激光与光电子学进展 先进成像

基于 Vision Transformer的小儿肺炎辅助诊断

赵爽[1]，魏国辉[2]，赵文华[2*]，马志庆[2]

[1]山东中医药大学实验室管理处，山东 济南 250355；

[2]山东中医药大学智能与信息工程学院，山东 济南 250355

摘要 为改善基层医疗机构儿童肺炎诊疗水平，提高基层医生分析临床医学影像的效率和质量，提出了一种基于Vision Transformer(ViT)的小儿肺炎辅助诊断模型。首先利用ResUNet对儿童胸片进行肺区域分割，将左右肺区域从胸片中分割出来以降低其他组织对肺炎诊断的干扰。然后，将分割后的图像输入改进的混合ViT模型进行诊断，该模型使用传统卷积神经网络的特征映射作为Transformer的输入，并在卷积神经网络中引入自注意力机制，增强卷积以加强其获取全局相关性的能力。最后，对卷积神经网络的骨干网络和Transformer模型进行端到端的训练，使模型能够达到良好的图像分类结果。在Chest X-Ray Images肺炎标准数据集上进行了实验，实验结果表明，所提模型的肺炎识别准确率、精确率和召回率分别达到97.27%、97.69%和98.60%。即该模型具有较好的可行性，可使基层儿童肺炎的临床诊断准确率得到很大提升。

关键词 图像处理；图像分类；儿科肺炎；残差网络；自注意力机制；Transformer

中图分类号 R445 **文献标志码** A **DOI**：10.3788/LOP213019

Assistant Diagnosis of Pediatric Pneumonia Based on Vision Transformer

Zhao Shuang[1], Wei Guohui[2], Zhao Wenhua[2*], Ma Zhiqing[2]

[1]*Laboratory Management Office, Shandong University of Traditional Chinese Medicine, Jinan* 250355, *Shandong, China*;

[2]*College of Intelligence and Information Engineering, Shandong University of Traditional Chinese Medicine, Jinan* 250355, *Shandong, China*

Abstract To improve the diagnosis and treatment level of pneumonia in children in primary medical institutions and doctors' efficiency and quality in analyzing clinical medical images, an auxiliary diagnosis model of pneumonia in children, based on the Vision Transformer (ViT), is proposed. First, ResUNet is used to segment the lung region in the chest film of children, and the left and right lung regions are separated from the chest film to mitigate the interference of other tissues during pneumonia diagnosis. Further, the segmented image is input into the improved hybrid ViT model for diagnosis. This model uses the feature map of the traditional convolutional neural network (CNN) as the input of the Transformer and introduces the self-attention mechanism into the CNN to improve convolution to enhance its ability to obtain global correlation. Finally, the backbone network of the CNN and Transformer model are trained end-to-end so that the proposed model can achieve good image classification results. Experiments were conducted on the Chest X-Ray Images pneumonia standard dataset. The experimental results show that the accuracy, precision, and recall of the proposed model for pneumonia recognition reach 97.27%, 97.69%, and 98.60% respectively. In other words, the model has good feasibility and can significantly improve the clinical diagnosis accuracy of pneumonia in children at the grass-root level.

Key words image processing; image classification; pediatric pneumonia; residual network; self-attention mechanism; Transformer

1 引 言

根据世界卫生组织的报告，全球每年患肺炎死亡的儿童人数超过死于疟疾、痢疾和麻疹的儿童人数总和[1]。患者如果尽早接受治疗并且免疫系统没有受到破坏可以痊愈，因此肺炎的早期诊断极为重要。当前，

收稿日期：2021-11-22；**修回日期**：2021-12-29；**录用日期**：2022-01-05；**网络首发日期**：2022-01-16

基金项目：山东省研究生教育质量提升计划课题(SDYJG1943)

通信作者：*zhaowh0621@163.com

肺部X光检查是肺炎诊断的主要途径，这种方法在临床用药及流行病学研究中发挥着重要作用[2]。然而与成人不同，儿童在不同的年龄段，其肺炎症状在胸片上的表现呈现多样化情况。利用胸片进行肺炎诊断是一项具有挑战性的工作，需要医生具备较高的医疗影像判断能力。近年来，计算机辅助诊断技术作为一种辅助医疗工具，被广泛应用于医疗影像学的领域[3]。计算机辅助诊断技术在减少肺炎的漏诊误诊、提高诊断准确率等方面发挥着积极的作用。

