第 43 卷 第 13 期/2023 年 7 月/光学学报 研究论文

光学学报

基于轻量化方向 Transformer 模型的肺炎 X 光片辅助诊断

周涛[1,2]，叶鑫宇[1,2*]，刘凤珍[1,2]，陆惠玲[3]

[1]北方民族大学计算机科学与工程学院，宁夏 银川 750021；

[2]北方民族大学图像图形智能处理国家民委重点实验室，宁夏 银川 750021；

[3]宁夏医科大学医学信息与工程学院，宁夏 银川 750004

摘要 为满足轻量化卷积神经网络(CNN)对肺炎 X 光片中方向和语义信息提取的需求，提出一种基于轻量化方向 Transformer 的肺炎 X 光片辅助诊断模型。首先，构造 CNN 结合 Transformer 的密集连接架构，实现深浅层中局部和全局信息的结合；其次，设计方向卷积捕获不同大小、形状特征的空间和方向信息，并降低 Transformer 学习全局特征的计算复杂度；然后，为每个样本特征采用专门的卷积核，降低方向卷积参数量，并保持高效计算；最后，通过构造均衡聚焦损失函数来提高模型肺炎识别能力。在肺炎 X 光片数据集中，所提出模型以较低的模型参数量、计算量，以及较短的运行时间，获得了 98.87% 准确率和 98.85% AUC 值的最佳性能，在 3 个公共肺炎相关数据集中均获得较强的鲁棒性和较优的泛化能力。

关键词 图像处理；密集局部和全局特征；方向 Transformer；轻量化卷积；肺炎 X 光片

中图分类号 TP391.41 **文献标志码** A **DOI**：10.3788/AOS230447

1 引 言

肺炎是最常见的传染性疾病之一，早期的准确识别对于肺炎诊断和治疗至关重要，然而肺炎检测不仅取决于医疗技术，还取决于放射科医生的经验[1]。X 光片成本低且易于获取，成为诊断肺炎疾病最常见的分析方法，手动分析肺炎 X 光片耗时长且可靠度不高[2]，而基于深度学习和医学影像的计算机辅助诊断肺炎取得较高的精准度。收集了 10 万张肺炎 X 光片的数据集 ChestX-ray8[2]被公开并使用卷积神经网络(CNN)进行分析和识别。由迁移学习[3]和 CNN 创建的肺炎 X 光片检测模型[4]具有较好的性能。通过不同扩张率的跳跃卷积改进的 GhostNet[5] X 光片肺炎识别模型，融合了高层次语义和低层次细节信息，并获得了 99.66% 准确率。在 X 光片胸部成像过程中，患者体位和吸气深度等因素导致肺炎的成像结果容易与其他疾病的成像结果混淆，并且现有的方法忽略了肺部 X 光片中影像的方向特征，如肺炎常见的发病部位在肺的中叶和下叶，通过方向特征可以较好地识别出肋骨和肺部区域。此外，充分提取空间信息模型也更易识别出肺炎病灶。

李翔等[6]基于空间注意力机制的深度学习模型获得了 2.91% 的精度提升，但基于 CNN 的模型难以学习像素间的全局关系。Transformer 可更好地对全局信息进行建模[7]：Okolo 等[8]提出增强 Transformer 分类模型，在 4 个肺炎 X 光片数据集中获得了近 3% 的精度提升；Park 等[9]将利用 DenseNet 提取肺炎 X 光片中的特征再嵌入到 Transformer，获得较好肺炎诊断性能；Yuan 等[10]对图像的局部结构信息与全局相关性进行同时建模，获得较高的计算效率和较优的性能；Liang 等[11]利用 CNN 提取局部特征，同时利用 Transformer 提取全局特征，并进行特征互补融合以提高模型性能，而更好的结合方式可以更充分地提取肺炎 X 光片的特征信息。采用更大数据集和设计更复杂结构可以提高模型性能，但都会增加计算和存储资源消耗，限制其在特殊场景中的应用。DenseNet 结合分组卷积可以轻量化，但性能难以提高[12]。Chen 等[13]对并行的 CNN 局部特征和 Transformer 全局特征进行双向融合，以较少的计算量实现了较高的精度。Mehta 等[14]将 CNN 中的卷积替换为少量参数的 Transformer 全局操作，构建出轻量、高效的模型。

现有模型在提取肺炎 X 光片的方向信息和全局语义信息方面存在一定难度，轻量化程度不高，为此，本文提出一种基于轻量化方向 Transformer

收稿日期：2023-01-06；**修回日期**：2023-02-10；**录用日期**：2023-02-21；**网络首发日期**：2023-03-09

基金项目：国家自然科学基金（62062003）、宁夏自然科学基金（2022AAC03149）

通信作者：*3303626778@qq.com

(LDTransformer)的肺炎X光片辅助诊断模型。首先，构造了一种CNN结合Transformer的密集连接架构，分4个阶段学习肺炎特征，使用密集连接实现深浅层中局部和全局信息的结合，学习更多可区分的特征；其次，针对肺炎X光片中影像的方向特征提取困难问题，设计横向、纵向和扩张卷积并行的结构，对不同大小、形状特征的空间和方向信息进行捕获；然后，针对模型轻量化程度不高的问题，在CNN部分设计轻量化卷积，将输入特征图分块处理并使用通道注意的混合器进行融合，改进所有样本特征共享的普通卷积，为每个样本特征采用专门的卷积核，增强网络特征提取能力并保持高效的计算效率；最后，构造均衡聚焦损失函数，提高小样本和错误分类样本的权重，并降低过分类样本的权重，提高模型性能和稳定性。

2　本文模型

所提出的LDTransformer整体框架如图1所示，分4个阶段学习肺炎特征，每个阶段采用CNN局部特征提取和Transformer全局特征提取交叉堆叠并进行密集连接，CNN部分采用轻量化方向卷积，最终特征利用全连接和Softmax分类层进行识别。本节将详细介绍模型的具体结构。

