文章编号：1673-3193(2025)05-0549-12　　J. North Univ. China, Nat. Sci. Ed., 2025,46(5):549-560.

基于膨胀卷积与图注意聚合的多模态医学图像融合

靳凯欣，王丽芳，郭　威，韩　强，郁晓庆

(中北大学 计算机科学与技术学院 生物医学成像与影像大数据山西省重点实验室，山西 太原 030051)

摘　要：针对目前基于深度学习的多模态医学图像融合方法存在高级特征提取不足和低级特征容易丢失的问题，本文提出了基于膨胀卷积与图注意聚合的多模态医学图像融合方法。该方法由双分支编码器、融合模块和解码器三部分组成。双分支编码器由基于卷积的低级编码器和基于图卷积的高级编码器构成。其中，基于卷积的低级编码器主要采用膨胀卷积来减少纹理细节等低级特征的丢失，同时为高级编码器提供初始化的节点特征。基于图卷积的高级编码器则主要采用图注意聚合模块，有效捕获深层语义等高级特征。图注意聚合模块结合边缘编码的多头注意力构建节点邻接矩阵，再基于此邻接矩阵通过图卷积对节点进行深层聚合。融合模块对提取到的特征进行融合，最后使用解码器重建融合图像。将该方法与6种先进的图像融合方法在主观视觉和客观评价指标上进行了对比。结果显示，该方法在EN上相较于IGNet方法提升了2.4%，在AG和MI上相较于DATFuse方法分别提升了3.53%和5.06%，在SD、SF和SCD上相较于SwinFusion方法分别提升了1.18%，6.24%和3%，同时该方法得到的融合图像保留了更多的纹理细节信息。综合实验结果表明，该方法实现了多模态医学图像的有效融合，为临床诊断提供了更可靠的图像支持。

关键词：多模态医学图像融合；双分支编码器；膨胀卷积；图卷积；多头注意

中图分类号：TP391　　**文献标识码**：A　　**doi:** 10.62756/jnuc.issn.1673-3193.2025.01.0006

引用格式：靳凯欣，王丽芳，郭威，等. 基于膨胀卷积与图注意聚合的多模态医学图像融合[J]. 中北大学学报(自然科学版)，2025，46(5)：549-560.

JIN Kaixin，WANG Lifang，GUO Wei，et al. Multimodal medical image fusion based on dilated convolution and graph attention aggregation[J]. Journal of North University of China(Natural Science Edition)，2025，46(5)：549-560.

Multimodal Medical Image Fusion Based on Dilated Convolution and Graph Attention Aggregation

JIN Kaixin，WANG Lifang，GUO Wei，HAN Qiang，YU Xiaoqing

(School of Computer Science and Technology，Shanxi Provincial Key Laboratory of Biomedical Imaging and Imaging Big Data，North University of China，Taiyuan 030051，China)

Abstract：Existing deep learning-based multimodal medical image fusion methods suffer from insufficient high-level feature extraction and easy loss of low-level features. To tackle these problems，this paper proposed a multimodal medical image fusion method based on dilated convolution and graph attention aggregation. The method was comprised of three components：a dual-branch encoder，a fusion module，and a decoder. The dual-branch encoder consisted of a convolution-based low-level encoder and a graph-

收稿日期：2025-01-06

基金项目：山西省“1331工程”科技创新计划(20210222)；山西省重点研发项目(202202010101008)；山西省重点研发项目(202102010101011)；山西省省筹资金资助回国留学人员科研项目(2024-118)

作者简介：靳凯欣(2000－)，女，硕士生，主要从事多模态医学图像融合、图像处理的研究。

通信作者：王丽芳(1977－)，女，教授，博士，主要从事图像融合、图像处理的研究。E-mail：727690392@qq.com。

convolution-based high-level encoder. The convolution-based low-level encoder employed dilated convolution to mitigate the loss of low-level features like texture details and provided initialized node features for the high-level encoder. The graph-convolution-based high-level encoder mainly adopted the graph attention aggregation module to effectively capture high-level features such as deep semantics. The graph attention aggregation module constructed a node adjacency matrix by integrating multi-head attention with edge encoding and then performed deep aggregation of nodes through graph convolution based on this adjacency matrix. The fusion module fused the extracted features, and the decoder reconstructed the fused image. The method was compared with six state-of-the-art image fusion methods on subjective vision and objective evaluation metrics. The results show that this method improves 2.4% on EN compared to the IGNet method, 3.53% and 5.06% on AG and MI compared to the DATFuse method, and 1.18%, 6.24%, and 3% on SD, SF, and SCD compared to the SwinFusion method, respectively, while the fused image obtained by this method retains more texture detail information. The comprehensive experimental results demonstrate that this method achieves effective fusion of multimodal medical images, offering more reliable image support for clinical diagnosis.

Key words: multimodal medical image fusion; dual-branch encoder; dilated convolution; graph convolution; multi-head attention

0 引 言

在临床医学中，多模态医学图像融合技术展现出了无可替代的重要性，突破了单一模态图像在揭示人体复杂生理结构时的信息局限性[1]。例如，计算机断层扫描(Computed Tomography，CT)或核磁共振成像(Magnetic Resonance Imaging，MRI)擅长于捕捉特定类型的信息，CT在骨骼成像上表现出高空间分辨率，但软组织对比度不足，而MRI则在软组织成像上更优，但骨骼成像分辨率较低。多模态图像融合技术整合了两者优势，可生成既包含详尽骨骼结构又具备丰富软组织细节的图像[2]，为临床诊断和治疗提供更全面的信息支持。

