激光与光电子学进展 先进成像

基于双路径交叉融合网络的肺结节CT图像分类方法

杨萍，张鑫*，温帆，田吉，何宁
北京联合大学智慧城市学院，北京 100101

摘要 针对肺结节计算机断层(CT)图像具有的细节多样性以及类间相似性的问题，构建了一种集卷积神经网络(Convolutional neural network, CNN)和Transformer优势的双路径交叉融合网络对肺结节进行更精确的分类。首先，以窗口多头自注意力和滑动窗口多头自注意力为基础，构建全局特征块，用于捕获结节的形态特征；以大核注意力为基础构建局部特征块，用于提取结节的纹理、密度等内部特征。其次，设计特征融合块用于融合上一阶段的局部与全局特征，使每一条路径都能获得更综合的判别信息。然后，引入KL(Kullback-leibler)散度来增加不同尺度特征之间的分布差异性，优化网络性能。最后，采用决策层融合的方法获得分类结果。在LIDC-IDRI数据集上进行实验，网络的分类准确率、召回率、精确率、特异性、受试者操作特征(ROC)曲线下的面积(Area under curve, AUC)分别为94.16%、93.93%、93.03%、92.54%、97.02%。实验结果表明，所提方法具有较好的肺结节良恶性分类能力。

关键词 肺结节良恶性分类；CT图像；局部-全局特征；Transformer；注意力机制
中图分类号 TP391 **文献标志码** A **DOI**: 10.3788/LOP231413

Pulmonary Nodule Computed Tomography Image Classification Method Based on Dual-Path Cross-Fusion Network

Yang Ping, Zhang Xin*, Wen Fan, Tian Ji, He Ning
Smart City College, Beijing Union University, Beijing 100101, *China*

Abstract Pulmonary nodule computed tomography (CT) images have diverse details and interclass similarity. To address this problem, a dual-path cross-fusion network combining the advantages of convolutional neural network (CNN) and Transformer is constructed to classify pulmonary nodules more accurately. First, based on windows multi-head self-attention and shifted windows multi-head self-attention, a global feature block is constructed to capture the morphological features of nodules; then, a local feature block is constructed based on large kernel attention, which is used to extract internal features such as the texture and density of nodules. A feature fusion block is designed to fuse local and global features of the previous stage so that each path can collect more comprehensive discriminative information. Subsequently, Kullback-Leibler (KL) divergence is introduced to increase the distribution difference between features of different scales and optimize network performance. Finally, a decision-level fusion method is used to obtain the classification results. Experiments are conducted on the LIDC-IDRI dataset, and the network achieves a classification accuracy, recall, precision, specificity, and area under curve (AUC) of 94.16%, 93.93%, 93.03%, 92.54%, and 97.02%, respectively. Experimental results show that this method can classify benign and malignant pulmonary nodules effectively.

Key words classification of benign and malignant pulmonary nodules; CT image; local-global feature; Transformer; attention mechanism

1 引 言

《2020年世界癌症报告》显示，全球肺癌的发病率为11.4%，在所有癌症中排行第二；由肺癌导致的死亡病例达到180万例，致死率占比18%，在所有癌症中排行第一，肺癌已成为威胁人类健康的第一癌症杀手[1]。虽然肺癌的发病率和致死率较高，但有关研究显示在肺癌早期的检测和诊断可有效降低死亡率及延长患者的生存时间[2]。早期的肺癌检查主要以诊断肺结节为主，从结节内部看，良性肺结节的纹理和密度比

收稿日期：2023-05-30；**修回日期**：2023-06-23；**录用日期**：2023-07-24；**网络首发日期**：2023-08-18
基金项目：国家自然科学基金(62172045, 62272049)
通信作者：*1254211375@qq.com

恶性肺结节更加均匀；从形态上看，良性肺结节边缘规整，恶性肺结节则具有毛刺、分叶等不规则的特点[3]，且有发展成为恶性肿瘤进而引发肺癌的可能性。

由于胸部CT（Computed tomography）图像成像环境复杂，目标结构多变，因此提取判别性特征是实现肺结节良恶性分类的关键步骤之一。相较于手工设计的特征提取方法，深度学习可以自动捕捉肺结节CT图像中层次化的特征表示，在提取更具表达能力的肺结节特征方面表现出明显的优势。

为了让网络学习到不同尺度下的肺结节特征，Shen等[4]和Xu等[5]从原始CT图像中裁剪出3种尺寸大小的感兴趣区域，并分别输入到3个网络中进行训练，以提取多尺度和多样性的肺结节特征。虽然这种方法取得了不错的分类效果，但是需要为每个尺度的输入图像设计一个单独的网络，增加了网络的复杂性。为了充分利用结节空间信息，Xie等[6]将提取到的3D肺结节CT图像的9种视角的图像切片作为输入，并联合多个基于知识协作的子网络对肺结节进行分类，以获取更丰富的结节信息。相较于直接使用3D图像作为输入，采用提取多视角图像切片的方法，可以在不增加网络参数量的同时提高网络对肺结节的表达能力。上述讨论的方法都是简单地通过堆叠卷积和池化等操作实现对输入数据的特征提取和降维，为了让网络能够关注到重要的图像区域，减少冗余信息。有一些方法引入注意力机制来提高网络的分类性能。Jiang等[7]构建了一个结合上下文注意力和空间注意力机制的3D双路径网络，分类准确率达到90.24%。同时另一种以ResNeXt网络为骨干，嵌入通道注意力机制的多层级特征融合网络也被用于肺结节良恶性分类[8]。以上基于卷积神经网络（Convolutional neural network, CNN）的方法，通过增加网络层数来扩大感受野进而提取图像的全局特征。然而，随着网络加深，反复的卷积和池化操作会导致图像分辨率下降，在一定程度上将损失肺结节的形状、大小等特征，进而丢失从形态上进行良恶性分类的依据。

Transformer[9]结构中的多头自注意力（Multi-head self-attention, MSA）机制具有长距离依赖的特性，可以对全局信息进行关注和利用。Liu等[10]通过对特征图进行拉平，引入Transformer中的编码器用于提取肺结节的全局特征。同时，提出了一种专门用于图像处理任务的Vision Transformer（ViT）[11]，它将图像块视为一系列的向量序列，然后使用MSA提取图像的全局特征。Wang等[12]则在ViT的基础上，提出了一种改进ViT的肺结节分类网络，引入2D平移滑动窗口，通过提取更紧凑的特征索引以及权重学习对角矩阵来解决过拟合的问题，在LIDC-IDRI数据集上进行实验，受试者操作特征（ROC）曲线下的面积（Area under curve, AUC）达到了96.04%。此外，由于Swin Transformer[13]具有捕获全局上下文信息且降低计算量的优势，Sun等[14]通过实验证明了Swin Transformer在肺结节分类任务上有较好的分类性能。Transformer网络在图像处理中较少考虑像素之间的局部空间信息，导致在捕捉肺结节CT图像的局部特征和细节方面相对较弱。