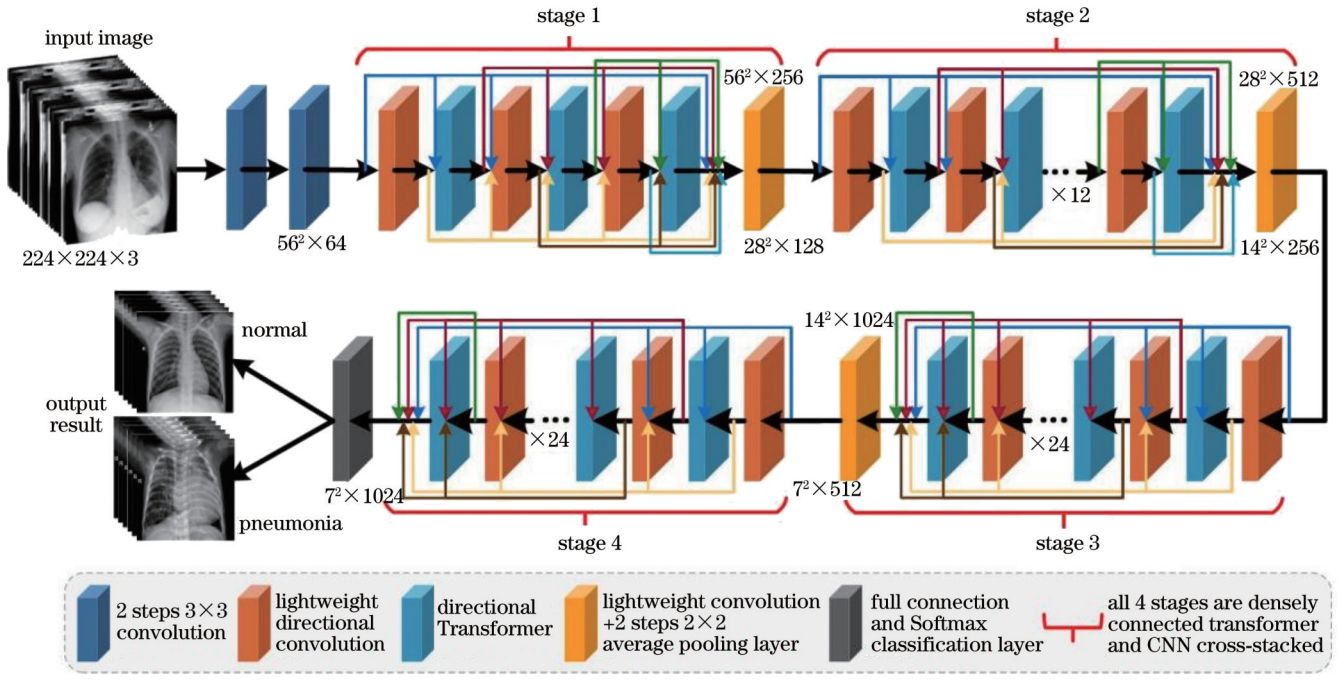

图1　LDTransformer整体框架结构图

Fig. 1　Overall frame structure of LDTransformer

2.1　方向Transformer

CNN通过卷积核对局部信息进行学习和提取，而肺炎X光片影像中的方向特征难以捕获，且CNN缺乏学习全局上下文信息的能力。Naseer等[7]研究发现Transformer在提取全局特征时具有更强的鲁棒性，但模型参数量和计算量都较大，训练过程中需要大量的内存、耗时较长，且难以充分学习浅层特征，会忽略部分小病灶。为此，本文构造了一种CNN结合Transformer的密集连接架构，如图1所示。将串行级联CNN与Transformer进行密集连接，结合CNN捕获局部信息和Transformer捕获全局上下文信息的优势，同时使用密集连接方式将局部信息流和全局信息流传递到后续层，实现在深浅层特征中对局部和全局信息进行结合，学习更多可区分的特征。对于输入特征图$\boldsymbol{x}$，在层数固定为N的前提下，传统Transformer中第一层梯度信息可以简易表示为

$$\frac{\partial L}{\partial \boldsymbol{w}_1}=\prod_i^N \boldsymbol{w}_i \times \boldsymbol{x}, \tag{1}$$

式中：$\boldsymbol{w}_i$表示权重矩阵；L为特征矩阵。通常情况下，$\boldsymbol{w}_i$的数值比较小，网络顶层的参数很难更新，导致梯度消失。使用残差连接的Transformer中第一层梯度信息变为

$$\frac{\partial L}{\partial \boldsymbol{w}_1}=\prod_i^N (\boldsymbol{w}_i+1)\times \boldsymbol{x}。 \tag{2}$$

残差连接Transformer在梯度信息中始终有加权值1，可以避免梯度消失问题，但当多次连乘时，归一化层会导致加权值变成一个系数，仍然会存在梯度消失问题。密集连接的Transformer中第一层梯度信息变为

$$\frac{\partial L}{\partial \boldsymbol{w}_1}=\prod_i^N [\boldsymbol{w}_i \times \boldsymbol{x}, \boldsymbol{x}], \tag{3}$$

式中：$[\boldsymbol{w}_i \times \boldsymbol{x}, \boldsymbol{x}]$为特征图进行通道拼接。密集连接Transformer通过将前续层信息传递给后续层来实现

特征重用，浅层与深层特征以拼接方式进行融合，残差相加导致信息破坏的现象不会发生，且每层直接利用梯度以及初始输入信息，进一步减轻梯度消失和网络退化问题。针对网络浅层进行 Transformer 全局特征相关性计算时存在的资源消耗多和特征提取能力弱的问题，设计了图 2 所示的方向 Transformer。