随着深度学习技术的不断进步及数据处理能力的不断提升，国内外学者纷纷在肺炎图像领域展开了关于深度学习的研究。Rajpurkar等[4]设计了一个121层卷积神经网络CheXNet，并在目前最大的X线胸片数据集Chest-X-Ray14[5]上进行了训练、测试，所得结果与4位从事学术研究的放射学家的标注结果的对比表明，由CheXNet得到的F1 score指标超过了放射科医生的平均水平。Varshni等[6]进一步使用在ImageNet上预训练好的Xception[7]、RestNet-50[8]、DensNet-121[9]等网络对Chest-X-Ray14数据集进行测试，并利用不同分类器对其性能进行评估，area under curve（AUC）最高达到了0.80。Kermany等[10]开发了Chest X-Ray数据集，利用预训练的AlexNet对肺炎图像进行迁移学习，识别准确率达到92.80%。梁高博[11]提出引入扩张卷积以代替部分普通卷积层的扩张卷积神经网络模型DCNET、结合注意力机制以提升网络判别能力的网络RES_SE_DCNET，并采用RES_SE_DCNET在大规模数据集Chest-X-Ray14上进行预训练，然后进行迁移学习的分类方法Finetuning_RES_SE_DCNET的测试，准确率分别达到89.10%、89.74%和90.71%。这种计算机辅助肺炎诊断方案可改善基层医疗机构诊疗水平，然而受到儿童肺炎影像数据较为稀缺的限制，目前国内对儿童肺炎辅助诊断算法的系统性研究相对较少，仍然需要开展进一步的研究以满足临床诊断要求。

目前在代表分类领域最高权威的ImageNet数据集[12]图像分类竞赛中，Vision Transformer（ViT）模型[13]以88.55%的准确率成功登顶第一的宝座，标志着Transformer[14]类的网络结构也可以很好地完成由卷积神经网络主导的分类任务。本文利用改进的混合ViT模型对分割后的儿童胸片进行诊断。该模型以卷积神经网络和Transformer相结合的方式，首先采用改进的卷积神经网络对儿童胸片进行特征提取，在卷积神经网络中引入自注意力机制（self-attention）以增强卷积，对比直接输入投影的图像块可获得更详细的图像特征，然后输入Transformer模型进行分类。结合卷积神经网络和Transformer模型的优势对Transformer模型和卷积神经网络的骨干网络进行端到端的训练，达到良好的图像分类效果，为基层医疗机构肺炎的早期筛查提供较为准确的辅助诊断意见。

2 网络结构设计

首先利用ResUNet对儿童胸片进行分割，然后将分割后的图像使用R-MSA32进行特征提取，最后输入ViT模型进行诊断。对卷积神经网络的骨干网络和Transformer模型进行端到端的训练，使模型达到良好的图像分类结果。整体网络结构如图1所示。

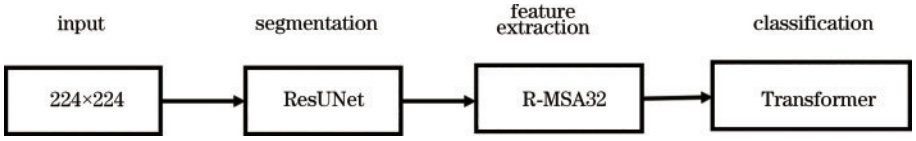

图1 整体网络结构

Fig. 1 Overall network architecture

2.1 分割

首先基于ResUNet分割出胸片中的左右肺区域，降低胸片中背景噪声对肺炎诊断的干扰。ResUNet以U-Net[15]结构为基础，U-Net结构如图2所示，其左侧可视为一个编码器，右侧可视为一个解码器。编码器和编码器各有4个子模块，每个子模块包含2个卷积层。

ResUNet则将U-Net的各个子模块替换成ResNet中的残差结构，如图3所示，在加深网络结构的同时避免过拟合，增加网络的非线性结构，从而获得更高的分割精度。

将ResUNet模型在公开数据集结核病标准数字图像数据库[16]进行预训练，该数据集来自美国国立医学图书馆、美国国立卫生研究院、贝塞斯达、马里兰州和中国深圳广东医学院深圳第三人民医院。将数据集中704张（附有掩膜[17]）图片随机分成566张图片组成的训练集、138张图片组成的测试集。模型训练完成后，迁移学习到5856张儿童胸片上，分割出两肺区域，得到只包含左右肺区域和背景的黑白掩膜，再结合掩膜与原图像，完成肺炎胸片上感兴趣区域的分割。

2.2 ViT模型

ViT模型[13]结构如图4所示。首先将图像分割成一个个图像块，然后将每个图像块组成一个向量，如图4虚线内所示。具体地，记输入图像为$\boldsymbol{X}$，$\boldsymbol{X}\in \mathrm{R}^{H\times W\times C}$，其中$H$和$W$分别是图像的高和宽，$C$为通道数，对于RGB图像通道数就是3。用$P\times P$大小的图像块去分割整个图像可以得到$N$个图像块，其中