考虑到肺结节的形态复杂，且良恶性肺结节之间存在一定的相似性，不仅需要对局部信息进行建模，用于描述肺结节的纹理、密度等内部细节特征，还需要捕获全局特征用于描述肺结节的边缘毛刺、大小等形态特征，这与医生在临床上诊断肺结节良恶性的方式是一致的，所以如何同时准确捕获肺结节局部和全局特征对肺结节的良恶性诊断非常重要。本文利用CNN获取局部特征的优势来捕获肺结节内部的纹理和密度等信息，使用Transformer的长距离依赖优势来获取肺结节的形态信息，并将二者进行融合，设计了一个集CNN和Transformer优势的双路径交叉融合网络，以获取更丰富和多样性的肺结节特征，同时将特征优化后的散度损失函数KL（Kullback-leibler）与交叉熵损失函数（Cross entropy loss）相结合，构成混合损失函数以提高网络的辨别能力。

2 所提方法

2.1 总体结构

将3 pixel×224 pixel×224 pixel大小的肺结节CT图像作为输入图像。总体网络结构如图1所示，由特征提取模块和分类模块构成。

特征提取模块由全局路径、局部路径、特征优化3个部分构成，其中全局和局部路径包括4个阶段，每阶段特征图大小分别为96×56×56、192×28×28、384×14×14、768×7×7，单位为pixel，各部分结构的详细说明如下。

1）全局路径：首先，图像经过patch partition后被划分为56×56个图像块，每个图像块由4×4个相邻像素构成。然后，由linear embedding和全局特征块构成第一阶段，linear embedding将每个图像块沿通道方向拉平，全局特征块则用于提取肺结节的形态特征。第二、三、四阶段均由特征融合块、patch merging以及全局特征块顺序排列而成，其中特征融合块用于融合上一阶段的局部与全局特征，patch merging在对图像进行下采样的同时使通道数翻倍，从而更好地帮助网络捕获图像中的空间信息。

2）局部路径：第一阶段包含下采样层和局部特征块，其中下采样层使用核大小为7、步长为4的卷积，将图像从3 pixel×224 pixel×224 pixel变换为96 pixel×56 pixel×56 pixel，局部特征块用于提取肺结节内部的纹理、密度等细节信息。后三个阶段均由特征融合块、下采样层和局部特征块组成，与第一阶段的下采样操作不同，后三个阶段的下采样层使用核大小为3、步长为2的卷积，保证每一个阶段得到的特征图大小与全

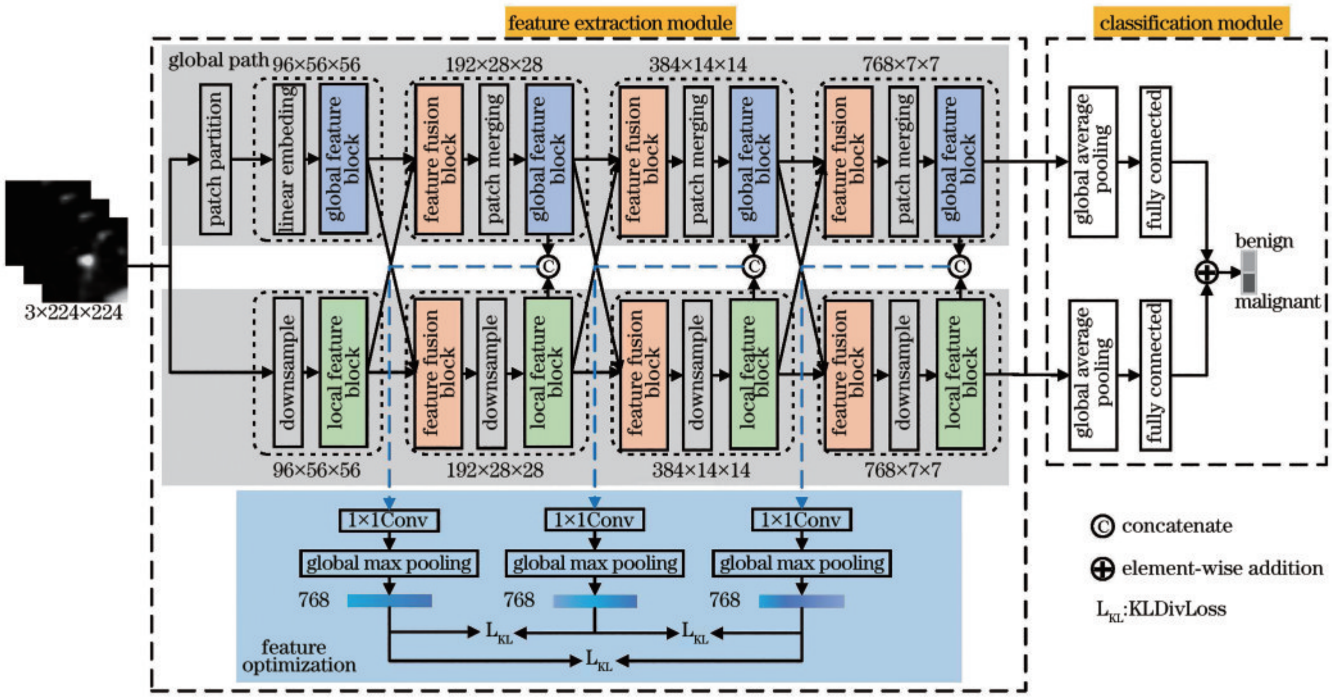

图 1　双路径交叉融合网络的总体结构图

Fig. 1　Overall structure diagram of dual path cross fusion network

局路径中的一致。

3）特征优化：通过最大化 KL 散度来调整不同阶段提取到的特征分布，增加不同阶段所捕获特征的差异性，以帮助网络提取到更加多样性的特征，优化网络性能。由于网络在第一阶段所提取到的判别信息太少，具有局限性，因此只计算后三个阶段所捕获的特征向量之间的 KL 散度。在计算 KL 散度之前，首先将每个阶段获取的全局与局部特征图进行通道拼接得到新的特征图；其次，再使用 1×1 卷积对特征进行整合再利用；然后，使用全局最大池化操作得到长度为 768 的特征向量；最后，计算每两个特征向量之间的 KL 散度，总 KL 散度损失为

$$L_{\mathrm{KL}}=-\sum_{i=1}^{2}\sum_{j=i+1}^{3}\sum_{x=0}^{L-1}\boldsymbol{p}_i(x)\ln\frac{\boldsymbol{p}_i(x)}{\boldsymbol{p}_j(x)},\qquad(1)$$

式中：$\boldsymbol{p}$ 为不同阶段的输出特征向量；L 为特征向量长度；i,j 为求和符号的下限。

分类模块采用等权重决策层融合[15]的方法融合两个路径的分类结果，将融合结果用于模拟临床上医生从不同角度对肺结节进行诊断，以获取多方面的判定结果。所提方法使用全局平均池化操作对特征进行变换和处理，使用全连接层进行结果的预测。