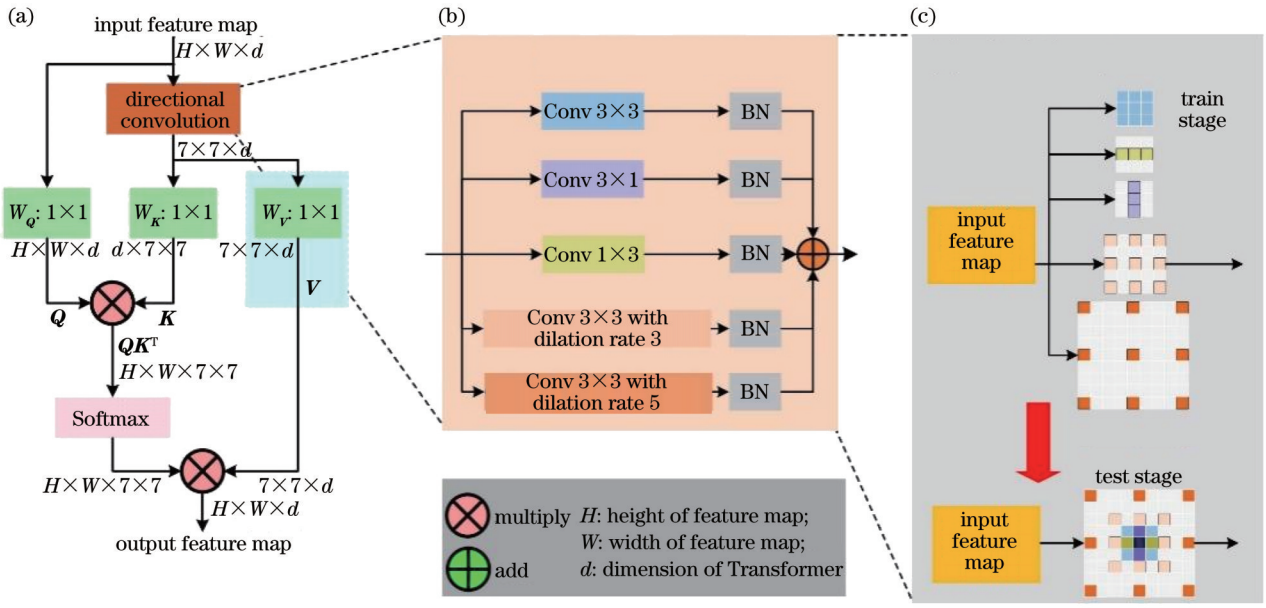

图 2　方向 Transformer 结构。(a)方向 Transformer 层；(b)方向卷积；(c)计算过程

Fig. 2　Structure of directional Transformer. (a) Directional Transformer layer; (b) directional convolution; (c) calculation process

如图 2(a)所示，Transformer 通过缩放点积[7]和 Softmax 函数对 1×1 卷积获得的查询($\boldsymbol{Q}$)和键值($\boldsymbol{K}$)进行计算，生成注意力权重并作为输入特征值($\boldsymbol{V}$)，利用方向卷积将输入特征的空间尺寸压缩到最小分辨率，然后通过 1×1 卷积获得键值($\boldsymbol{K}$)和特征值($\boldsymbol{V}$)。最终方向 Transformer 输出特征图 $\boldsymbol{X}$ 的计算式为

$$\boldsymbol{X}=\mathrm{Softmax}\left(\frac{\boldsymbol{QK}^{\mathrm{T}}}{\sqrt{d_K}}\right)\boldsymbol{V}, \tag{4}$$

式中：d_K 为键值($\boldsymbol{K}$)的维度。如图 2(b)所示，方向卷积由 3×3、1×3、3×1，以及扩张率为 3 和 5 的 3×3 卷积核并行而成，五分支结构可以拟合更多信息，其中：1×3 横向和 3×1 纵向卷积核用于捕获 X 光片中的方向信息；扩张率为 3 和 5 的卷积核在捕获方向信息的同时具有更大的感受野，也可以学习 X 光片中不同形状、大小的空间信息。方向卷积将浅层特征尺度压缩到最低分辨率，从而降低 Transformer 计算复杂度，提升网络对图像旋转和翻转的鲁棒性，提高语义判别能力和缓解类别混淆。图 2(b)中的“+”表示对 5 个分支输出特征相加，多分支结构通过拟合更多信息来提高特征表达能力，在训练结束后 5 个分支卷积核直接相加融合，由于融合时的累加过程是线性的，相关结果不会改变，即卷积操作具有可叠加性，最终推理阶段使用与 3×3 卷积核相同的计算量，可获得更强的特征提取能力。方向卷积在推理阶段的具体计算过程如图 2(c)所示，将 BN 的参数与卷积核的偏置项结合，然后将融合核和偏置项相加，得到一个单层，从而实现 BN 和多分支融合。

2.2　轻量化 CNN 卷积

基于 CNN 的方法或者基于 Transformer 的方法提高性能，主要通过更大数据集和更复杂的网络设计，无疑都增加了计算成本消耗，占用的计算资源和存储资源都较大。本实验在 Transformer 部分采用方向卷积压缩图像尺寸。尽管 CNN 中现有的深度可分离卷积、瓶颈结构、分组卷积、通道混洗等轻量化技术可以较好地降低模型参数量，但忽略了模型的计算效率，如随着分组卷积中分组数的增多，模型计算耗时的缩减程度远不如参数量的缩减程度。为了尽可能降低计算资源消耗，在 CNN 部分设计轻量化卷积，以提高方向卷积中各分支卷积的计算效率，将输入特征图进行分块逐样本卷积处理并使用混合器进行融合，结构如 3 所示。不同于所有样本特征共享的普通卷积，轻量化卷积为每个样本特征采用专门的卷积核，有效增强了网络的特征提取能力，同时保持了高效的计算效率。其中，混合器基于通道注意力设计，沿着特征图的空间维度进行平均计算，以此生成通道注意图，强调包含肺炎特征的通道，并减弱特征激活度低的通道。

如图 3(a)所示，普通卷积的计算式为

$$\boldsymbol{X}=\boldsymbol{w}\times\boldsymbol{x}+\boldsymbol{b}, \tag{5}$$

式中：$\boldsymbol{x}$ 和 $\boldsymbol{X}$ 分别为输入和输出特征图；$\boldsymbol{w}$ 和 $\boldsymbol{b}$ 为可学习参数，这两个可学习参数的维度为 $H_c\times W_c\times C\times O$，其中 H_c 和 W_c 分别为卷积核的高度和宽度，C 和 O 分别为输入和输出特征图的通道数。普通卷积中每个

