumonia Detection[J]. Electronics,2022,11(24):4157.

[14] Guddati P,Dash S,Tripathy R K. FPGA Implementation of the Proposed DCNN Model for Detection of Tuberculosis and Pneumonia Using CXR Images[J]. IEEE Embedded Systems Letters,2024,16(4):445-448.

[15] Sriwiboon N. Efficient and Lightweight CNN Model for COVID-19 Diagnosis from CT and X-Ray Images Using Customized Pruning and Quantization Techniques[J]. Neural Computing and Applications,2025,37:13059-13078.

[16] 金亚东,范菁,郗恩康,等. 基于联邦学习的空天地一体化通信技术研究综述[J]. 计算机应用研究,2025,42(11):3215-3224.

张宇涵(2004—),男,汉族,广西南宁人,东南大学电子科学与工程学院 2023 级本科生,253408368@qq.com;

王亚鹏(2004—),男,汉族,江苏扬州人,东南大学电子科学与工程学院 2023 级本科生,1013803100@qq.com;

张 萱(2004—),女,汉族,广东广州人,东南大学吴健雄学院 2023 级本科生,3049364290@qq.com;

孙 豪(1996—)男,汉族,东南大学电子学院电路与系统系,讲师,硕导,博士,主要研究方向为 FPGA 高能效计算、边缘智能与芯片系统设计,重点聚焦于智慧交通和机器人多模态智能感知等前沿应用,hao.sun@seu.edu.cn。