2.2　全局特征块

从全局出发，通过全局特征块可有效提取肺结节的形状、大小等形态信息，以获取肺结节之间明显的视觉差异性特征。以 Swin Transformer 中的窗口多头自注意力（Windows multi-head self-attention，W-MSA）和滑动窗口多头自注意力（Shifted windows multi-head self-attention，SW-MSA）为基础构建全局特征块，图 2 展示了全局特征块的结构。

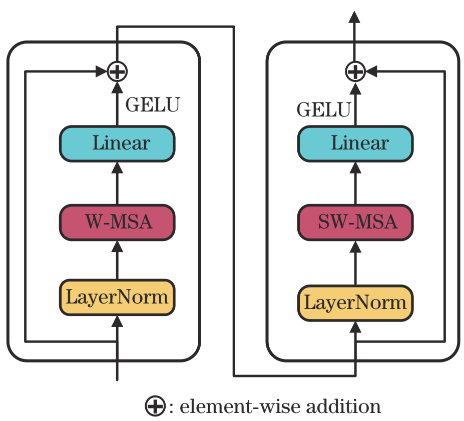

图 2　全局特征块

Fig. 2　Global feature block

相对于 ViT 结构中的 MSA，W-MSA 是在 $M\times M$ 大小的窗口内进行 MSA 操作，可有效将 MSA 的平方计算复杂度降低到线性计算复杂度。MSA 和 W-MSA 的计算复杂度表达式为

$$\Omega(\mathrm{MSA})=4HWC^2+2(HW)^2C,\qquad(2)$$

$$\Omega(\mathrm{W\text{-}MSA})=4HWC^2+2M^2HWC,\qquad(3)$$

式中：H 和 W 分别为图像的高和宽；C 为特征图的通道数；M 为窗口大小，设置 $M=7$。

在全局特征块的第一阶段，特征图经过层归一化 LN（LayerNorm）后进入 W-MSA，然后依次通过线性层（Linear）、高斯误差线性单元（Gaussian error linear units，GELU）激活函数进行处理。处理后的特征图

进入带有SW-MSA的第二阶段，通过SW-MSA可以实现窗口与窗口之间的信息交互，从而提取全局上下文信息。并在每一个阶段中添加残差连接，以保留更多的原始信息，有助于网络的梯度传播。式(4)描述了以上过程，设定 $\boldsymbol{X}_{\mathrm{G}} \in \mathbb{R}^{C \times L}$ 为输入特征图，L 为 $H \times W$，

$$\begin{cases} \boldsymbol{G}_1 = \sigma\left\{f\left\{\text{W-MSA}\left[\mathrm{LN}(\boldsymbol{X}_{\mathrm{G}})\right]\right\}\right\} + \boldsymbol{X}_{\mathrm{G}} \\ \boldsymbol{G}_2 = \sigma\left\{f\left\{\text{SW-MSA}\left[\mathrm{LN}(\boldsymbol{G}_1)\right]\right\}\right\} + \boldsymbol{G}_1 \end{cases}, \tag{4}$$

式中：$\boldsymbol{G}_1 \in \mathbb{R}^{C \times L}$ 和 $\boldsymbol{G}_2 \in \mathbb{R}^{C \times L}$ 分别为全局特征块的第一阶段和第二阶段的输出结果；LN为LayerNorm操作；f为线性操作；σ为GELU激活函数。

2.3　局部特征块

从局部出发，通过局部特征块可有效提取肺结节的纹理、密度等内部细节信息，以获取肺结节之间的细微的差异性特征。通过使用适当大小的卷积核能够捕获区域内的局部上文信息，标准的卷积操作采用静态权重，对通道维度上的适应性较差，同时在增加卷积核大小时，参数量也将大幅增加，但大核注意力(Large kernel attention，LKA)[16]可以很好地解决以上问题，不仅具有局部感受野，而且可以在空间和通道维度上实现更好的自适应，因此在局部路径中引入LKA构建局部特征块。图3展示了局部特征块的结构。

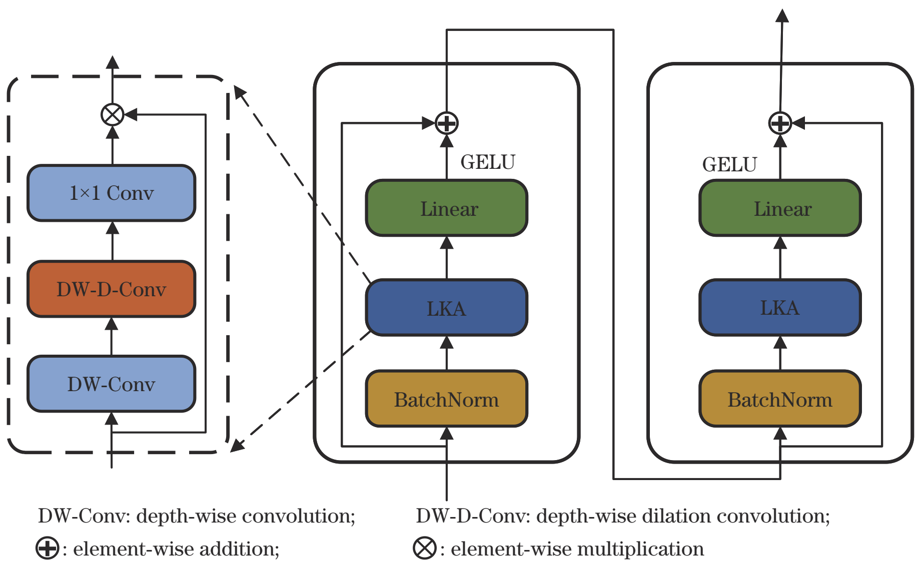

图3　局部特征块

Fig. 3　Local feature block

LKA可以将一个 $K \times K$ 卷积分解为具有空间局部性质、$(2d-1)\times(2d-1)$ 大小的深度卷积；具有空间长距离性质、$\left\lceil \frac{K}{d} \right\rceil \times \left\lceil \frac{K}{d} \right\rceil$ 大小的深度扩张卷积；在通道维度上操作的 1×1 卷积。其中，d 为膨胀率。为了与全局特征块中的窗口大小保持一致，设定 $K=7$，$d=2$。通过以上分解操作，LKA可以在较低的计算成本下建立局部领域内的像素点之间的关系，捕获更丰富的局部细节信息，具体实现过程的公式化表述为

$$\begin{cases} \boldsymbol{A}_{\text{Attention}} = \mathrm{Conv}_{1\times1}\{\mathrm{DW\text{-}D\text{-}}\\ \quad \mathrm{Conv}\left[\mathrm{DW\text{-}Conv}(\boldsymbol{X}_{\mathrm{LKA}})\right]\}, \\ \boldsymbol{A}_{\text{Output}} = \boldsymbol{A}_{\text{Attention}} \times \boldsymbol{X}_{\mathrm{LKA}} \end{cases} \tag{5}$$