现有的多模态医学图像融合方法分为传统的融合方法和基于深度学习的融合方法。传统的融合方法存在融合质量不佳和计算复杂的问题，例如：基于空间域的方法[3]会带来空间畸变和光谱畸变的问题，基于变换域的方法[4]存在参数过多和参数设置复杂的问题。基于深度学习的融合方法[5-6]具有强大的特征提取能力和数据表达能力，而卷积神经网络(Convolutional Neural Network，CNN)和Transformer作为深度学习的重要分支，在特征提取中发挥着不同作用。其中，CNN通过其内部卷积核的局部感受野，能够充分提取局部重要信息[7-8]，如Shao等[9]提出的自适应频域优化的渐进式医学图像融合网络，利用CNN充分保留了医学图像的纹理细节等局部信息，但是CNN在特征提取上存在全局上下文信息容易丢失的问题。Transformer[10]通过其注意机制更好地建立长期依赖关系，解决了全局上下文信息丢失的问题，如Chen等[11]提出了一种用于高级视觉任务的HitFusion，通过采用Transformer学习了源图像之间的跨特征相关性和长距离依赖性，但其存在局部信息容易丢失的问题。为了充分提取全局上下文信息且不丢失局部信息，Li等[12]提出了混合密集连接CNN与Transformer网络，该网络融合了Transformer和CNN结构，以充分提取图像中的全局和局部信息。然而，Transformer在处理医学图像时，因其采用单一尺度的注意力机制，难以捕捉特征层次化演变，进而导致病灶深层语义特征与组织结构的多尺度上下文等高级特征提取不足。

图神经网络(Graph Neural Networks，GNN)作为深度学习的一个重要分支，在处理非结构化数据方面展现了强大的语义信息表示能力。非结构化数据以节点和边构成的图结构形式存在。GNN通过聚合节点及其邻居节点的特征信息，使其在处理具有多层次、多关系的数据时，能够更准确地捕捉特征间的关联性和演变规律，从而能提取出更丰富的高级特征[13]。例如，Xu等[14]提出的对比图池化，借助GNN捕获图的上下文与邻居节点信息，提取出丰富的脑组织结构等高级特征。

鉴于GNN的这些优势，它被用于了多模态图像融合领域[15]。其中，Li等[16]采用了图表示学习方法对图像进行特征学习，深入挖掘疾病结构等高级特征，使得融合后的图像能够更精确地反映源图像中的语义信息。然而，上述方法在特征提取上依赖堆叠的图神经网络结构，存在明显局限。简单的堆叠结构在处理图像时会引发过度平滑问题，致使颜色、边缘以及局部纹理等用于初步识别病灶的低级特征丢失。同时，对于像病理组织学特征和疾病阶段特征这类有助于深入理解疾病的高级特征，也难以做到充分提取与利用。鉴于低级和高级特征对融合结果的准确性和信息完整性都十分关键，有效融合并利用这些特征成为亟待解决的问题。

针对上述多模态医学图像融合方法中存在着高级特征提取不足和低级特征容易丢失的问题，本文提出一种基于膨胀卷积和图注意聚合的多模态医学图像融合方法。该方法首先采用基于CNN的低级编码器和基于GCN的高级编码器分别进行低级和高级特征提取。其中，在基于CNN的低级编码器中，通过使用基于膨胀卷积的多尺度模块高效地提取图像中不同尺度和感受野的低级特征，同时为高级编码器提供更全面的节点特征表示。在基于GCN的高级编码器中，通过使用图注意聚合模块基于结合边缘编码的多头注意力为节点构建了具有全局信息的邻接矩阵，再使用图卷积神经网络聚合深层节点特征，以充分捕获高级特征。然后使用分级Softmax多模态融合网络进行特征融合，最后通过解码器对图像进行重建。对比实验表明，本文方法在主观视觉评价和客观评价指标上均优于最先进的6种图像融合方法。

1　相关工作

1.1　膨胀卷积

膨胀卷积是一种创新的卷积技术，其与普通卷积的核心区别在于：普通卷积以连续无间隔的方式对输入采样，卷积核紧密作用于相邻像素，感受野随网络层数呈线性扩展；而膨胀卷积通过引入膨胀率（Dilation Rate），在卷积核元素间增加“空洞”实现间隔采样[17]，能够在不改变参数量、不降低特征分辨率的前提下，指数级扩大感受野。以3*3卷积核为例，当Dilation=1（即未引入膨胀率）时，卷积核以连续无间隔的方式对输入特征图进行采样；当Dilation=2或Dilation=3时，3*3膨胀卷积的实际感受野可拓展至5*5或7*7，如图1所示。凭借这一特性，膨胀卷积无需增加卷积核物理尺寸或显著提升计算负担[18]，即可同时实现全局信息捕获与细节特征保留。

相较于普通卷积操作，膨胀卷积确保了较大的感受野，这对于解析复杂图像结构、捕捉长距离依赖关系至关重要，特别是在图像检测[19]、分割[20]等高级视觉任务中，由于膨胀卷积能够覆盖更广泛的区域，它能够在不丢失细节的情况下，更好地保留图像中的关键特征，从而减少了特征丢失。

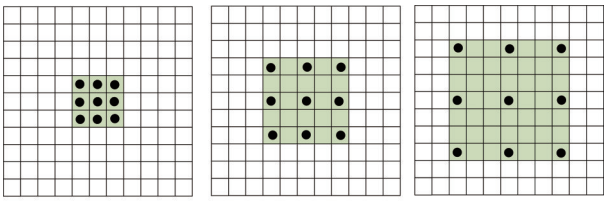

图1　具有不同膨胀系数的膨胀卷积

Fig. 1　Dilated convolutions with varying dilation rates

1.2　图卷积神经网络（GCN）

图神经网络（GNN）不同于CNN，CNN是对欧几里得空间中的数据集进行处理，它将图像视为固定维度的矩阵；而GNN用于非欧几里得空间的数据分析[21]，早期应用于知识图谱[22]和蛋白质[23]等非结构化数据处理。通过图像分割[24]将图像转换为图结构，GNN也可以有效处理结构化数据，因此在图像分类[25-26]、目标检测[27]等领域得到了广泛应用。