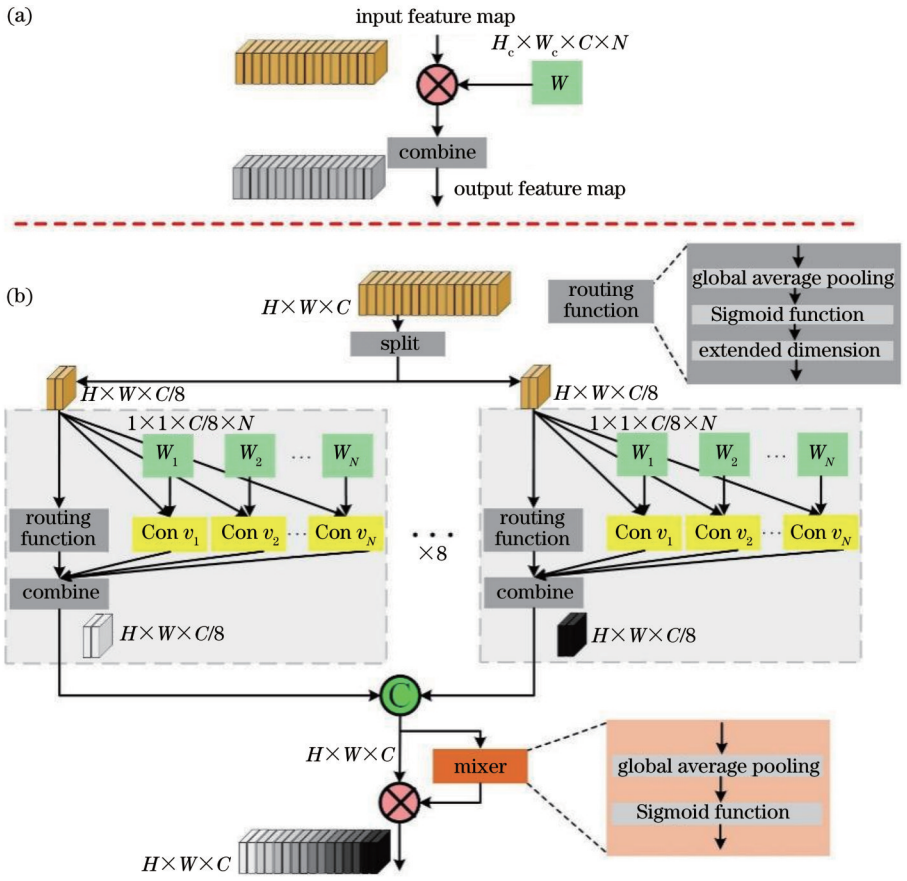

图 3　轻量化卷积结构。(a)普通卷积;(b)轻量化卷积

Fig. 3　Structure of lightweight convolution. (a) General convolution; (b) lightweight convolution

参数都需要与输入特征进行乘和加,而轻量化卷积是在使用卷积之前,将每个样本特征的卷积核计算分解为N个部分,分别计算再进行线性组合,这种卷积核计算方式同样需要应用于输入图像的每个像素,而这样操作可以使每个卷积核只需要并行计算一次,因此计算效率很高。对于输入特征图$\boldsymbol{x}$,轻量化卷积的计算式为

$$\boldsymbol{X}=(\alpha_1\boldsymbol{w}_1+\alpha_2\boldsymbol{w}_2+\cdots+\alpha_N\boldsymbol{w}_N)\times\boldsymbol{x}。\tag{6}$$

可将式(6)改写为图 3(b)所示的形式,即

$$\boldsymbol{X}=\alpha_1(\boldsymbol{w}_1\times\boldsymbol{x})+\alpha_2(\boldsymbol{w}_2\times\boldsymbol{x})+\cdots+\alpha_N(\boldsymbol{w}_N\times\boldsymbol{x}),\tag{7}$$

式中:α_i,$i\in[1,N]$为具有可学习参数的路由函数,用于学习每个样本的标量权重;N为单次运算卷积核的数量,每个卷积核$\boldsymbol{w}_i\left(i\in[1,N]\right)$与普通卷积中的卷积核具有相同维度。增加一个额外的参数时仅需要一个额外的乘加运算,所需推理成本较小,因此可通过增大N来增强网络特征提取能力。路由函数需要能够有效区分输入样本和保持高计算效率,因此使用全局平均池化(GAP)、全连接层(FC)、Sigmoid激活函数直接从输入特征图中计算出依赖于样本的路由权重。其中全连接层将池化后的输入映射成N的权重,常规卷积运算只在局部感受野上进行,而路由函数允许使用全局上下文信息对局部运算进行自适应。α_i的计算公式为

$$\alpha_i=\mathrm{Sigmoid}\left\{f_{\mathrm{FC}}\left[f_{\mathrm{GAP}}(\boldsymbol{x}_i)\right]\right\}。\tag{8}$$

为使CNN卷积进一步轻量化,借鉴分组卷积的思路,将输入特征图按通道分为8个部分,分别进行特征提取,并使用混合器进行组合,自适应增强包含肺炎特征的通道。在CNN的方向卷积中使用轻量化技术,可以使网络更高效地学习肺炎X光片的方向特征。

2.3　均衡聚焦损失函数

Kim 等[15]在深度网络中使用具有合页损失的ReLU函数,在平滑决策边界、平滑条件类和边际条件3种情况下都能实现模型快速收敛。聚焦损失[16]是在交叉熵损失上增加一个调制因子,对简单类别的损失贡献进行衰减,均衡网络对不同难易程度类别的学习,但忽略了过分类样本对模型性能的影响。为此,本实验构造了均衡聚焦损失函数,该函数可提高模型性能,降低整体分类误差,确保模型鲁棒性。聚焦损失中p_{t}是当真实标签为正类的概率,r为增加权重的调制因子,本文参考文献[16],选择$r=2$,则聚焦损失函数的计算公式为

$$L_1=-(1-p_{\mathrm{t}})^r\log(p_{\mathrm{t}})。\tag{9}$$

合页损失常用于最大间隔分类任务,将预测误差控制在$-1\sim1$范围内,让样本刚好能正确分类。当样本与分割线的距离超过1时,不对其进行反向传播,从

而降低过分类样本的权重，此时合页损失的计算公式为

$$L_2=\max(0,1-p\times t), \tag{10}$$

式中：t为真实标签，正类设置为1，负类设置为0。

均衡聚焦损失使用均衡因子λ对聚焦损失和合页损失进行整合，可以提高小样本类别和错误分类样本的权重，同时降低过分类样本的权重，使最终分类器更专注于降低整体分类误差。均衡聚焦的计算公式为

$$L_3=\lambda\times\max(0,1-p\times t)-(1-p_t)^r\log(p_t)。\tag{11}$$

3 实验和讨论

3.1 数据集和实验指标

本实验使用的数据集来自广州市妇幼保健院的ChestXRay2017[17]和来自X光片数据集[18]，其中ChestXRay2017数据集包含1583幅正常图像和4273幅肺炎图像，X光片数据集包含2313幅正常图像和2313幅肺炎图像。按6∶2∶2的比例把样本图像分成训练集、验证集和测试集进行实验。本次实验环境为Windows Server 2019系统，其内存为256 GB，搭载两块3 GHz的36核处理器，并采用两块泰坦第Ⅴ代显卡，基于GPU的Pytorch框架搭建网络，使用Adam优化器进行优化，采用0.01的初始学习率和每10周期衰减0.9的策略，设置权重衰减值为1×10^{-4}，训练周期为250，训练批处理大小为48。

根据模型预测结果分类错误和正确的个数，得到真正(TP)类、假正(FP)类、假负(FN)类和真负(TN)类。准确率(A)为全部类预测正确的比例，精确率(P)为正类且模型预测正确占所有正类的比例，召回率(R)为模型预测的正类占所有正类的比例，F_1分数的计算公式为

$$F_1=2\times\frac{P\times R}{P+R}。\tag{12}$$

ROC曲线以敏感度即真正类率(TPR，即召回率)为纵轴、假正类率(FPR)为横轴，将ROC曲线下方的面积定义为AUC，ROC曲线越靠近左上角，AUC值越大，表示模型的排序和分类性能越好。所有评价指标均是值越大表示模型性能越好。FPR、特异度(TNR)可表示为

$$\kappa_{\mathrm{FPR}}=\frac{N_{\mathrm{FP}}}{N_{\mathrm{FP}}+N_{\mathrm{TN}}}, \tag{13}$$

$$\kappa_{\mathrm{TNR}}=\frac{N_{\mathrm{TN}}}{N_{\mathrm{FP}}+N_{\mathrm{TN}}}, \tag{14}$$