式中：$\boldsymbol{X}_{\mathrm{LKA}} \in \mathbb{R}^{C \times H \times W}$ 为输入特征图；$\boldsymbol{A}_{\text{Attention}} \in \mathbb{R}^{C \times H \times W}$ 为生成的注意力图；$\boldsymbol{A}_{\text{Output}} \in \mathbb{R}^{C \times H \times W}$ 为输出结果。

类似于全局特征块的处理流程，特征图在第一个局部特征块，经批量归一化BN(BatchNorm)后进入LKA，然后经过线性层和GELU激活函数后，进入第二个局部特征块，用公式表示为

$$\begin{cases} \boldsymbol{L}_1 = \sigma\left\{f\left\{\mathrm{LKA}\left[\mathrm{BN}(\boldsymbol{X}_{\mathrm{L}})\right]\right\}\right\} + \boldsymbol{X}_{\mathrm{L}} \\ \boldsymbol{L}_2 = \sigma\left\{f\left\{\mathrm{LKA}\left[\mathrm{BN}(\boldsymbol{L}_1)\right]\right\}\right\} + \boldsymbol{L}_1 \end{cases}, \tag{6}$$

式中：$\boldsymbol{X}_{\mathrm{L}} \in \mathbb{R}^{C \times H \times W}$ 为输入特征图；$\boldsymbol{L}_1 \in \mathbb{R}^{C \times H \times W}$ 和 $\boldsymbol{L}_2 \in \mathbb{R}^{C \times H \times W}$ 分别为局部特征块第一阶段和第二阶段的输出结果；BN为BatchNorm操作；LKA的操作如式(4)所示。

2.4　特征融合块

为实现全局信息与局部信息在网络中交互，从而获得更加丰富和具有区分度的特征表达信息，提高分类性能，在全局路径和局部路径中加入了特征融合块，并将一种轻量级的无参注意力机制[17]引入到特征融合块中，以增强网络对重要特征的敏感性。图4展示了特征融合块的结构，其中，$\mathrm{Avg}_{\mathrm{C}}$ 表示沿通道维度的平均池化，$\mathrm{Avg}_{\mathrm{HW}}$ 表示沿空间维度的平均池化。

输入的 $\boldsymbol{G}$ 和 $\boldsymbol{L}$ 分别代表来自上一阶段的全局特征图和局部特征图。$\boldsymbol{G}$ 和 $\boldsymbol{L}$ 先通过元素相加的方式进行融合得到特征图 $\boldsymbol{F} \in \mathbb{R}^{C \times H \times W}$，然后进行通道维度和空

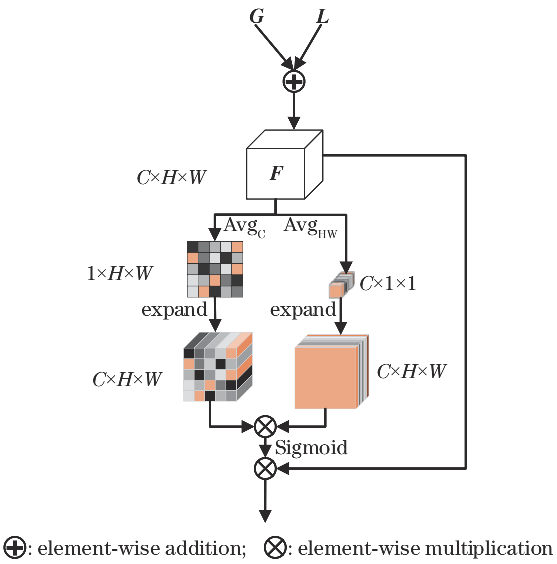

图 4　特征融合块

Fig. 4　Feature fusion block

间维度上的平均池化，生成空间注意力图 $\boldsymbol{A}_{\mathrm{sp}} \in \mathbb{R}^{1\times H\times W}$ 和通道注意力图 $\boldsymbol{A}_{\mathrm{ch}} \in \mathbb{R}^{C\times 1\times 1}$，

$$\boldsymbol{A}_{\mathrm{sp}}(x_{H\times W}) = \frac{1}{C}\sum_{i=1}^{C} x_{H\times W}(i), \tag{7}$$

$$\boldsymbol{A}_{\mathrm{ch}}(y_C) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} y_C(i,j), \tag{8}$$

式中：$x_{H\times W} \in \mathbb{R}^{C}$ 代表每个空间元素；$y_C \in \mathbb{R}^{H\times W}$ 代表每个通道。

随后将得到的注意力图通过广播机制沿着缩小的维度扩展到原始大小，并以元素相乘操作重新组合，最后使用 Sigmoid 激活函数来增强特征的非线性变换能力，并将得到的结果作用于特征图 F。以上整个过程描述为

$$\begin{cases} \boldsymbol{F} = \boldsymbol{G} + \boldsymbol{L} \\ \boldsymbol{F}' = \omega(\boldsymbol{A}_{\mathrm{sp}} \times \boldsymbol{A}_{\mathrm{ch}}) \times \boldsymbol{F} \end{cases}, \tag{9}$$

式中：ω 为 Sigmoid 激活函数；$\boldsymbol{F}' \in \mathbb{R}^{C\times H\times W}$ 为特征融合块的输出结果。

通过以上操作可以在不增加网络参数量的同时实现通道和空间注意力的功能提升，从而可以在不同通道和位置之间提取重要的融合信息，增强网络对不同特征的感知能力。

2.5　混合损失函数

将交叉熵损失函数和 KL 散度损失函数结合构成一个混合损失函数，不但可以计算网络预测错误的损失，而且对特征提取过程能起到深层监督作用，弥补单一损失函数的不足。混合损失函数为

$$L = \alpha L_{\mathrm{CE}} + \beta L_{\mathrm{KL}}, \alpha + \beta = 1, \tag{10}$$

式中：α 和 β 为平衡因子；L_{CE} 为总交叉熵损失函数；L_{KL} 为 KL 散度损失函数。实验证明当 α 取 0.9、β 取 0.1 时，所提网络分类效果最好。由于所提网络最终预测的结果来自两条路径，因此分类错误的总交叉熵损失函数 L_{CE} 表示为

$$\begin{cases} L_{\mathrm{CE}}^{j} = -\dfrac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{k} q_i \ln(y_i), j = 1, 2 \\ L_{\mathrm{CE}} = \sum_{j=1}^{2} L_{\mathrm{CE}}^{j} \end{cases}, \tag{11}$$

式中：N 为样本总数；k 为类别数；q_i 为真实标签；y_i 为预测概率。总 KL 散度损失 L_{KL} 的计算如式（1）所示。

3　数据集与实验设置

3.1　数据预处理

使用公共数据集 LIDC-IDRI[18] 作为训练和测试数据，包含 1010 位病人的 1018 个 CT 扫描和记录专家标注信息的 XML（Extensible markup language）文件。由于 CT 图像来自不同的机构和设备，需要将 CT 图像重采样至 1 mm×1 mm×1 mm，以保证数据的一致性。每个肺结节由 1 到 4 位专家标注恶性程度，恶性程度的值从 1 到 5。为了确保大多数专家同意肺结节的存在，选取至少有 3 位专家标注的肺结节，且根据专家标注的恶性程度的中位数来划分肺结节良恶性。当中位数小于 3 时，判定肺结节为良性；大于 3 时为恶性；等于 3 时说明专家不确定肺结节的良恶性，将其排除掉。最后共得到 848 幅肺结节 CT 图像，其中包括 442 幅良性肺结节和 406 幅恶性肺结节图像。同时对所获取的数据进行了以下预处理操作：