$$N=\frac{HW}{P^2}。\tag{1}$$

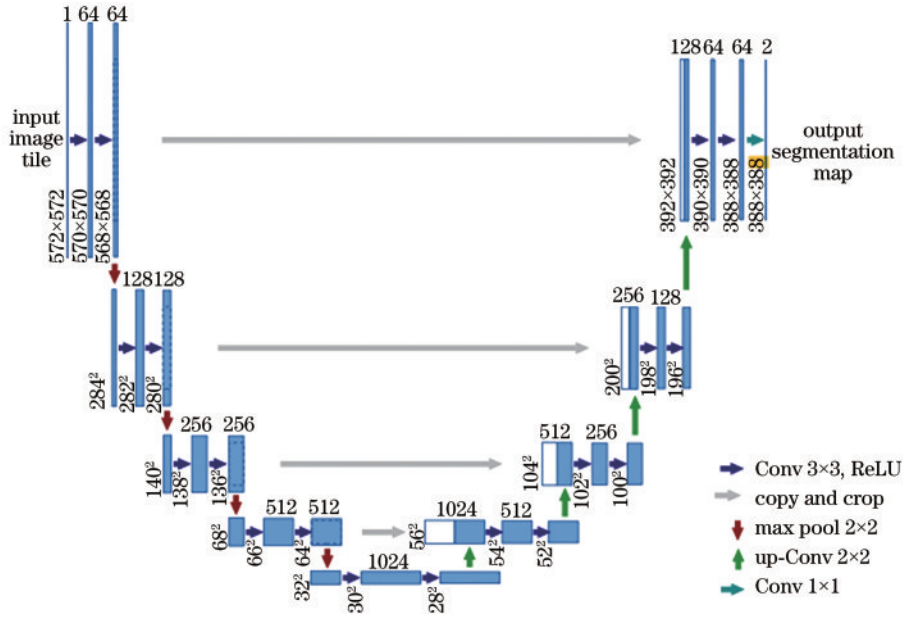

图 2　U-Net 结构图

Fig. 2　U-Net architecture

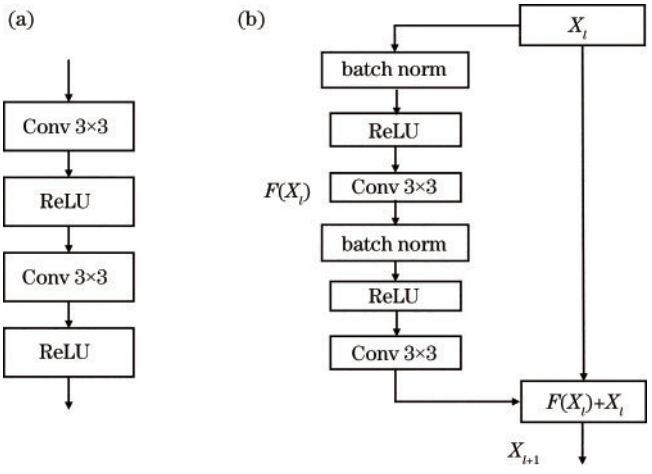

图 3　示例图。(a) U-Net 子模块；(b)残差模块

Fig. 3　Sample graph. (a) U-Net sub-module; (b) residual module

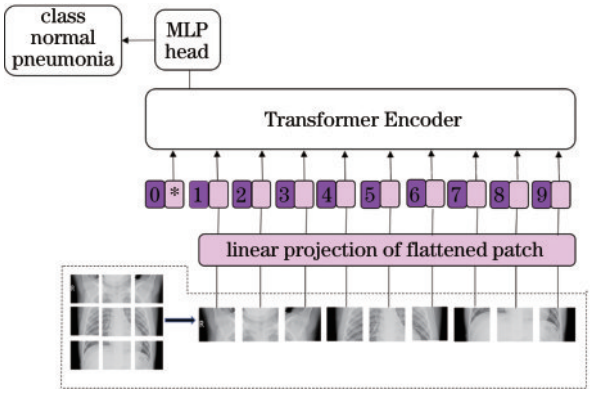

图 4　ViT 模型结构

Fig. 4　ViT model architecture

每个小图像块的像素大小是 $P\times P\times C$，转化为向量后就是 P^2C 维的向量，将 N 个图像块的向量连接在一起就得到了一个 $N\times P^2C$ 的二维矩阵。然后将上述过程得到的 N 个 P^2C 维的向量进行线性变换，将向量维度降为 D。综上所述，原本 $H\times W\times C$ 维的图像被转化成 N 个 D 维的向量。

由于 Transformer 模型本身是没有位置信息的，需要将位置信息嵌入模型中去。图 4 中，编号为 0～9 的框表示各个图像块的位置信息 $\boldsymbol{E}_{\text{pos}}$，编号为 0～9 右侧的框则是图像块向量经过线性变换之后的 D 维向量。将位置信息 $\boldsymbol{E}_{\text{pos}}$ 和 D 维向量结合输入 Transformer 模型。而带*的框不是通过某个图像块产生的，其对应的是整个图像的表示，记作 $\boldsymbol{X}_{\text{class}}$。

上述这两个过程可描述为

$$\boldsymbol{z}_0=\left[\boldsymbol{X}_{\text{class}};\ \boldsymbol{X}_p^1\boldsymbol{E};\ \boldsymbol{X}_p^2\boldsymbol{E};\ \cdots;\ \boldsymbol{X}_p^n\boldsymbol{E}\right]+\boldsymbol{E}_{\text{pos}},\ \boldsymbol{E}\in\mathrm{R}^{(P^2C)\times D},\ \boldsymbol{E}_{\text{pos}}\in\mathrm{R}^{(N+1)\times D},\tag{2}$$