图卷积神经网络（Graph Convolutional Neural Network，GCN）属于GNN的一类，是在图结构中引入卷积操作，通过迭代聚合邻居节点的特征向量，更新节点的隐藏状态，从而学习节点的特征[28]。在GCN的第l层中，输出的隐藏表示H^l，如式（1）所示。

$$H^l=\sigma(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{l-1}W^l), \tag{1}$$

式中：$\hat{A}=A+I_N$为添加自连接的邻接矩阵，A为邻接矩阵，I_N为单位矩阵；$\hat{D}_{ij}=\sum_j\hat{A}_{ij}$，为度矩阵；$W$为特定层可训练的权重矩阵；$\sigma(\cdot)$为激活函数；$H^{l-1}$为第$l-1$层的输出。

2　基于膨胀卷积与图注意聚合的多模态医学图像融合

本文提出的基于膨胀卷积和图注意聚合的多

模态医学图像融合方法由双分支编码器、融合模块和解码器三部分组成，如图2所示。

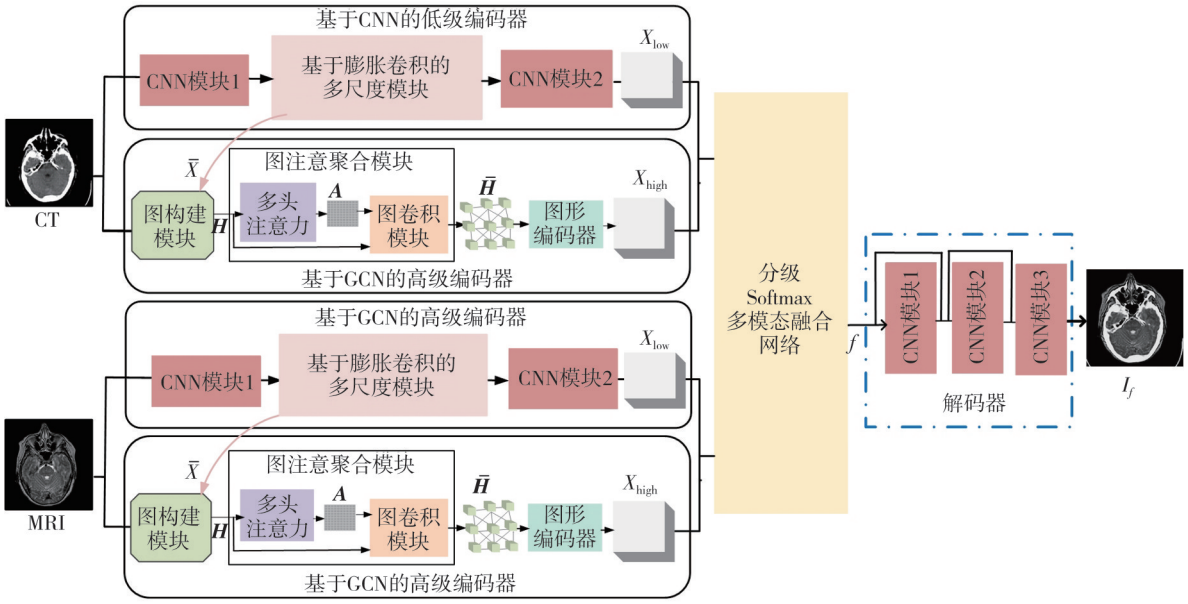

图2 基于膨胀卷积和图注意聚合的多模态医学图像融合

Fig. 2 Multimodal medical image fusion based on dilated convolution and graph convolution

在双分支编码器中使用基于CNN的低级编码器和基于GCN的高级编码器来分别提取低级和高级特征。在基于CNN的低级编码器中，使用基于膨胀卷积的多尺度模块提取不同尺度和感受野的纹理细节等低级特征X_{low}。在基于GCN的高级编码器中，图构建模块结合多尺度模块的特征图$\bar{X}$生成节点特征矩阵$\boldsymbol{H}$，然后图注意聚合模块通过结合边缘编码的多头注意力构建邻接矩阵$\boldsymbol{A}$，并利用图卷积聚合节点的深层结构信息，得到高级特征$\bar{H}$，最后图形编码器将图结构信息转化为特征图X_{high}。融合模块采用分级Softmax多模态融合网络整合特征f，最后由解码器重建融合图像I_f。

2.1 基于CNN的低级编码器

基于CNN的低级编码器由两层CNN模块和基于膨胀卷积的多尺度模块构成，如图2所示。

首先，CNN模块1通过两个卷积块提取浅层特征$X\in\mathbf{R}^{H*W*N/(2*2)}$。其中，卷积块1由1*1卷积和Relu函数构成，卷积块2由stride=2、kenerl size=3*3、padding=1的卷积、BatchNorm和Relu函数构成。

在多尺度模块中使用膨胀卷积从浅层特征X中捕获更为丰富且多尺度的信息，如图3所示。该模块采用3个不同膨胀率(1、3、5)的平行路径，每个路径包含两组膨胀卷积块，每块由BatchNorm、3*3膨胀卷积和Relu函数构成。此外，两层卷积间添加了kenerl size=3*3，stride=1，padding=1的平均池化层，以减少冗余信息并保持图像尺寸不变。经过多尺度模块后得到不同尺度和感受野的低级特征。最后，这些特征通过CNN模块2整合。其中，CNN模块2由两个卷积块构成，每个卷积块均由3*3卷积、BatchNorm和Relu函数构成，最终得到低级特征$X_{low}\in\mathbf{R}^{H*W*N/(2*2)}$。