1）以肺结节为中心，裁剪 64 pixel×64 pixel×64 pixel 大小的图像块，以去除掉 CT 图像中大量的噪声。

2）从轴向、冠状和矢量面三个角度提取 2D 切片[19]，以获取 CT 图像多个视角的空间信息。

3）对每个角度的 2D 切片进行数据增广操作，包括水平翻转、旋转（90°，180°，270°）、添加高斯噪声等操作，以防止网络过拟合，提高网络的泛化能力。

4）采用双线性插值方法将图像放大到 224 pixel×224 pixel，并将每幅图像复制 3 份，在通道维度上合并成 3 pixel×224 pixel×224 pixel 的图像，最后根据十折交叉验证的方法划分数据集。以上预处理过程如图 5 所示。

3.2　评价标准

采用准确率（Accuracy；R_{A}）、召回率（Recall；R）、精确率（Precision；P）、特异性（Specificity；R_{S}）、ROC 曲线下的面积 AUC 等 5 个指标来评价网络的分类性能：

$$R_{\mathrm{A}} = \frac{N_{\mathrm{TP}} + N_{\mathrm{TN}}}{N_{\mathrm{TP}} + N_{\mathrm{TN}} + N_{\mathrm{FP}} + N_{\mathrm{FN}}}, \tag{12}$$

R_{A} 表示正确区分良恶性结节的个数占总共结节个数的比例。

$$R = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FN}}}, \tag{13}$$

R 表示所有恶性结节中被识别为恶性的比例。

$$P = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}}}, \tag{14}$$

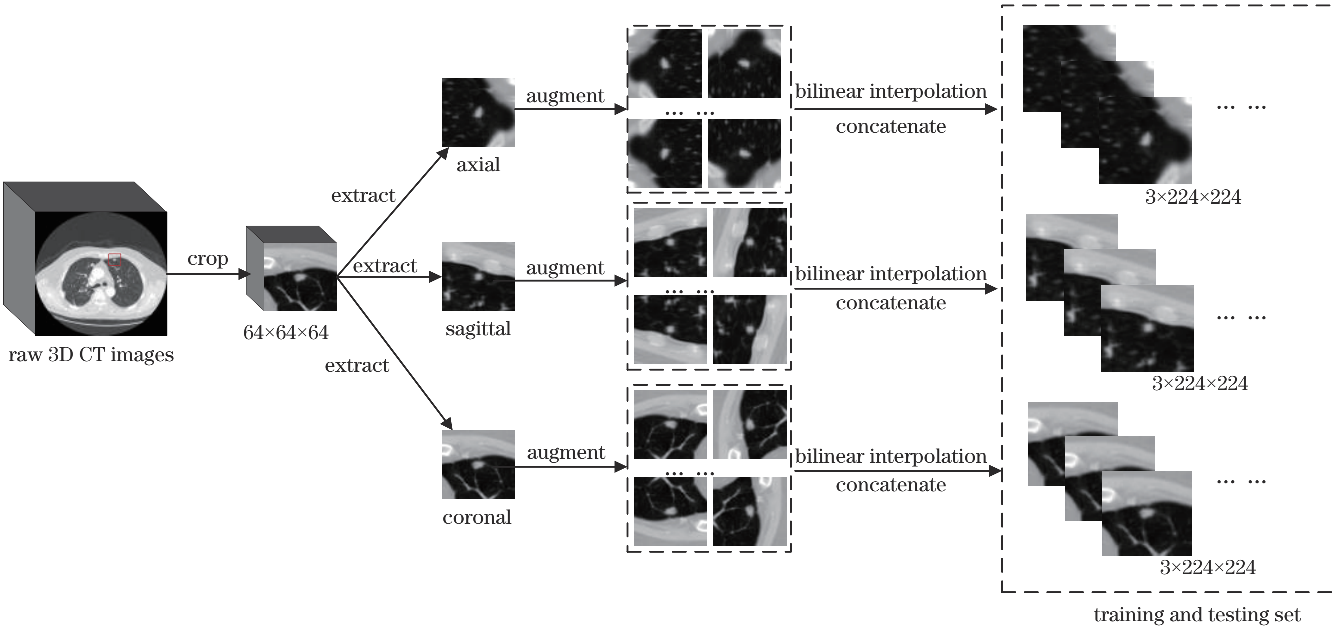

图 5　数据预处理过程

Fig. 5　Data preprocessing process

P 表示预测为恶性结节的样本中实际为恶性结节的比例。

$$R_S = \frac{N_{TN}}{N_{TN} + N_{FP}}, \tag{15}$$

R_S 表示所有良性结节中被识别为良性的比例。

以上公式中 N_{TP} 代表真阳性数量、N_{TN} 代表真阴性数量、N_{FP} 代表假阳性数量、N_{FN} 代表假阴性数量。

3.3　参数设置

采用 Python3.9、PyTorch1.12 框架进行编程，在 32 GB 内存、NVIDIA GeForce RTX 3090Ti、Intel Core i5-12600KF 3.70 GHz CPU 处理器和 Window 10 系统上进行网络的训练与测试，其他详细参数设置如表 1 所示。

表 1　参数设置

Table 1　Parameter setting

Training parameter	Parameter value
Batch size	64
Train epochs	50
Optimizer	AdamW
Learning rate	10^{-4}
Learning rate schedule	Cosine annealing
Optimizer momentum	β_1, β_2=0.9, 0.9999
Weight decay	0.01

4　实验结果与分析

4.1　不同模型的对比

为了验证所提方法的有效性，对比所提方法与图像分类领域中的 5 个经典模型（ResNet50[20]、DenseNet121[21]、ConvNeXt-Base[22]、ViT-Base[11]、Swin Transformer-Base[13]）在 LIDC-IDRI 数据集上的分类性能，其中 ResNet50、DenseNet121、ConvNeXt 是基于 CNN 的模型，ViT 和 Swin Transformer 是基于 Transformer 的模型。各项分类指标如表 2 所示，表 3 为不同模型的参数量、FLOPs（Floating point operations）以及推理时间的对比结果，图 6 则通过 ROC 曲线下的面积 AUC 展示了不同模型在肺结节良恶性分类任务上的表现。