式中：$\boldsymbol{X}_p^1\boldsymbol{E};\ \boldsymbol{X}_p^2\boldsymbol{E};\ \cdots;\ \boldsymbol{X}_p^n\boldsymbol{E}$ 是 P^2C 维的图像块向量右乘矩阵 $\boldsymbol{E}$ 得到的 D 维向量。这 N 个 D 维向量和同样是 D 维向量的 $\boldsymbol{X}_{\text{class}}$ 结合在一起就得到了一个 $(N+1)\times D$ 维矩阵，再加上维位置信息 $\boldsymbol{E}_{\text{pos}}$ 就是 Transformer encoder 的输入 $\boldsymbol{z}_0$。

Transformer Encoder[14] 是由 Encoder 模块重复堆叠 L 次组成的，Encoder 模块的结构如图 5 所示。

对于 Encoder 模块的第 l 层，记其输入为 $\boldsymbol{z}_{l-1}$，输出为 $\boldsymbol{z}_l$，则整个计算过程为

$$\boldsymbol{z}_l=\mathrm{MLP}\left[\mathrm{LN}\left(\boldsymbol{z}_l'\right)\right]+\boldsymbol{z}_l',\ l=1,\cdots,N,\tag{3}$$

$$\boldsymbol{z}_l'=\mathrm{MSA}\left[\mathrm{LN}\left(\boldsymbol{z}_{l-1}\right)\right]+\boldsymbol{z}_{l-1},\ l=1,\cdots,N,\tag{4}$$

式中：MSA[18] 为多头自注意力，是图 5 中的 multi-head self-attention；MLP[19] 为多层感知机，是图 5 中的 multi-layer perceptron；LN[20] 为层归一化，是图 5 中的 layer norm。

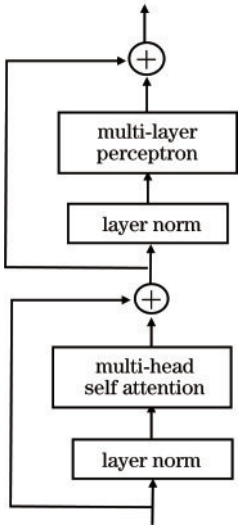

图 5　Encoder模块结构图

Fig. 5　Encoder module architecture

MLP Head 是最终用于分类的层结构，由全连接层和激活函数组成。

2.3　混合 ViT 网络模型

混合 ViT 网络模型首先用传统的卷积神经网络来提取特征，然后通过 ViT 模型得到最终的结果。以 R50＋ViT[13]网络为例，这里的特征提取部分 R50 采用的是改进的 ResNet50，表 1 为 R50 特征提取阶段网络结构，网络主要在以下 3 个部分进行了更改：

1）通过标准化卷积层中的权重来加速深度网络训练，即表 1 中的 stdConv[21]；

2）使用组归一化（GN）[22]替换批归一化（BN）[23]；

3）将 stage 4 中的 3 个模块移到 stage 3 中。

输入图像尺寸为 224 × 224 × 3，最后输出一个 14 × 14 × 768 维向量，与 ViT 模型中 Transformer Encoder的输入相对应。

表 1　R50特征提取阶段网络结构

Table 1　Network structure of R50 feature extraction stage

Output size	R50
112 × 112	stdConv，7 × 7，64，stride 2
	max pool，3 × 3，stride 2
56 × 56	$\begin{bmatrix}\text{stdConv},1\times1,64\\ \text{stdConv},3\times3,64\\ \text{stdConv},1\times1,256\end{bmatrix}\times3$
28 × 28	$\begin{bmatrix}\text{stdConv},1\times1,128\\ \text{stdConv},3\times3,128\\ \text{stdConv},1\times1,512\end{bmatrix}\times4$
14 × 14	$\begin{bmatrix}\text{stdConv},1\times1,256\\ \text{stdConv},3\times3,256\\ \text{stdConv},1\times1,1024\end{bmatrix}\times9$
14 × 14	Conv，1 × 1，768，stride 1

2.4　改进的混合 ViT 网络模型

将 R50 网络 stage 3 中 3×3 卷积更改为 MSA[18]模块，同时将 9 个模块缩减为 3 个，形成一个新的网络 R-MSA32，其中 self-attention 用于计算特征中不同位置之间的权重，从而达到更新特征的效果。首先将输入特征通过随机初始化映射矩阵生成 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 等 3 个特征，然后将 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 进行点乘得到 attention map，再将 attention map 与 $\boldsymbol{V}$ 点乘得到加权后的特征，如图 6 所示。不同的随机初始化映射矩阵可以将输入向量映射到不同的子空间，这可以让模型从不同角度理解输入的特征，因此同时几个 self-attentionn 的组合效果可能会优于单个 self-attention，这种同时计算多个 self-attention 的方法被称为 multi-head self-attention。每个 head 都会产生一个输出特征，再把多个合并的多维自注意力特征进行降维，最后得到一个新的特征。表 2 为 R-MSA32 特征提取阶段网络结构。