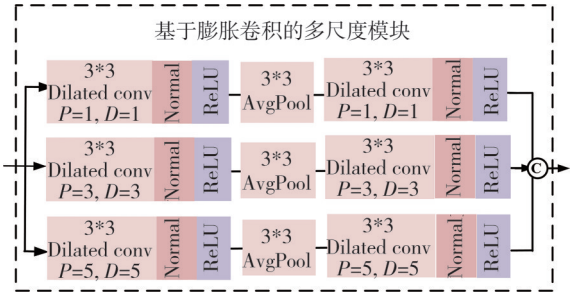

图3 基于膨胀卷积的多尺度模块

Fig. 3 Multi-scale module based on dilated convolution

2.2 基于GCN的高级编码器

本文采用基于GCN的高级编码器充分学习图结构信息，以捕获更全面的高级特征。整个分支分为三部分：图构建模块、图注意聚合模块和图形编码器，如图2中的基于GCN的高级编码器所示。

2.2.1 图构建模块

图构建模块是将图像构建为图结构并生成节

点特征矩阵，整个过程划分为图像分割和节点初始化两个阶段，如图 4 所示。

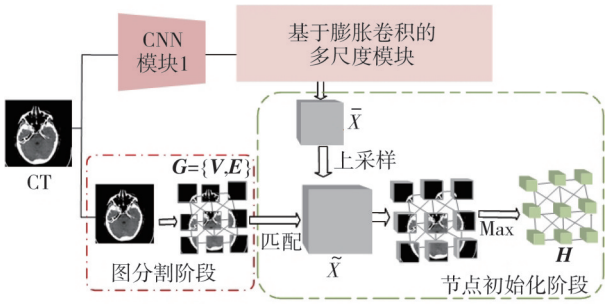

图 4　图构建模块

Fig. 4　Graph construction module

在图像分割阶段，首先将图像 $I_{ct}\in\mathbf{R}^{H*W*C}$ 平均分割为多个图像块，以减少计算复杂度，将分割后的图像块视作图结构中的节点，节点表示为 $V=\{v_1,v_2,\cdots,v_n\}$；对于每个节点 v_i，根据K-近邻算法来计算图像块之间的欧氏距离，选出 k 个最近邻居节点 $N(V_i)$，为节点 v_i 与其邻居节点 $N(V_i)$ 之间添加一条边 E_{ij}，由此得到图结构表示 $G=\{V,E\}$。

在节点初始化阶段，本文利用膨胀卷积提取多尺度特征的能力，将提取的特征作为节点初始特征 $H=\{h_1,h_2,\cdots,h_n\}$。通过这种方式，节点能够捕捉丰富的多尺度信息，提升节点特征的表达能力，同时为图注意聚合提供高质量的输入。如图4所示，将多尺度模块的特征图 $\bar{X}$ 上采样至图像大小得到 $\tilde{X}$，将 $\tilde{X}$ 与图像块（即节点）进行匹配，并在 $\tilde{X}$ 中选取出最大像素值作为相应的节点特征，确保每个节点捕获对应区域最显著的特征，最终得到节点特征矩阵 $H\in\mathbf{R}^{n*d}$，如式（2）所示。

$$H_{ij}=\max_{(p,q)\in P_k}\tilde{X}_{j,(p,q)},\tag{2}$$

式中：$H_{ij}\in\mathbf{R}^{n*d}$ 为第 i 个节点的第 j 个特征；P_k 为第 k 个图像块；(p,q) 为 P_k 中每个像素的位置；$\tilde{X}_{j,(p,q)}$ 为位置 (p,q) 第 j 个通道中特征图的值。

2.2.2　图注意聚合模块

图注意聚合模块主要由多头注意力和图卷积块组成，如图 2 所示。

多头注意力通过结合边缘编码共同构建邻接矩阵，捕捉节点的全局结构信息和连接特性，优化图卷积对节点深层特征的聚合。其中，多头注意力通过并行多个注意力头，计算图中节点全局结构信息。边缘编码通过高斯函数对节点之间的边进行权重赋值，以此更好地反映节点之间的相似性和关联性，具体的计算式如（3）所示。

$$E_{ij}=\begin{cases}\phi(\|h_i-h_j\|), & \text{if } h_j\in N(h_i),\\ 0, & \text{otherwise},\end{cases}\tag{3}$$

式中：E_{ij} 为节点 i 和 j 之间的边缘编码；$\phi(x)=e^{-rx^2}$ 为高斯函数；$\|h_i-h_j\|^2$ 为节点之间的欧几里得距离；r 是一个经验设定的值，赋值为 0.2；$N(h_i)$ 为节点 h_i 的邻接节点集合。

多头注意力的具体结构如图 5 所示，首先将节点特征 $H\in\mathbf{R}^{n*d}$ 送入到具有 M 个头的多头注意力中进行分头计算，然后由线性变换得到向量序列 $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$，再计算每个注意力头中 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 值相似度矩阵 $\bar{\boldsymbol{S}}$，同时结合边缘编码 E，由式（4）～式（5）表示。

$$\bar{\boldsymbol{S}}_{ij}=\frac{(h_iW_Q)(h_jW_k)^{\mathrm{T}}}{\sqrt{d}},\tag{4}$$

$$\boldsymbol{S}_{ij}=\bar{\boldsymbol{S}}_{ij}+E_{ij},\tag{5}$$

式中：$\boldsymbol{W}_Q$，$\boldsymbol{W}_K$ 分别为 $\boldsymbol{Q}$，$\boldsymbol{K}$ 的投影权值；d 为节点特征的维度。计算 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 的相似度矩阵 $\boldsymbol{S}$ 后，再经过 softmax 层后与 V 值进行注意力关系的计算，依次得到 M 个头的关系矩阵 $G=\{\boldsymbol{G}_1,\boldsymbol{G}_2\cdots,\boldsymbol{G}_M\}$，然后使用一组可学习的权重向量 $\boldsymbol{W}=[\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_M]\in\mathbf{R}^{1*M}$，学习最终的邻接矩阵 $\boldsymbol{A}\in\mathbf{R}^{n*n}$，由式（6）～式（7）表示。