表 2　不同模型的分类指标的对比

Table 2　Comparison of classification metrics of different models

unit: %

Model	Accuracy	Recall	Precision	Specificity	AUC
ResNet50	86.21	83.99	87.04	88.25	91.28
DenseNet121	84.91	81.00	86.69	88.40	90.75
ConvNeXt-Base	89.90	89.70	89.48	90.00	94.60
ViT-Base	90.44	89.66	90.36	91.18	94.74
Swin Transformer-Base	90.34	89.50	90.32	91.10	94.59
Proposed method	**94.16**	**93.93**	**93.03**	**92.54**	**97.02**

从表 2 和表 3 的综合结果可以看出，与其他模型相比，所提方法在肺结节良恶性分类任务中能够以较少的参数量实现更高的分类性能，分类准确率达到了 94.16%。同时，所提方法以识别一幅图片仅耗费 10.3 ms 的优势在推理时间上展现出较强的竞争力，这表明该方法具有高效性，可以快速地处理与分析输入数据。一般来说，FLOPs 越大，模型的复杂度越高，推理时间也会相应增加，但通过分析不同模型的 FLOPs 和推理时间之间的关系可以得出，仅根据

表 3 不同模型的参数量、FLOPs、推理时间的对比

Table 3 Comparison of parameter number, FLOPs and inference time of different models

Model	Parameter /M	FLOPs /G	Inference time /ms
ResNet50	23.5	4.1	13.7
DenseNet121	**7.0**	**2.9**	19.3
ConvNeXt-Base	87.6	15.4	15.3
ViT-Base	85.8	16.9	13.8
Swin Transformer-Base	86.7	15.2	19.9
Proposed method	28.2	4.0	**10.3**

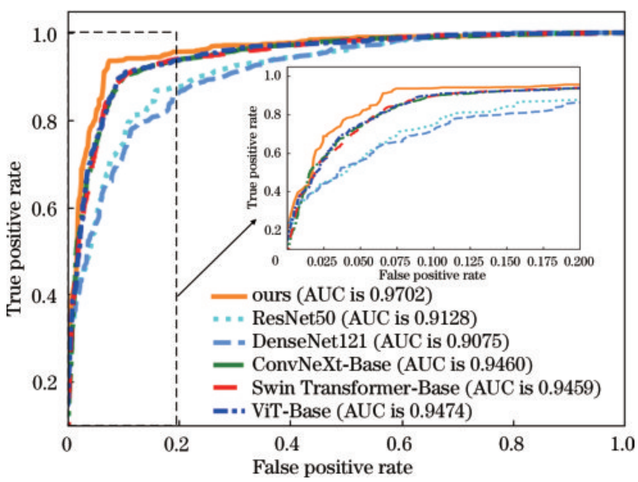

图 6 不同模型的 ROC 曲线

Fig. 6 ROC curves of different models

FLOPs 来比较模型的推理时间是不准确的，需要考虑其他因素的影响，比如模型结构、优化策略等。

与所对比的模型类似，所提方法构建了一种具有层次化结构的网络用于提取不同级别的特征，实现逐级的信息传递与处理，但相较于纯 CNN 或 Transformer 结构，所提方法提出的集 CNN 和 Transformer 优势的混合网络结构，可以同时对局部特征和全局上下文信息进行建模，从而有效捕获肺结节的纹理、密度等内部特征以及形状、大小等形态特征，提高网络的分类能力。从图 6 也可以看出，所提方法的 AUC 值达到了 97.02%，整体性能要优于其他模型。

4.2 不同研究方法的对比

表 4 展示了所提方法与近期其他学者的研究结果对比，为了公平起见，与所提方法对比的研究方法均采用 LIDC-IDRI 数据集进行训练和测试，同时所提方法与文献[10]、文献[23]使用相似的预处理方法和相同的数据集规模。从表中数据可以看出，所提方法在准确率、召回率、精确率、AUC 等 4 个指标中取得明显优势，反映出所提网络相对比其他方法具有更强大的特征提取能力，可以更好地对良性和恶性肺结节进行分类。文献[10]利用 CNN 与 Transformer 设计了一个用于提取局部和全局特征的串行网络，从对比结果可以看出，所提方法的综合指标要优于文献[10]，说明所设计的并行网络结构可以更有效地融合局部与全局特征，提高网络对不同层次的细节和语义信息的理解能力。文献[12]提出了一种改进 ViT 的肺结节分类网络，但由于缺乏 CNN 固有的归纳偏置，降低了其在分类任务上的性能，而所提方法由卷积操作构成的局部路径，能够提取具有旋转、平移不变性的局部特征，弥补了 Transformer 结构对局部空间结构建模的不足。文献[23]使用纯 CNN 构建了一个用于提取局部与全局特征的分类网络，但其一直保持特征图分辨率不变，会导致特征重复和信息冗余，从而影响分类性能，因此该方法在肺结节良恶性分类任务上表现不佳。文献[24]采用了神经架构搜索的方式自动构建肺结节分类网络，虽然特异性比所提方法高出了 2.50 百分点，但是召回率仅为 85.37%，远低于所提方法的 93.93%，说明网络更容易将恶性结节判定为良性结节，这在临床实践中，会提高患者错过最佳治愈机会的风险。另外，所提方法在肺结节良恶性分类任务上的表现也优于文献[25-27]的方法。

表 4 与不同研究方法的对比

Table 4 Comparison with different research methods unit: %

Type	Method	Accuracy	Recall	Precision	Specificity	AUC
CNN	Ref. [23]	88.46	88.66	87.38	—	95.62
	Ref. [24]	90.77	85.37	—	95.04	—
	Ref. [25]	92.81	92.36	92.59	—	96.17
	Ref. [26]	91.07	90.93	91.18	91.22	95.84
	Ref. [27]	91.25	89.10	91.59	93.39	91.25
Transformer	Ref. [12]	93.33	—	—	—	96.04
CNN+Transformer	Ref. [10]	92.92	93.84	91.62	—	96.28
	Proposed method	94.16	93.93	93.03	92.54	97.02

4.3 平衡因子 α 和 β 的取值

图 7 展示了混合损失函数中 α 和 β 的不同取值对分类性能的影响。当仅使用交叉熵损失即 $\alpha=1$ 和 $\beta=0$ 时，分类准确率为 92.04%；当 $\alpha=0$ 和 $\beta=1$ 时，总损失仅由 KL 散度损失进行优化，网络不能收敛；当 $\alpha=0.9$ 和 $\beta=0.1$ 时，分类准确率最高达到 94.16%，相较

于仅使用交叉熵损失时，分类准确率提高了 2.12 百分点。从图 7 可以看出，引入 KL 散度损失并恰当设置其在总损失中的权重，可以优化网络以提高分类性能，但当 KL 散度损失权重增大时，可能会导致网络过分关注不同特征向量之间的差异性，而忽略掉因分类错误产生的损失，从而降低分类性能。