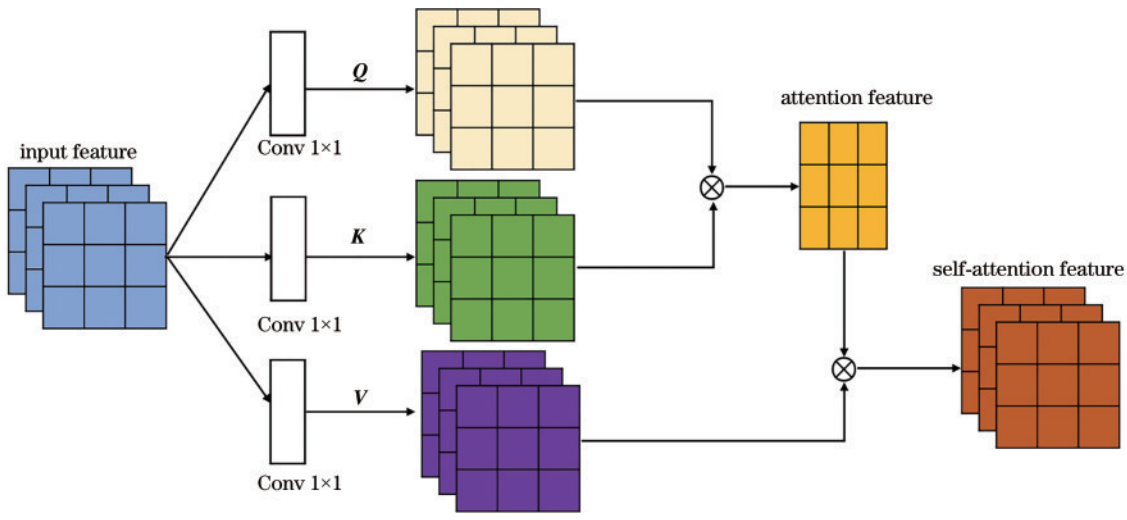

图 6　自注意力机制

Fig. 6　Self-attention mechanism

表 2　R-MSA32特征提取阶段网络结构

Table 2　Network structure of R-MSA32 feature extraction stage

Output size	R-MSA 32
112 × 112	stdConv, 7 × 7, 64, stride 2
	max pool, 3 × 3, stride 2
56 × 56	$\begin{bmatrix}\text{stdConv},1\times1,64\\ \text{stdConv},3\times3,64\\ \text{stdConv},1\times1,256\end{bmatrix}\times3$
28 × 28	$\begin{bmatrix}\text{stdConv},1\times1,128\\ \text{stdConv},3\times3,128\\ \text{stdConv},1\times1,512\end{bmatrix}\times4$
14 × 14	$\begin{bmatrix}\text{stdConv},1\times1,256\\ \text{stdConv, MSA},256\\ \text{stdConv},1\times1,1024\end{bmatrix}\times3$
14 × 14	Conv, 1 × 1, 768, stride 1

使用在公共 ImageNet-21k 数据集上预训练的权值初始化改进后混合 ViT 网络模型的部分权重，在儿童胸片数据集上进行迁移学习。

3　实　　验

3.1　数据集

用于训练和评估的数据集是 Chest X-Ray[10]，该公共数据集是基于广州妇女儿童医学中心 1～5 岁儿科患者的 X 射线扫描数据库制作的。Chest X-Ray 数据集包含从儿童医院收集和标记的总共 5856 张胸部 X 射线图像，即胸片。图像格式为 JPEG。为保证实验过程中训练集与测试集图像的独立性，提高模型的泛化能力，按 8∶1∶1 的比例将数据集随机分为训练集、验证集和测试集。具体的 X 射线扫描图像如图 7 所示。

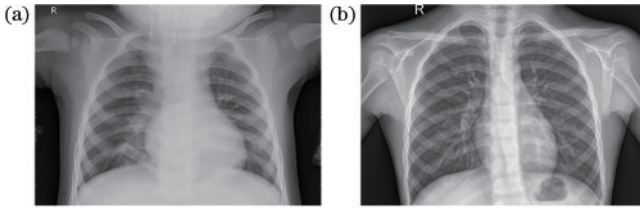

图 7　Chest X-Ray 数据集示例图。(a)肺炎影像；(b)正常影像

Fig. 7　Sample of Chest X-Ray dataset. (a) Pneumonia image; (b) normal image

3.2　数据预处理

为适应卷积神经网络的输入，将图像尺寸调整为 224 × 224，在训练集采用图像翻转和随机角度的旋转变换等仿射变换方法来进行数据增强，以增加卷积神经网络对肺炎影像特征尺度和方向上的鲁棒性。

3.3　实验环境

实验使用 Python 语言基于 PyTorch 框架进行编程，实验平台为基于 GeForce GTX1080Ti 的 Ubuntu 系统。在 CPU 环境下进行数据预处理，在 GPU 上训练模型以加快数据的计算，提高实验效率。

3.4　网络配置

网络配置方面，损失函数使用双稳态逻辑损失函数[24]，有效减小因人为对图像的标记错误造成的不良影响。优化方法采用自适应矩估计(Adam)算法[25]，经过多次实验，将初始学习率设置为 0.0001，并采用小批量数据(batch_size 为 16)的方式训练模型，epoch 设置为 50。