式中：N_{FP}、N_{TN}分别为FP类、TN类样本的数量。

3.2 消融实验与分析

为了评估模块的有效性，在DenseNet121基础上依次进行5组实验：在消融实验1中交叉添加Transformer；在消融实验2中，将Transformer改进为方向Transformer；在消融实验3中，使用分组卷积进行轻量化对比；在消融实验4中，使用轻量化方向卷积；在消融实验5中，将实验4的轻量化方法扩展到整个网络。5组实验结果的对比如表1所示，不同模型的热力图如图4所示，其中：红色伪彩程度越深，表示网络对这个区域的关注度越高；括号内数字序号为识别错误样本的序号。

表1 消融实验的具体结果

Table 1 Specific results of ablation experiments

Model	Parameter amount	Calculation amount	Total time /s	A /%	AUC /%	R /%	F_1 /%	P /%
Base	6.96×10^6	2.87×10^9	16670	93.44±2.1	93.60±2.0	92.69±1.6	94.33±1.7	96.04±1.8
1	5.09×10^6	2.18×10^9	17285	94.30±3.3	94.26±3.5	94.49±3.4	95.14±3.3	95.78±3.2
2	4.68×10^6	1.81×10^9	16381	95.83±2.3	95.63±2.2	96.74±2.2	96.47±2.3	96.19±2.4
3	3.75×10^6	1.27×10^9	15564	95.49±2.7	95.44±2.8	95.73±2.8	96.16±2.7	96.59±2.6
4	1.26×10^6	4.95×10^8	13479	97.22±2.0	97.17±2.3	97.41±2.1	97.63±1.9	97.85±2.0
5	2.53×10^5	3.98×10^7	12047	97.54±1.7	97.53±1.9	97.64±1.7	97.91±1.6	98.19±1.6

在实验1中，模型参数量和计算量分别降低了26.88%和23.94%，运行总时间并未缩短，表明Transformer较CNN计算效率更低，而准确率和AUC值分别提升了0.92%和0.71%，表明密集连接的交叉堆叠可获得更好的整体性能。如图4第2行所示，该模型可以学习全局特征，关注范围较大。在实验2中，由于方向卷积压缩Transformer输入尺寸，资源占用和消耗均小幅降低，准确率和AUC值分别提升了0.92%和0.71%，模型实际运行性能更稳定，表明方向卷积可捕获不同形状、大小的空间和方向信息。从图4第3行可以看出，该模型关注区域更大，能够更好地建模全局相关性和捕获影像的方向特征，识别错误率明显降低。在实验3中，参数量和计算量分别降低了19.99%和30.02%，而运行时间仅缩短了4.99%，CNN中分组卷积虽然可以轻量化，但会导致性能降低。在实验4中，参数量和计算量分别缩减了73.07%和72.68%，方向卷积的轻量化效果更佳，模型性能更为稳定，准确率和AUC值分别提升了1.81%和1.61%，表明混合器可以自适应增强肺炎特征，而运行时间缩减了17.72%，表明轻量化卷积在实际运行中可以更高效运行。在实验5中，模型参数量和计算量急剧下降至2.53×10^5和3.98×10^7，参数量、计算量和运行时间分

别缩减了 79.91%、91.97% 和 10.62%，表明 1×1 卷积的轻量化有效性更高，且准确率和 AUC 值分别提升了 0.32% 和 0.37%。从图 4 第 4 行看出，模型能更好地聚焦于胸腔区域，较好地忽视肺部边缘区域，使模型能更加精准地定位到肺炎病灶区域。