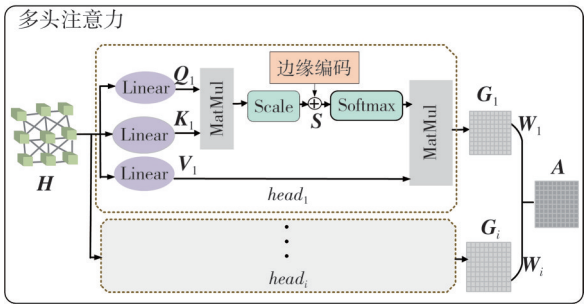

图 5　多头注意力

Fig. 5　Multi-head attention

$$\boldsymbol{G}=Attn(\boldsymbol{H})=softmax(\boldsymbol{S})\boldsymbol{V},\tag{6}$$

$$\boldsymbol{A}=\sum_{i=1}^{M}w_i\boldsymbol{G}_i,\tag{7}$$

图卷积块主要结合了图卷积网络（GCN）和前馈神经网络（FFN）的优势，实现了对图中节点特征的深度聚合，不仅显著提升了节点特征的表达能力，还增强了模型对图结构的处理能力。其中，图卷积块由三层包含GCN和FFN的GCF块堆叠而成，如图 6 所示。GCN块由图卷积和Relu函数构成，用于聚合节点特征；FFN由两个卷积块组成，卷积块中包含BatchNorm、1*1卷积和Relu函数，避免了堆叠GCN出现过度平滑现

象[29]。图卷积块的公式由式(8)~式(10)表示。

$$GCF(H^{l-1},A)=H^{l}=FFN(\hat{H}^{l})+\hat{H}^{l},\quad(8)$$

$$FFN(\hat{H}^{l-1})=Conv(Conv(\hat{H}^{l})),\quad(9)$$

$$\hat{H}^{l}=GCN(H^{l-1})+H^{l-1},$$

$$GCN(H^{l-1})=\sigma(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{l-1}\boldsymbol{W}^{l}),\quad(10)$$

式中：H^{l-1}为$l-1$层GCF模块的输出；$\hat{D}$为度矩阵；$\hat{A}=A+I_N$为添加自连接的邻接矩阵，A为邻接矩阵，I_N为单位矩阵；$\boldsymbol{W}$为可学习的权重；σ为激活函数；$Conv$为卷积块。

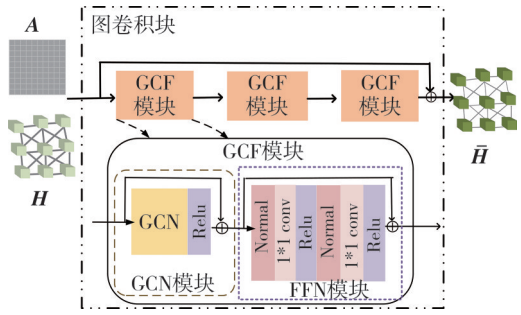

图6 图卷积块

Fig. 6 Graph convolution block

2.2.3 图形编码器

在图形编码器中，通过构建图像块与像素之间的映射矩阵$\boldsymbol{Q}$[30]，可以有效地将图结构信息转换为特征图$\boldsymbol{X}_{high}\in\mathbf{R}^{H*W*N}$，由式(11)~式(12)所示。

$$Q_{i,j}=\begin{cases}1, & \text{if } \hat{X}_i\in P_j,\\ 0, & \text{otherwise},\end{cases}$$

$$\hat{\boldsymbol{X}}=Flatten(I),\quad(11)$$

$$\boldsymbol{X}_{\text{high}}=\text{Re}\,shape(\boldsymbol{Q}\bar{H}),\quad(12)$$

式中：P_j为第j个图像块；$I\in\mathbf{R}^{H*W*C}$为原始图像；$Flatten(\cdot)$为将空间维度平坦化；$\hat{\boldsymbol{X}}_i\in\mathbf{R}^{HW*C}$为平坦后的原始像素中第$i$个像素；$Q_{i,j}$为第$i$个像素和第$j$个图像块之间的关系；$\text{Re}\,shape(\cdot)$为恢复数据的空间维度。

2.3 分级Softmax多模态融合网络

本文采用分级Softmax多模态融合网络进行特征融合，其结构如图7所示。首先融合不同模态的同级别特征，确保同级别信息的有效整合；再融合不同级别的特征，实现跨级别信息的交互与利用。

在分级Softmax多模态融合网络中，首先分别将不同模态的低级特征$\boldsymbol{X}_{\text{low}}\in\mathbf{R}^{H*W*N/(2*2)}$上采样至高级特征$\boldsymbol{X}_{\text{high}}\in\mathbf{R}^{H*W*N}$相同尺寸，然后对不同模态的高低级特征分别进行串联，以得到不同级别的特征f_1和f_2。然后使用Softmax加权对f_1和f_2进行

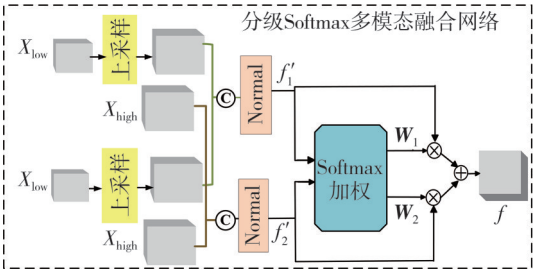

图7 分级Softmax多模态融合网络

Fig. 7 Hierarchical softmax multimodal fusion network

特征融合，将f_1和f_2进行归一化处理，以减少后续计算的复杂度得到f_1'和f_2'。然后分别对f_1'和f_2'中各个通道信息进行求和，将多通道特征压缩为单通道特征，来捕获不同级别的整体信息，由式(13)所示。