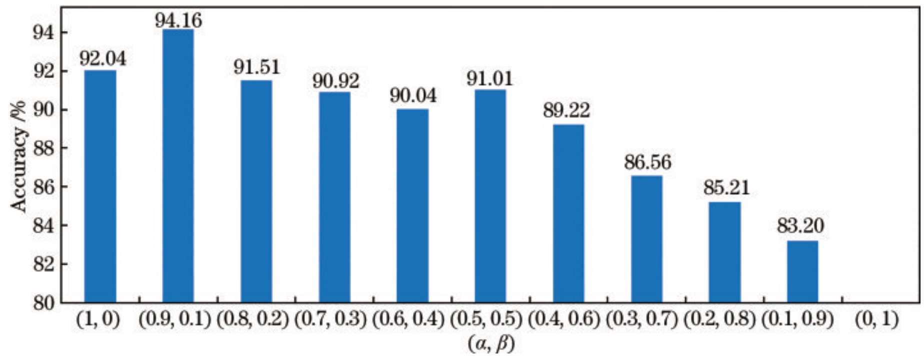

图 7　α 和 β 的取值对分类性能的影响

Fig. 7　Influence of α and β values on classification performance

4.4　消融实验

讨论网络各部分对分类性能的影响，消融实验结果如表 5 所示。M1 和 M2 表示将交叉熵作为损失函数，仅以局部路径或全局路径构成的分类网络时，其分类准确率分别为 85.42%、85.26%，其他各项指标在分类任务上也表现出不足。M3 和 M4 表示将 M1 和 M2 中的交叉熵损失函数替换为混合损失函数，分类效果得到改善，说明网络识别结节的能力得到提高，同时也证明相较于单一的损失函数，混合损失函数起到了深层监督作用，获得更好的训练效果。M5 表示联合局部路径和全局路径搭建的双路径网络用于肺结节良恶性的判定，分类准确率达到 90.89%，这表明，相较于 M1 和 M2 的单路径串行网络，双路径网络可以从局部和全局两个角度获取综合的结节特征信息，从而提高网络的判别能力。M6 相比于 M5，增加了特征融合块，分类性能进一步提高，这表明将局部和全局特征在分类过程中融合可以为网络提供更全面、更丰富的特征表示，证明了特征融合块的有效性。M7 的结果表明，所设计的网络分类性能达到最优，准确率、召回率、精确率、特异性、AUC 分别为 94.16%、93.93%、93.03%、92.54%、97.02%。另外，从实验结果也可以看出，所设计的网络在召回率和精确率之间达到了较高水平的平衡，可以有效地降低误判和漏判的风险，保证网络预测结果的可靠性。

表 5　消融实验结果

Table 5　Results of the ablation experiment　　unit: %

Model	Local path	Global path	Feature fusion block	Hybrid loss function	Accuracy	Recall	Precision	Specificity	AUC
M1	√				85.42	81.05	86.54	86.38	90.82
M2		√			85.26	83.97	85.42	86.43	90.56
M3	√			√	86.45	85.66	86.36	87.18	91.06
M4		√		√	87.39	86.86	86.44	86.95	91.61
M5	√	√			90.89	89.67	90.93	91.62	95.31
M6	√	√	√		92.04	92.61	90.56	91.65	96.23
M7	√	√	√	√	**94.16**	**93.93**	**93.03**	**92.54**	**97.02**

4.5　可视化结果分析

所提方法使用 Grad-CAM（Gradient-weighted class activation mapping）算法[28]生成热力图以理解网络的推理过程，网络对某些区域的感兴趣程度越高，在热力图中对应区域的颜色越深。图 8 展示了良恶性肺结节的 Grad-CAM 可视化和分类结果。

如图 8 所示，高热力值主要集中在结节的内部并逐渐扩散到边界，表明网络在进行特征提取的过程中不仅关注到了结节的纹理、密度等内部特征，还关注了结节的形状、大小等形态特征，这与医生在临床上诊断肺结节良恶性的关注点一致，证明所提方法能够更好地融合局部与全局信息，有助于网络学习更多的鉴别性特征以得到更精准的分类结果。从图 8 中可以发现，同一种类别的结节之间存在较大的差异，比如恶性结节“D”、“E”比结节“B”、“C”的形态更规则、密度更均匀。同时，良恶性结节之间在形态和内部特征上也存在一定的相似性，比如恶性结节“C”和良性结节“L”。但所提方法都能够高概率地正确识别出不同类型的结节，这表明所提方法在一定程度上解决了肺结节分类任务中所面临的类间相似性、类内差异性问

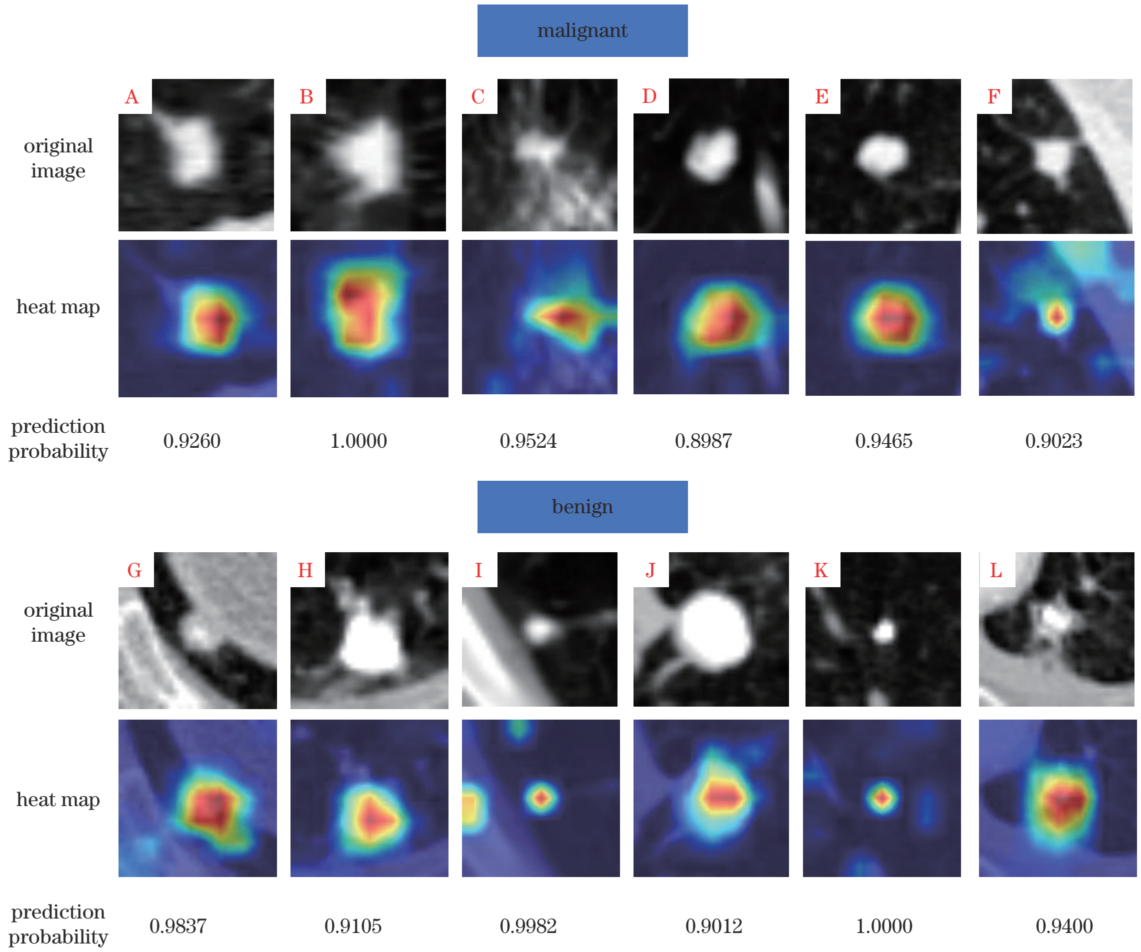

图 8 良恶性结节的 Grad-CAM 可视化和分类结果

Fig. 8 Grad-CAM visualization and classification results of benign and malignant nodules