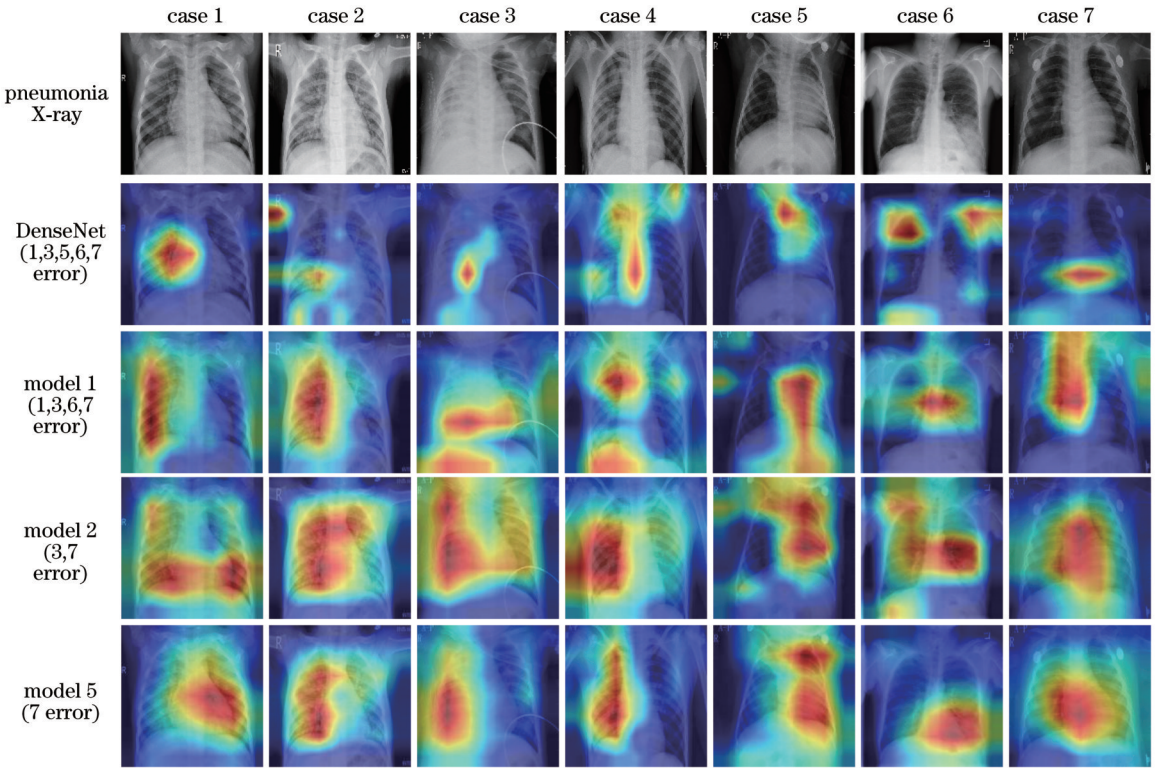

图 4　消融模型在肺炎 X 光片上的热力图

Fig. 4　Thermograms of ablation model on pneumonia X-ray

3.3　损失函数实验与分析

在本文模型上进行了 5 组不同损失函数的实验，具体结果如表 2 所示。从表 2 可以看出：使用合页损失后，模型训练变得更加稳定，表明通过不奖励过拟合样本，也就是降低易分类类别中极易分类样本的权重，可以确保整体分类误差；使用针对标签噪声的对称交叉熵损失，可小幅提升模型性能；使用聚焦损失时，准确率和 AUC 值分别提升了 0.89% 和 0.92%，但模型稳定性明显降低；使用对聚焦损失进行重新加权的类平衡聚焦损失时，模型训练变得更稳定；使用所设计的均衡聚焦损失时，准确率和 AUC 值分别提升了 1.36% 和 1.35%，表明在改变错误分类和小样本的权重时，降低过分类样本的权重，确保整体分类误差，可以明显提高分类预测精度和模型稳定性。

表 2　使用不同损失函数时的评价指标对比

Table 2　Comparison of evaluation indices with different loss functions

Model	*A* /%	AUC /%	*R* /%
Cross entropy loss	97.54±1.7	97.53±1.9	97.64±1.7
Hinge loss[15]	97.88±0.9	97.86±1.1	97.98±1.0
Symmetric cross entropy loss[16]	98.01±1.7	97.99±1.9	98.09±1.8
Focal loss[16]	98.41±1.9	98.43±2.0	98.31±1.9
Class balance focal loss[19]	98.48±1.5	98.51±1.7	98.31±1.8
Balanced focal loss	98.87±0.8	98.85±1.0	98.99±0.9

3.4　对比实验与分析

在 3.1 节所述的肺炎 X 光片数据集上，将本文模型——LDTransformer 与 3 个 CNN 轻量化模型、3 个 Transformer 轻量化模型和 5 个 CNN 结合 Transformer 轻量化模型进行对比，具体分类结果如表 3 所示。相比于其他轻量化模型，本文模型以较低的资源消耗获得了最佳的性能，说明本文模型具有较高的效率和较强的肺炎识别能力。

相比于 3 个 CNN 轻量化模型，本文模型较参数量最小的 EfficientNetb0 总耗时快了 22.25%，并获得明

表 3　肺炎 X 光片数据集中各模型的对比结果

Table 3　Comparison results of each model in pneumonia X-ray dataset

Model	Parameter amount	Calculation amount	Total time /s	A /%	AUC /%	R /%	F_1 /%	P /%
EfficientNetb0[3]	4.48×10^4	1.35×10^7	15155	91.52±2.5	91.68±2.8	90.78±3.1	92.65±2.6	94.61±2.4
MobileNetV3[13]	1.66×10^6	6.24×10^7	14757	92.05±1.8	92.03±2.3	92.13±2.2	93.17±1.9	94.25±1.8
GhostNet[5]	3.90×10^6	1.48×10^8	15398	94.17±2.0	94.05±2.2	94.71±2.4	95.03±2.1	95.36±2.0
DeiT-S[8]	2.19×10^6	4.24×10^9	21986	92.78±3.3	92.63±3.6	93.48±3.1	93.85±3.2	94.22±3.3
T2T-ViT-19[9]	2.14×10^6	4.33×10^9	22179	94.62±2.6	94.37±2.9	95.84±2.4	95.46±2.5	95.09±2.5
XCiT-S24-16T[20]	4.76×10^7	8.95×10^9	25142	96.02±1.9	95.94±2.1	96.40±1.8	96.62±1.8	96.84±1.9
ConvViT[10]	4.22×10^5	9.08×10^7	15817	93.17±3.9	92.99±3.8	94.04±4.1	94.20±3.8	94.36±4.0
Mobile-Former[13]	4.58×10^6	9.26×10^7	16124	93.37±2.8	93.28±2.7	93.81±2.8	94.43±2.7	94.88±2.9
Mobile-ViT[14]	5.63×10^6	1.75×10^9	18978	95.19±2.3	95.19±2.3	96.00±2.4	95.27±2.3	94.55±2.2
EdgeNeXt-S[21]	5.58×10^6	9.57×10^8	16535	96.02±2.5	95.92±2.4	96.51±2.3	96.62±2.4	96.73±2.6
NextVit-S[22]	3.18×10^7	5.79×10^9	20136	96.75±2.1	96.63±2.0	97.30±2.2	97.25±2.1	97.19±2.0
LDTransformer	2.53×10^5	3.98×10^7	11783	98.87±0.8	98.85±1.0	98.99±0.9	99.04±0.9	99.10±0.8

显的性能提升；较 GhostNet 模型的准确率、AUC 值、召回率、F_1 分数和精确率分别提高了 4.75%、5.10%、4.52%、4.22% 和 4.92%。相比于 3 个纯 Transformer 轻量化模型，本文模型的计算效率大幅提高，同时获得了更优的性能：较 DeiT-S 模型的训练时间缩短了 46.41%，5 项指标均提升近 6.56%；较建模局部结构信息与全局相关性的 T2T-ViT-19 模型，5 项指标均提高约 4.49%；较利用交叉协方差矩阵进行键和查询之间交互的 XCiT-S24-16T 模型，5 项指标分别提高了 2.96%、3.03%、2.68%、2.51% 和 2.33%，尽管 XCiT-S24-16T 采用跨通道而不是向量划分方式进行计算，但还是难以忽视缺乏局部特征的影响。

相比于 5 个 CNN 结合 Transformer 轻量化模型，本文模型以较少的参数量、计算量，以及较短的训练时间在肺炎识别方面具有明显优势。相较于卷积归纳偏差自注意力的 ConvViT，本文模型以 25.50% 的训练时间获得近 6.12% 的性能提升；Mobile-Former 需要学习的向量数量很少，但其训练时间并未明显降低；相较于 Transformer 向 CNN 嵌入的 Mobile-ViT，本文模型以更高的效率获得近 3.87% 的精度提升；相较于深度可分离全局注意力模型 EdgeNeXt-S，本文模型以 4.16% 的计算量获得 2.97% 准确率和 3.05% AUC 值的提升；相较于工业部署场景中设计的 NextVit-S，本文模型以较少的资源消耗在 5 项指标上分别提高了 2.19%、2.30%、1.74%、1.84% 和 1.97%。