$$S_1=\sum\nolimits_j f'_{1,j},\quad S_2=\sum\nolimits_j f'_{2,j},\quad(13)$$

式中：S_1和S_2为不同级别的单通道特征；$f'_{1,j}$和$f'_{2,j}$为不同级别的第j个通道的特征，对得到的S_1和S_2使用Softmax函数，以获得不同级别特征的权重$\boldsymbol{W}_1$和$\boldsymbol{W}_2$，由式(14)所示。

$$\boldsymbol{W}_i=\frac{\exp(S_i)}{\sum_i\exp(S_i)},\quad(14)$$

式中：$\boldsymbol{W}_i, i\in\{1,2\}$，为不同级别的权重值；exp表示指数运算。最终，融合不同级别权重值与相应特征得到融合特征f，由式(15)表示。

$$f=\boldsymbol{W}_1 f'_1+\boldsymbol{W}_2 f'_2。\quad(15)$$

2.4 解码器

解码器由3个CNN模块构成，将融合特征f重构得到图像I_f，如图2中的解码器所示。每个CNN模块均由两层的3*3的卷积、BatchNorm和Rule函数构成，用于提取和恢复特征，从而实现对融合特征的重建，同时为了避免梯度消失，在CNN模块1和2与CNN模块2和3之间采用了残差连接的方式，最后得到重建融合图像I_f。

2.5 无监督训练

本文方法采用的是无监督训练，旨在通过对单一图像重建的方式，对双分支编码器与解码器进行训练。在训练过程中使用由重建损失L_{recon}和梯度损失L_{gradient}构成的总损失函数L_{total}进行无监督训练。其中，L_{recon}避免了源图像在编码和解码过程中信息的丢失，L_{gradient}保留了源图像更多的纹理细节信息，由式(16)~式(18)表示。

$$L_{\text{total}}=L_{\text{recon}}+\beta L_{\text{gradient}},\quad(16)$$

$$L_{\text{recon}}=\| I_{ct}-I'_{ct}\|_2^2, \tag{17}$$

$$L_{\text{gradient}}=\frac{1}{HW}\| \nabla I_{ct}-\nabla I'_{ct}\|_1, \tag{18}$$

式中：β为调优参数；$\| \ \|_2$为L_2范数；∇I_{ct}为源图像的梯度信息；$\nabla I_{ct}'$为重构图像的梯度信息；$\| \ \|_1$为L_1范数。

3　实　验

3.1　数据集及参数设置

本文的数据集来自美国哈佛医学院官方网站所开源的正常脑图像和脑肿瘤疾病图像，从中挑选出脑部纹理清晰、细节特征丰富的数据集，其中包含CT、MRI等医学图像共10 000张作为训练集，已配准的成对CT/MRI图像共20对作为测试集。所有图像大小统一调整为256×256。

参数设置：GPU为GeFore RTX3090搭载24 GB；环境框架为Pytorch；训练时使用Adam优化器，batch-size设置为16，epoch设置为100，初始学习率为1×10^{-4}，学习率每20个epoch衰减0.5，损失函数L_{total}中β设置为1。

3.2　评价指标

为了客观评价多模态医学图像融合方法的性能，本文选取了7个常用指标：熵、标准差、平均梯度、互信息、边缘信息保留、空间频率和差异相关和。

熵（Entropy，EN）：衡量融合图像的信息丰富程度，值越高表示纹理细节和多样性越多，计算公式为

$$EN=-\sum_{i=0}^{L-1}p(i)\log_2 p(i), \tag{19}$$

式中：L为融合图像灰度级数；$p(i)$为灰度值i出现的概率。

标准差（Standard Deviation，SD）：评估图像信息丰富度，值越大表示图像灰度分布更分散，信息量更多，融合图像质量越好，计算公式为

$$SD=\sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(F(i,j)-\bar{F})^2}, \tag{20}$$

式中：M和N分别为图像的高和宽；F为融合图像；$\bar{F}$为图像平均灰度。

平均梯度（Arerage Gradient，AG）：描述图像的边缘特征，值越高表示边缘特征越明显和丰富，计算公式为

$$AG=\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\sqrt{\frac{\nabla F_x^2(i,j)+\nabla F_y^2(i,j)}{2}}, \tag{21}$$

式中：∇I_x^2和∇I_y^2分别为融合图像F在x，y轴熵的梯度。

互信息（Mutual Information，MI）：度量源图像与融合图像之间的信息特征相似性，值越高表示信息特征越丰富，计算公式为

$$MI(A,B)=H(A)+H(B)-H(A,B), \tag{22}$$

式中：$H()$为计算图像的熵；A、B分别为源图像。

边缘信息传递因子（Edge Information Transmission Factor，$Q^{AB/F}$）：测量源图像与融合图像的边缘信息量，值越大表示边缘信息保存得越好，计算公式为

$$Q^{AB/F}=\frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(Q^{AF}(i,j)\boldsymbol{\omega}^A(i,j)+Q^{BF}(i,j)\boldsymbol{\omega}^B(i,j))}{\sum_{i=1}^{N}\sum_{j=1}^{M}(\boldsymbol{\omega}^A(i,j)+\boldsymbol{\omega}^B(i,j))}, \tag{23}$$

式中：$Q^{AF}(i,j)$和$Q^{BF}(i,j)$为不同源图像A、B与融合图像之间的边缘信息保存值；$\boldsymbol{\omega}^A(i,j)$和$\boldsymbol{\omega}^B(i,j)$为权重。

空间频率（Spatial Frequency，SF）：评估融合图像的灰度变化率，值越大表示图像越清晰，计算公式为

$$\begin{aligned}SF&=\sqrt{RF^2+CF^2},\\ CF&=\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(F(i,j)-F(i-1,j))^2},\\ RF&=\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(F(i,j)-F(i,j-1))^2},\end{aligned} \tag{24}$$