3.5　评价标准

选择准确率(accuracy)、召回率(recall)和精确率(precision)这 3 个指标评价模型性能。准确率即预测患者是否患有肺炎与实际结果之比，体现模型的预测能力；召回率为实际肺炎患者预测为患有肺炎的概率，召回率越高则漏诊的概率越低；精确率是指正确预测为肺炎的患者占全部预测为患有肺炎的比例。准确率、召回率与精确率的表达式分别为

$$R_{\text{accuracy}}=\frac{N_{\text{TP}}+N_{\text{TN}}}{N_{\text{TP}}+N_{\text{TN}}+N_{\text{FN}}+N_{\text{TN}}},\tag{5}$$

$$R=\frac{N_{\text{TP}}}{N_{\text{TP}}+N_{\text{FN}}},\tag{6}$$

$$P=\frac{N_{\text{TP}}}{N_{\text{TP}}+N_{\text{FP}}},\tag{7}$$

式中：N_{TP}、N_{TN}、N_{FP}、N_{FN}分别代表正阳性、正阴性、假阳性与假阴性的数量。

3.6　实验结果与分析

图 8 为 ResUNet 对结核病标准数字图像数据库进行分割的结果。对比掩膜和 ResUNet 预测图像可知，ResUNet 在捕捉形状信息方面效果良好。

图 9 展示了对胸片数据集使用训练好的分割模型的分割效果图，从图中可以看出，分割区域囊括了两肺的大部分区域，为后续肺炎诊断去除了背景噪声。

图 10 是测试集的混淆矩阵，其中竖轴的标签表示真实属性，而横轴的标签表示分类的预测结果。图 10 第 1 行第 1 列表示正常胸片被成功分类成正常胸片的图像数目，为 148 张；第 1 行第 2 列表示正常胸片被分类成肺炎患者的胸片的样本数目，为 6 张；第 2 行第 1 列表示肺炎患者的胸片被分类成正常胸片的样本数目，为 10 张；第 2 行第 2 列表示肺炎患者的胸片被分类成肺炎患者的胸片的样本数目，为 422 张。矩阵中的对角线上的数值代表被正确预测的样本数目，共 570 张。

表 3 是 ResNet50 网络模型、ViT 网络模型、混合 ViT 网络模型 R50＋ViT、ResUNet＋ R50＋ViT 与所提模型的分类准确率、精确率和召回率的对比情况。由于 ResNet50 网络模型、ViT 网络模型、混合 ViT 网络模型 R50＋ViT、ResUNet＋R50＋ViT 与所提模型的模型参数是逐步增加的，分类训练时长和测试时长也是逐步增加的，模型一个 epoch 的训练时长/测试时长分别为 130 s/42 s、156 s/46 s、185 s/49 s、185 s/49 s 及 188 s/50 s。

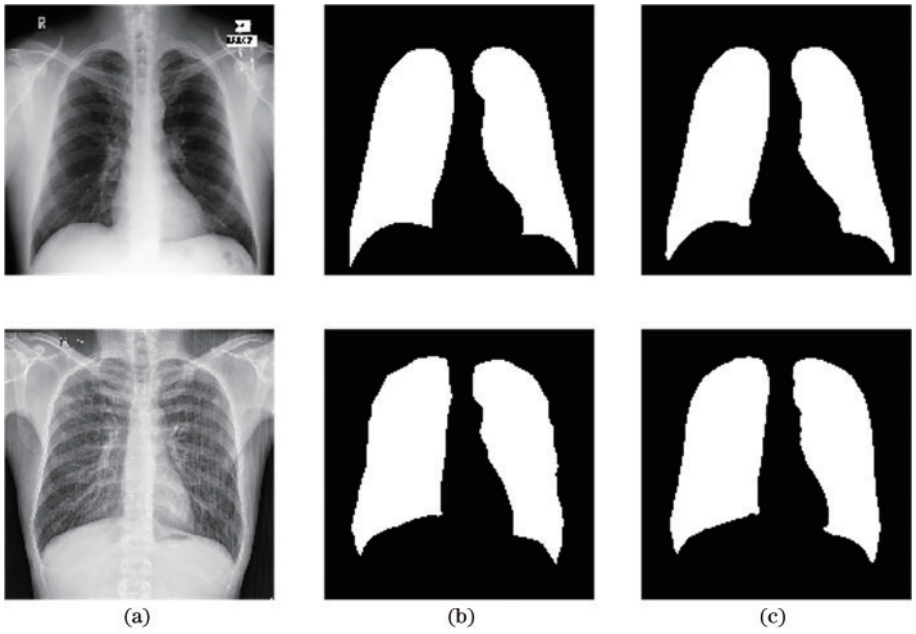

图 8　分割结果。(a)原图像;(b)掩膜;(c)预测图像

Fig. 8　Segmentation results. (a) Original images; (b) masks; (c) prediction images