图 5 所示为 12 种模型的 ROC 曲线和 AUC 值，可以看到，本文模型 LDTransformer 的 ROC 曲线位于左上角，具有明显的优势和较强的鲁棒性，能较好地识别肺炎全局病灶信息与局部病灶信息。

图 6 所示为 12 种模型的 P-R 曲线，以精确率（P）为横轴、召回率（R）为纵轴，PR 表示曲线下方的面积。可以看出，本文模型的曲线下方面积最大、性能最优。

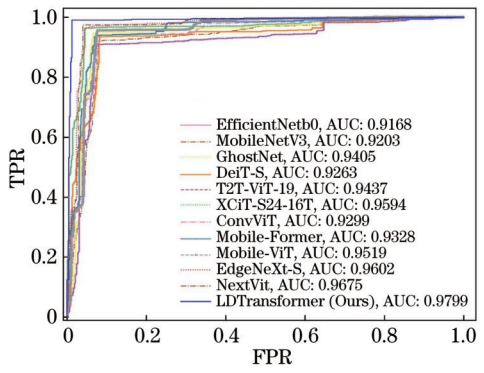

图 5　各模型的 ROC 曲线

Fig. 5　ROC curves of each model

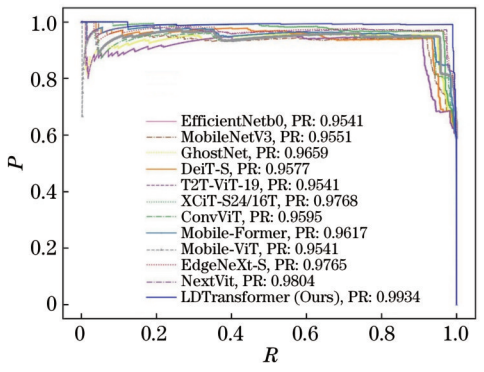

图 6　各模型的 P-R 曲线

Fig. 6　P-R curves of each model

3.5　不同公开数据集实验结果对比

为验证本文模型的泛化能力和鲁棒性，在两个公开的肺炎相关 X 光片数据集上进行对比实验，结果如表 4 和表 5 所示，其中：1 个公开数据集来自文献[23]，

共有 8552 幅正常图像、5674 幅肺炎图像和 7598 幅新冠肺炎图像；第 2 个公开数据集来自文献[24]，共有 510 幅正常图像和 510 幅新冠肺炎图像。可以看到，本文模型 LDTransformer 均取得较好的结果。

表 4　第 1 个公开数据集的结果

Table 4　Results of the first public dataset

Model	TPR /%	TNR /%	*A* /%	AUC /%
DenseNet121[12]	91.00	87.00	88.13	90.00
EfficientNetb0[3]	83.00	92.00	94.64	95.00
Covid-caps[23]	90.00	95.00	95.00	97.00
ViT-B32[23]	96.00	96.00	96.00	99.10
LDTransformer	98.55	98.97	98.71	99.53

表 5　第 2 个公开数据集的结果

Table 5　Results of the second public dataset

Model	*A* /%	AUC /%
CVDNet[24]	97.20	—
AF-CAP[24]	99.16	98.80
LDTransformer	99.63	99.17

4　结　　论

针对特征提取不充分和模型轻量化程度不足的问题，提出一种用于 X 光片辅助诊断肺炎的 LDTransformer 模型，构造 CNN 结合 Transformer 的密集连接架构，在深浅层中对局部和全局信息进行结合。设计横向、纵向和扩张卷积并行的方向卷积，学习不同形状、大小的空间和方向信息，为每个样本特征采用专门卷积核的轻量化，实现了较少资源消耗，设计均衡聚焦损失函数优化训练。所提 LDTransformer 在肺炎 X 光片数据集中以 2.53×10^5 的较低模型参数量、3.98×10^7 的最低模型计算量和 12647 s 的最快总速度获得 98.87% 准确率和 98.85% AUC 值的较高性能，并在两个公共肺炎 X 光片数据集中均获得最优精度，具有较好的泛化能力和较强的鲁棒性，热力图可视化对比结果进一步说明，本文模型以较高的轻量化程度，有效提取了肺炎 X 光片影像方向特征和全局特征。

参　考　文　献

[1] Zhou T, Ye X Y, Lu H L, et al. Dense convolutional network and its application in medical image analysis[EB/OL]. (2022-04-25)[2023-02-01]. https://doi.org/10.1155/2022/2384830.

[2] Naralasetti V, Shaik R K, Katepalli G, et al. Deep learning models for pneumonia identification and classification based on X-ray images[J]. Traitement Du Signal, 2021, 38(3): 903-909.

[3] 龚希, 陈占龙, 吴亮, 等. 用于高分辨遥感影像场景分类的迁移学习混合专家分类模型[J]. 光学学报, 2021, 41(23): 2301003.
Gong X, Chen Z L, Wu L, et al. Transfer learning based mixture of experts classification model for high-resolution remote sensing scene classification[J]. Acta Optica Sinica, 2021, 41(23): 2301003.

[4] Jain R, Nagrath P, Kataria G, et al. Pneumonia detection in chest X-ray images using convolutional neural networks and transfer learning[J]. Measurement, 2020, 165: 108046.

[5] Li H, Zeng N Y, Wu P S, et al. Cov-Net: a computer-aided diagnosis method for recognizing COVID-19 from chest X-ray images via machine vision[J]. Expert Systems With Applications, 2022, 207: 118029.

[6] 李翔, 何淼, 罗海波. 一种面向遮挡行人检测的改进 YOLOv3 算法[J]. 光学学报, 2022, 42(14): 1415003.
Li X, He M, Luo H B. Occluded pedestrian detection algorithm based on improved YOLOv3[J]. Acta Optica Sinica, 2022, 42(14): 1415003.

[7] Naseer M, Ranasinghe K, Khan S, et al. Intriguing properties of vision transformers[EB/OL]. (2021-05-21) [2022-10-09]. https://arxiv.org/abs/2105.10497.

[8] Okolo G I, Katsigiannis S, Ramzan N. IEViT: an enhanced vision transformer architecture for chest X-ray image classification[J]. Computer Methods and Programs in Biomedicine, 2022, 226: 107141.

[9] Park S, Kim G, Oh Y, et al. Multi-task vision transformer using low-level chest X-ray feature corpus for COVID-19 diagnosis and severity quantification[J]. Medical Image Analysis, 2022, 75: 102299.

[10] Yuan L, Chen Y P, Wang T, et al. Tokens-to-token ViT: training vision transformers from scratch on ImageNet[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2022: 538-547.