式中：CF和RF分别为列频率和行频率。

差异相关性总和（Sum of Correlations Differences，SCD）：测量融合图像与源图像之间的差异，值越大表示相关性越高，细节和结构特征保持得越好，计算公式为

$$SCD=\frac{\sum_{i,j}A(i,j)D_{A,F}(i,j)}{\sqrt{\sum_{i,j}A(i,j)^2\sum_{i,j}D_{A,F}(i,j)^2}}+\frac{\sum_{i,j}B(i,j)D_{B,F}(i,j)}{\sqrt{\sum_{i,j}B(i,j)^2\sum_{i,j}D_{B,F}(i,j)^2}}, \tag{25}$$

式中：$D_{A,F}$和$D_{B,F}$分别表示融合图像F与源图像A，B之间的差异；(i,j)表示图像中的像素位置。

3.3 对比实验

本文选取了6种基于深度学习的图像融合方法，包括TUFusion[31]、Swinfusion[32]、IGNet[33]、U2Fusion[7]、ITFuse[34]，和DATFuse[35]，将其与本文的方法进行定性和定量两个方面的比较，以全面客观地评价本文方法在多模态医学图像融合上的性能。

3.3.1 定性比较

为评估本文方法在多模态医学图像融合中的效果，选取脑肿瘤、脑中风、脑出血及脑梗四类典型病例的CT/MRI融合图像开展整体视觉定性比较。具体通过目视检查融合图像与原始模态图像的色彩合理性、边缘清晰度、组织对比度和尺寸的一致性以及模态互补性表现[36]。

本文方法与6种对比方法得到的融合结果如图8所示。各方法具体表现如下：TUFusion在脑肿瘤、脑中风等场景中虽保留了部分纹理细节与边缘特征，但脑肿瘤病灶局部、脑中风影像特定位置模糊，病灶区域清晰度不足，影响关键部位辨识；SwinFusion的融合图像亮度和对比度与源图像相近，但在脑肿瘤和脑梗死影像中丢失部分精细细节，边缘部分也有待完善；IGNet处理医学图像中的组织信息时精细结构不突出且整体亮度偏暗，如脑梗死融合图像因亮度不足，难以观察脑组织层次与病变细节；U2Fusion在脑肿瘤、脑出血等场景中保留了丰富的纹理细节，但脑部轮廓描绘有缺陷；ITFuse得到的融合结果整体颜色偏暗，在脑肿瘤等场景中纹理细节显著模糊；DATFuse在融合MRI图像时，对其细节突显不足，如脑梗死图像无法充分突出组织细节而影响病灶细微特征的捕捉。在脑肿瘤、脑中风等场景中，本文方法能有效保留CT与MRI的纹理细节、边缘等低级特征，同时清晰表征了精细结构与脑组织的空间关系，以及病灶位置与大小等高级特征。综上所述，本文方法在视觉效果上优于6种对比方法，可为医生诊断提供更优质的影像依据。

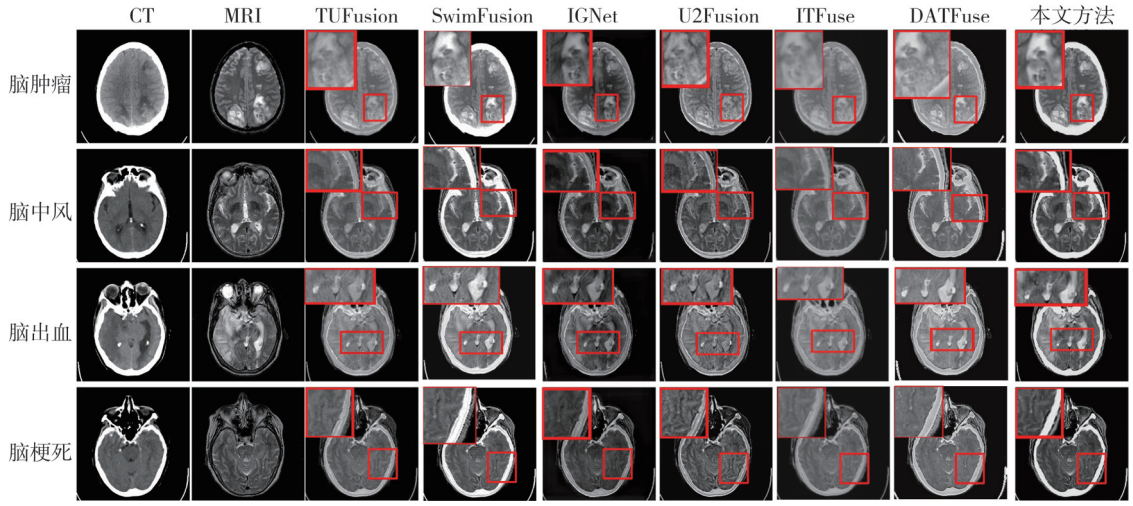

图8 本文方法和6种对比方法在CT/MRI数据集上的定性比较

Fig. 8 Qualitative comparison of the proposed method and six image fusion methods on the CT/MRI dataset

3.3.2 定量比较

将本文方法及6种对比方法在7种常用指标上进行定量比较，实验的测试结果如表1所示，其中，黑色加粗字体表示该指标的最优值。本文方法在7种客观评价指标上均表现优异，特别是在EN和SD指标上展示了卓越的图像内容和语义信息保留能力，同时，在QABF和AG指标上显示出对图像细节和边缘对比度的良好捕捉。SF、SCD和MI指标进一步凸显了本文方法在综合保留高级和低级视觉特征上的全面优势。同时，相较于表1中的次优值，本文方法在EN、SD、AG、SF、SCD和MI上分别提升了2.4%，1.18%，3.53%，6.24%，3%和5.06%。其中，提升比例=(最优值−次优值)/次优值×100%。这不仅体现了本文方法的实际应用潜力，也为其在图像处理领域的广泛应用奠定了坚实基础。