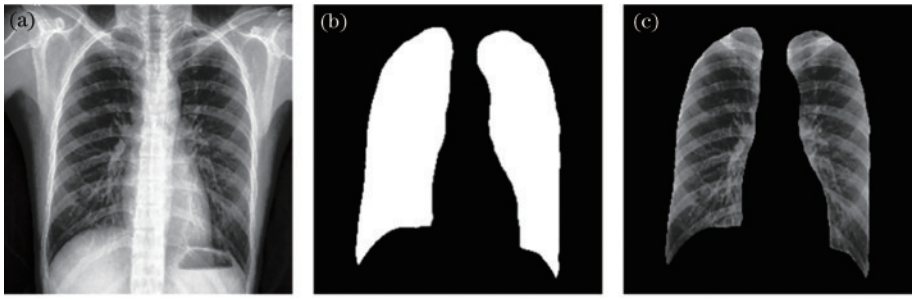

图 9　分割结果。(a)原图像;(b)掩膜;(c)预测图像

Fig. 9　Segmentation results. (a) Original image; (b) mask; (c) prediction image

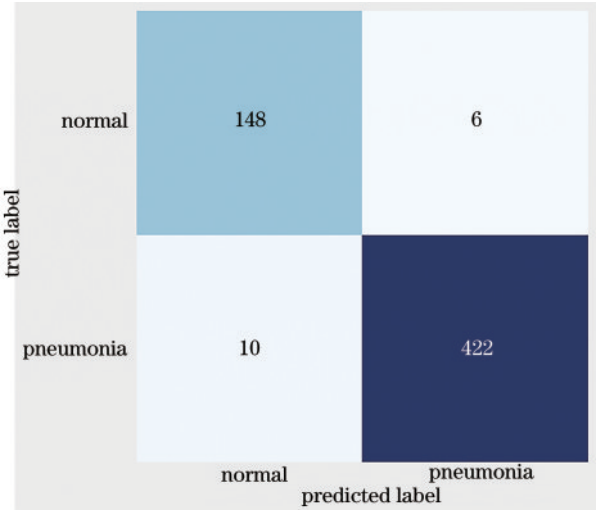

图 10　混淆矩阵

Fig. 10　Confusion matrix

表 3　ViT 网络模型与混合 ViT 网络模型对比

Table 3　Comparison between ViT network model and hybrid ViT network model　　unit: %

Model	Accuracy	Precision	Recall
ResNet50	94.70	94.32	96.49
ViT	95.21	96.49	96.94
R50+ViT	96.08	97.42	97.20
ResUNet+R50+ViT	96.58	98.13	97.22
Proposed model	97.26	98.60	97.69

表 3 数据表明,所提模型的准确率、精确率和召回率都高于其他模型,分别达到 97.26%、98.60% 和 97.69%,说明所提模型具有一定的可行性。与 ResNet50 网络模型相比,ViT 网络模型的性能更优。同时 ViT 网络模型经过足够的预训练并迁移到小规模数据的特定任务时,可以获得出色的结果。对比 ViT 网络模型和其他 3 个混合 ViT 网络模型发现,无论是识别准确率、精确率还是召回率,混合 ViT 网络模型都高于 ViT 网络模型,说明混合 ViT 网络模型可以更好地拟合肺炎图像。对比 R50+ViT 模型与 ResUNet+R50+ViT 模型发现,ResUNet+R50+ViT 模型的分类精度较高,分割后的胸片图像降低了背景图像的干扰,提高了分类精度。对比 ResUNet+R50+ViT 模

型，所提模型的分类效果较好，说明引入自注意力机制可在不增加计算成本的情况下增加感受野，提升分类效果。

为验证所提模型的有效性，在 Chest X-Ray 数据集上与现有的研究成果进行对比，选择 GIV3[26]、DenseNet′[27]及 AlexNet_S[28]等 3 种模型与所提模型进行对比。其中 GIV3 是一种改进的深度卷积神经网络模型，在 GoogLeNet 和 Inception V3 网络的基础上通过构建特征融合层，同时使用 random forest（RF）作为模型的分类器实现肺炎图像的识别分类；AlexNet_S 是利用两种结构、深度不同的神经网络模型 AlexNet 与 InceptionV3，使用知识蒸馏方法得到的优化模型；DenseNet′是一种改进的 DenseNet 模型算法，在 DenseNet 深度模型的基础上，在全连接层加入中心损失，在最后输出部分将交叉熵损失函数替换为 Focal 损失。

实验结果如表 4 所示，因所用评价指标不同，选择共同的评价识别准确率来比较模型性能。从表 4 可以看出，在与其他深度神经网络模型的对比中，所提模型在识别准确率评价指标上取得最优，表明了其有效性。

表 4　与现有的研究成果的对比

Table 4　Comparison with existing research results　unit: %

Model	Accuracy
DenseNet′	90.46
AlexNet_S	94.87
GIV3	96.77
Proposed model	97.26

4　结　　论

首先通过 ResUNet 对儿童胸片进行分割，分割出左右肺部，然后利用改进的卷积神经网络 R32 提取分割后的图像特征，并将其输入 Transformer 模型，实现儿童肺炎图像的识别分类。实验结果表明，所提模型的准确率、精确率和召回率分别达到了 97.26%、97.69% 和 98.60%，可有效提高基层儿童肺炎临床诊断的准确率。且所提模型在测试集上表现出比其他已有模型更好的识别效果。ViT 模型是一种很简单但很灵活的方法，将其抽象为一系列嵌入，可应用于任何类型的数据。在后续工作中，在肺炎诊断模型的完善下，可构建多分类诊断模型判断肺炎致病原类型为病毒还是细菌，进一步规范儿童肺炎临床用药。

参　考　文　献

[1] McAllister D A, Liu L, Shi T, et al. Global, regional, and national estimates of pneumonia morbidity and mortality in children younger than 5 years between 2000 and 2015: a systematic analysis[J]. The Lancet Global Health, 2019, 7(1): e47-e57.