[11] Liang S, Nie R C, Cao J D, et al. FCF: feature complement fusion network for detecting COVID-19 through CT scan images [J]. Applied Soft Computing, 2022, 125: 109111.

[12] 林昭苏, 王杨云逗, 王昊, 等. 基于 DenseNet 的散射成像景深拓展研究[J]. 光学学报, 2022, 42(4): 0436001.
Lin Z S, Wang Y Y D, Wang H, et al. Expansion of depth-of-field of scattering imaging based on DenseNet[J]. Acta Optica Sinica, 2022, 42(4): 0436001.

[13] Chen Y P, Dai X Y, Chen D D, et al. Mobile-former: bridging MobileNet and transformer[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 5260-5269.

[14] Mehta S, Rastegari M. MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer[EB/OL]. (2021-10-05)[2022-10-08]. https://arxiv.org/abs/2110.02178.

[15] Kim Y, Ohn I, Kim D. Fast convergence rates of deep neural networks for classification[J]. Neural Networks, 2021, 138: 179-197.

[16] Wang Y S, Ma X J, Chen Z Y, et al. Symmetric cross entropy for robust learning with noisy labels[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27 - November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2020: 322-330.

[17] Kermany D S, Zhang K, Goldbaum M. Labeled optical coherence tomography (OCT) and chest X-ray images for

classification[J]. Mendeley data, 2018, 2(2): 17632.
[18] Giełczyk A, Marciniak A, Tarczewska M, et al. Pre-processing methods in chest X-ray image classification[J]. PLoS One, 2022, 17(4): e0265949.
[19] Cui Y, Jia M L, Lin T Y, et al. Class-balanced loss based on effective number of samples[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2020: 9260-9269.
[20] El-Nouby A, Touvron H, Caron M, et al. XCiT: cross-covariance image transformers[EB/OL]. (2021-06-17)[2022-10-09]. https://arxiv.org/abs/2106.09681.
[21] Maaz M, Shaker A, Cholakkal H, et al. EdgeNeXt: efficiently amalgamated CNN-transformer architecture for mobile vision applications[M]//Karlinsky L, Michaeli T, Nishino K. Computer vision-ECCV 2022 workshops. Lecture notes in computer science. Cham: Springer, 2023, 13807: 3-20.
[22] Li J S, Xia X, Li W, et al. Next-ViT: next generation vision transformer for efficient deployment in realistic industrial scenarios[EB/OL]. (2022-07-12) [2022-10-08]. https://arxiv.org/abs/2207.05501.
[23] Chetoui M, Akhloufi M A. Explainable vision transformers and radiomics for COVID-19 detection in chest X-rays[J]. Journal of Clinical Medicine, 2022, 11(11): 3013.
[24] Balasubramanian K, Ananthamoorthy N P, Ramya K. An end-end deep learning framework for lung infection recognition using attention-based features and cross average pooling[J]. International Journal for Multiscale Computational Engineering, 2022, 20(2): 67-82.

Lightweight Directional Transformer for X-Ray-Aided Pneumonia Diagnosis

Zhou Tao[1,2], Ye Xinyu[1,2*], Liu Fengzhen[1,2], Lu Huiling[3]

[1]*School of Computer Science and Engineering, North Minzu University, Yinchuan* 750021, *Ningxia, China*;
[2]*The Key Laboratory of Images and Graphics Intelligent Processing of State Ethnic Affairs Commission, North Minzu University, Yinchuan* 750021, *Ningxia, China*;
[3]*School of Medical information and Engineering, Ningxia Medical University, Yinchuan* 750004, *Ningxia, China*

Abstract

Objective Computer-aided pneumonia diagnosis with chest X-rays based on convolutional neural networks (CNNs) is an important research direction. The presence of factors such as patient positions and inspiratory depth in chest X-rays images can lead to confusion with other diseases, and existing methods ignore the directional and spatial features of images in chest X-rays, such as the common onset of pneumonia in the middle and lower lobes of the lung. However, it is difficult to extract the directional information and global semantic information of pneumonia X-rays by a CNN. Additionally, the model is not sufficiently lightweight, and the time and space complexity is high. Hence, this paper proposes a lightweight directional Transformer (LDTransformer) model for pneumonia X-rays to assist in diagnosis.

Methods The densely connected architecture of CNN combined with the Transformer is constructed. It is composed of cross-stacking local feature extraction and global feature extraction, and its dense connections are used to achieve the combination of local and global information in deep and shallow layers. Next, lateral, vertical, and dilated convolutions in parallel with the directional convolution are designed to capture spatial and directional information of different shape sizes. The directional convolution is used to compress feature scales in the Transformer and learn global features and directional features of images with low computational complexity. After that, the lightweight convolution in CNN is designed. It employs a dedicated convolution kernel for each sample feature, learns features in chunks, and fuses them by a channel-noted blender to reduce the number of model parameters and maintain efficient computation while effectively increasing the feature extraction capability of the network. Finally, a balanced focal loss function is constructed to increase the weight of small and misclassified samples and decrease the weight of overclassified samples.

Results and Discussions The LDTransformer model achieves high recognition accuracy with good robustness and generalization in all three X-ray datasets of number, category, and difficulty. Smaller datasets make it difficult for the high-performance CNN and Transformer models to learn sufficiently, while the lightweight model using a combination of both can obtain high recognition accuracy (Table 6). Compared with various lightweight models of CNN and Transformer (Table 4), the model in this paper has advantages in terms of the number of parameters, computation, and training time. In particular, its lightweight design with a dedicated convolution kernel for each sample feature makes the operation efficiency significantly better than that of existing models. Finally, the performance of each component of the model in this paper is tested separately by ablation experiments and loss function comparison experiments, and the region of interest and

accuracy of the model are visualized by the heat map visualization in the ablation experiments (Fig. 4).

Conclusions Considering the inadequate feature extraction and insufficient model lightweight, this paper proposes a model for X-ray-aided pneumonia diagnosis to combine local and global information in deep and shallow layers. The directional convolution learns spatial and directional information of different shape sizes. The lightweight convolution with a dedicated convolution kernel for each sample feature is designed to reduce resource consumption, and a balanced focal loss function is constructed to optimize training. The proposed model achieves the accuracy of 98.87% and an AUC value of 98.85% under a small number of model parameters (2.53×10^5), the lowest model computation (3.98×10^7), and the fastest total speed (12647 s) in the pneumonia X-ray dataset. It effectively extracts the directional features and global features of pneumonia X-ray images with a high degree of lightweight.

Key words image processing; densely local and global features; directional Transformer; lightweight convolution; pneumonia X-ray