题。另外，良性结节“H”、“I”明显比恶性结节“D”、“E”、“F”的尺寸更大，说明仅依靠形状、大小等特征来判别肺结节可能会影响分类的准确率，因此需要综合多个角度的信息来准确判断结节的性质。

5 结 论

利用 CNN 与 Transformer 的优势，设计了一个用于肺结节良恶性分类任务的双路径交叉融合网络，可以有效地捕获不同尺度的肺结节的全局形态特征以及纹理、密度等内部特征，还设计了特征融合块以及混合损失函数，进一步优化了网络性能以提高网络的预测能力。通过双路径的信息融合，网络能够综合考虑局部和全局信息，对于复杂的肺结节 CT 图像具有更好的适应性和泛化能力。在 LIDC-IDRI 数据集上的实验结果表明，所提方法在肺结节良恶性分类任务上具有一定的优势。由于目前可用的肺结节 CT 图像数据较少，在网络训练过程中易产生过拟合的问题，因此在接下来的工作中，将会探索有关图像生成的研究方法，并生成高质量图像以扩充数据集。

参 考 文 献

[1] Sung H, Ferlay J, Siegel R L, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries[J]. CA: A Cancer Journal for Clinicians, 2021, 71(3): 209-249.

[2] Wu G X, Raz D J. Lung cancer screening[M]//Reckamp K L. Lung cancer. Cancer treatment and research. Cham: Springer, 2016, 170: 1-23.

[3] McWilliams A, Tammemagi M C, Mayo J R, et al. Probability of cancer in pulmonary nodules detected on first screening CT[J]. The New England Journal of Medicine, 2013, 369(10): 910-919.

[4] Shen W, Zhou M, Yang F, et al. Multi-scale convolutional neural networks for lung nodule classification[J]. Information Processing in Medical Imaging, 2015, 24: 588-599.

[5] Xu X Y, Wang C D, Guo J X, et al. MSCS-DeepLN: evaluating lung nodule malignancy using multi-scale cost-sensitive neural networks[J]. Medical Image Analysis, 2020, 65: 101772.

[6] Xie Y T, Xia Y, Zhang J P, et al. Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest CT[J]. IEEE Transactions

on Medical Imaging, 2019, 38(4): 991-1004.

[7] Jiang H L, Gao F, Xu X X, et al. Attentive and ensemble 3D dual path networks for pulmonary nodules classification[J]. Neurocomputing, 2020, 398: 422-430.

[8] Huang Y S, Wang T C, Huang S Z, et al. An improved 3-D attention CNN with hybrid loss and feature fusion for pulmonary nodule classification[J]. Computer Methods and Programs in Biomedicine, 2023, 229: 107278.

[9] Vaswani A, Shazeer N, Parmar N, et al. Attention is all You need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, December 4-9, 2017, Long Beach, California, USA. New York: ACM Press, 2017: 6000-6010.

[10] Liu D X, Liu F H, Tie Y, et al. Res-trans networks for lung nodule classification[J]. International Journal of Computer Assisted Radiology and Surgery, 2022, 17(6): 1059-1068.

[11] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale[EB/OL]. (2020-10-22) [2023-02-03]. https://arxiv.org/abs/2010.11929.

[12] Wang R, Zhang Y S, Yang J T. TransPND: a transformer based pulmonary nodule diagnosis method on CT image[M]//Yu S Q, Zhang Z X, Yuen P C, et al. Pattern recognition and computer vision. Lecture notes in computer science. Cham: Springer, 2022, 13535: 348-360.

[13] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2022: 9992-10002.

[14] Sun R N, Pang Y X, Li W F. Efficient lung cancer image classification and segmentation algorithm based on an improved swin transformer[J]. Electronics, 2023, 12(4): 1024.

[15] Cai J, Meng Z B, Khan A S, et al. Feature-level and model-level audiovisual fusion for emotion recognition in the wild[C]//2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), March 28-30, 2019, San Jose, CA, USA. New York: IEEE Press, 2019: 443-448.

[16] Guo M H, Lu C Z, Liu Z N, et al. Visual attention network[EB/OL]. (2022-02-20) [2023-02-03]. https://arxiv.org/abs/2202.09741.

[17] Körber N. Parameter-free average attention improves convolutional neural network performance (almost) free of charge[EB/OL]. (2022-10-14) [2023-02-03]. https://arxiv.org/abs/2210.07828.

[18] Armato S G III, McLennan G, Bidaut L, et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans[J]. Medical Physics, 2011, 38(2): 915-931.

[19] Onishi Y, Teramoto A, Tsujimoto M, et al. Multiplanar analysis for pulmonary nodule classification in CT images using deep convolutional neural network and generative adversarial networks[J]. International Journal of Computer Assisted Radiology and Surgery, 2020, 15(1): 173-178.

[20] 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 (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

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

[22] Liu Z, Mao H Z, Wu C Y, et al. A ConvNet for the 2020s[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: 11966-11976.

[23] Al-Shabi M, Lan B L, Chan W Y, et al. Lung nodule classification using deep Local-Global networks[J]. International Journal of Computer Assisted Radiology and Surgery, 2019, 14(10): 1815-1819.

[24] Jiang H L, Shen F H, Gao F, et al. Learning efficient, explainable and discriminative representations for pulmonary nodules classification[J]. Pattern Recognition, 2021, 113: 107825.

[25] Al-Shabi M, Shak K, Tan M. 3D axial-attention for lung nodule classification[J]. International Journal of Computer Assisted Radiology and Surgery, 2021, 16(8): 1319-1324.

[26] Huang H, Wu R Y, Li Y, et al. Self-supervised transfer learning based on domain adaptation for benign-malignant lung nodule classification on thoracic CT[J]. IEEE Journal of Biomedical and Health Informatics, 2022, 26(8): 3860-3871.

[27] Zhu Q K, Wang Y Q, Chu X P, et al. Multi-view coupled self-attention network for pulmonary nodules classification[M]//Wang L, Gall J, Chin T J, et al. Computer vision-ACCV 2022. Lecture notes in computer science. Cham: Springer, 2023, 13846: 37-51.

[28] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 618-626.