[2] Chaves G S, Freitas D A, Santino T A, et al. Chest physiotherapy for pneumonia in children[J]. The Cochrane Database of Systematic Reviews, 2019, 1(1): CD010277.

[3] Shiraishi J, Li Q, Appelbaum D, et al. Computer-aided diagnosis and artificial intelligence in clinical imaging[J]. Seminars in Nuclear Medicine, 2011, 41(6): 449-462.

[4] Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning[EB/OL]. (2017-11-14) [2021-04-08]. https://arxiv.org/abs/1711.05225.

[5] Wang X S, Peng Y F, Lu L, et al. ChestX-Ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 3462-3471.

[6] Varshni D, Thakral K, Agarwal L, et al. Pneumonia detection using CNN based feature extraction[C]//2019 IEEE International Conference on Electrical, Computer and Communication Technologies, February 20-22, 2019, Coimbatore, India. New York: IEEE Press, 2019.

[7] Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, MA. New York: IEEE Press, 2015.

[8] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[9] Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 2261-2269.

[10] Kermany D S, Goldbaum M, Cai W J, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning[J]. Cell, 2018, 172(5): 1122-1131.e9.

[11] 梁高博. 基于深度学习的儿科肺炎辅助诊断算法研究[D]. 泉州: 华侨大学, 2020.
Liang G B. Study of pediatric pneumonia assisted diagnosis algorithm based on deep learning[D]. Quanzhou: Huaqiao University, 2020.

[12] Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[13] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale[EB/OL]. (2020-10-22) [2021-04-05]. https://arxiv.org/abs/2010.11929.

[14] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems, December 4-9, 2017, Long Beach, California, USA. Massachusetts: The MIT Press, 2017: 5998-6008.

[15] Diakogiannis F I, Waldner F, Caccetta P, et al. ResUNet-A: a deep learning framework for semantic segmentation of remotely sensed data[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 162: 94-114.

[16] Jaeger S, Karargyris A, Candemir S, et al. Automatic tuberculosis screening using chest radiographs[J]. IEEE Transactions on Medical Imaging, 2014, 33(2): 233-245.

[17] Candemir S, Jaeger S, Palaniappan K, et al. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration[J]. IEEE Transactions on Medical Imaging, 2014, 33(2): 577-590.

[18] Voita E, Talbot D, Moiseev F, et al. Analyzing multi-head self-attention: specialized heads do the heavy lifting, the rest can Be pruned[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, July 28-August 2, 2019, Florence, Italy. Stroudsburg: Association for Computational Linguistics, 2019: 5797-5808.

[19] Gaudart J, Giusiano B, Huiart L. Comparison of the performance of multi-layer perceptron and linear regression for epidemiological data[J]. Computational Statistics & Data Analysis, 2004, 44(4): 547-570.

[20] Ba J L, Kiros J R, Hinton G E. Layer normalization[EB/OL]. (2016-07-21)[2021-05-06]. https://arxiv.org/abs/1607.06450.

[21] Qiao S Y, Wang H Y, Liu C X, et al. Micro-batch training with batch-channel normalization and weight standardization[EB/OL]. (2019-03-25) [2021-04-05]. https://arxiv.org/abs/1903.10520.

[22] Wu Y, He K. Group normalization[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11217: 3-19.

[23] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift [C]//ICML'15: Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37, July 6-11, 2015, Lille, France. New York: ACM Press, 2015: 448-456.

[24] Amid E, Warmuth M K, Anil R, et al. Robust Bi-tempered logistic loss based on bregman divergences[EB/OL]. (2019-06-08)[2021-08-09]. https://arxiv.org/abs/1906.03361.

[25] Newey W K. Adaptive estimation of regression models via moment restrictions[J]. Journal of Econometrics, 1988, 38(3): 301-339.

[26] 何新宇, 张晓龙. 基于深度神经网络的肺炎图像识别模型[J]. 计算机应用, 2019, 39(6): 1680-1684.
He X Y, Zhang X L. Pneumonia image recognition model based on deep neural network[J]. Journal of Computer Applications, 2019, 39(6): 1680-1684.

[27] 魏榕剑, 邵剑飞. 基于改进DenseNet网络的肺炎X光图像识别算法[J]. 电视技术, 2021, 45(6): 140-143.
Wei R J, Shao J F. Pneumonia X-ray image recognition algorithm based on improved DenseNet network[J]. Video Engineering, 2021, 45(6): 140-143.

[28] 邓棋, 雷印杰, 田锋. 用于肺炎图像分类的优化卷积神经网络方法[J]. 计算机应用, 2020, 40(1): 71-76.
Deng Q, Lei Y J, Tian F. Optimized convolutional neural network method for classification of pneumonia images[J]. Journal of Computer Applications, 2020, 40(1): 71-76.