此外，图9的折线图进一步展示了本文方法与对比方法在测试集上的趋势对比。从图9中可以看出，本文方法在指标上均呈现出明显的优势，验证了其在图像融合和特征保留方面的优越性。

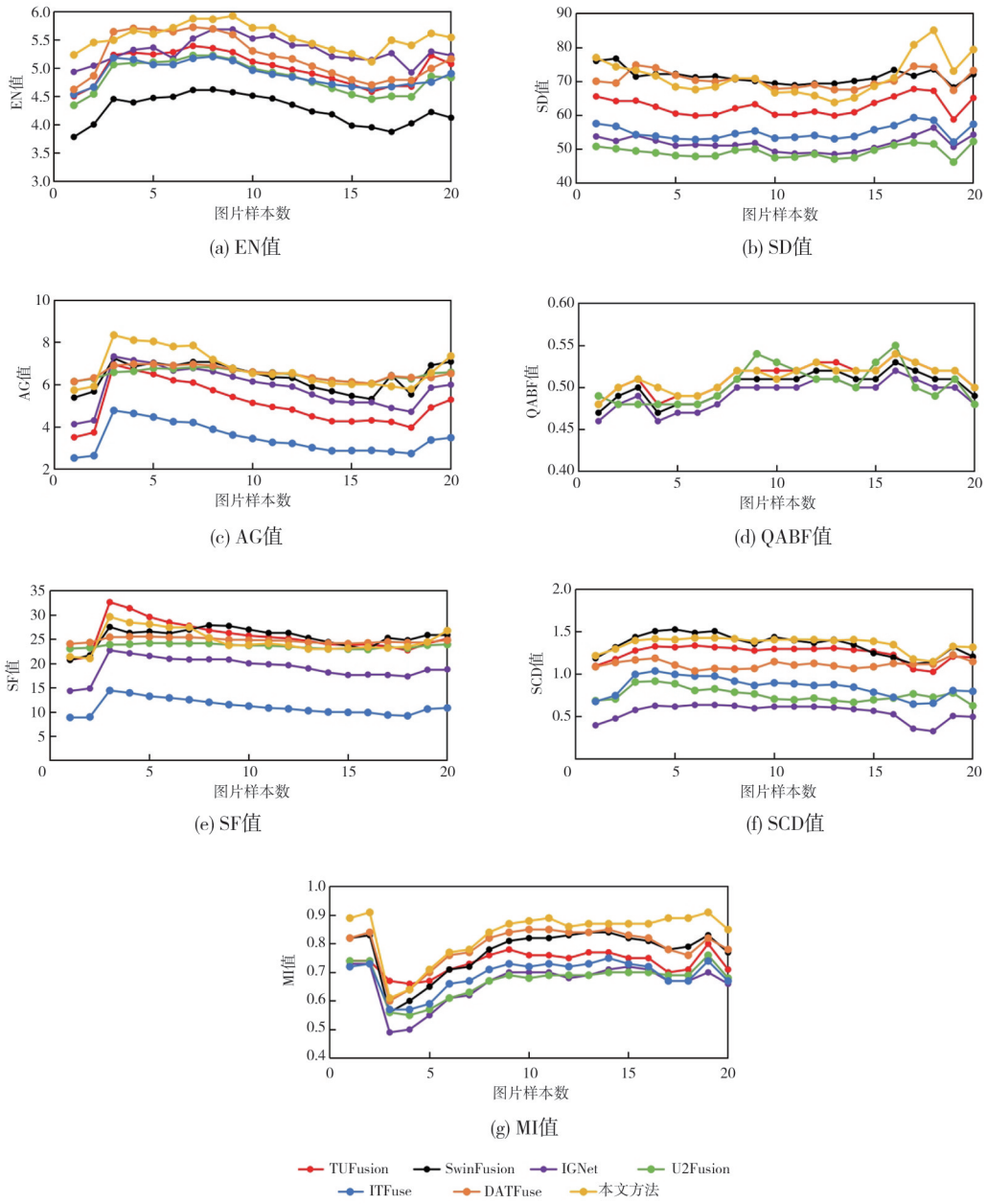

图 9　不同融合方法的评价指标-图片样本数折线图

Fig. 9　Line charts of evaluation index vs. image sample number for different fusion methods

表 1　对比实验的客观评价指标平均值

Tab. 1　Average values of objective evaluation metrics for comparative experiment

融合方法	EN	SD	AG	QABF	SF	SCD	MI
TUFusion	5.01	62.96	5.01	0.51	19.33	1.20	0.75
SwinFusion	4.27	71.97	6.38	0.50	25.32	1.34	0.77
IGNet	5.43	50.98	5.83	0.49	19.80	0.51	0.66
U2Fusion	4.89	55.38	6.47	0.49	23.29	0.76	0.67
ITFuse	4.91	55.06	3.45	0.26	11.09	0.85	0.70
DATFuse	5.21	70.56	6.51	0.30	24.56	1.25	0.79
本文方法	**5.56**	**72.82**	**6.74**	**0.51**	**26.90**	**1.38**	**0.83**

3.4　消融实验

消融实验旨在全面评估本文所提出的方法中各个关键组件以及损失函数权重对融合效果的具体贡献，从而验证该方法在多模态医学图像融合任务上的有效性。为了清晰评估各组件和损失函数权重的影响，本文设置了5组消融实验：

实验一：将低级 CNN 编码器的多尺度模块中的膨胀卷积替换为普通卷积，其余不变，探究膨胀卷积对低级特征提取的影响。

实验二：去除高级 GCN 编码器的多头注意力模块，其余不变，探究其对高级特征提取充分性的影响。

实验三：去除多头注意力中的边缘编码，其余不变，验证其对高级特征丰富性及融合图像质

量的影响。

实验四：将损失函数权重设为0，使训练时损失函数不影响参数更新，以明确梯度损失在融合过程的作用。

实验五：将损失函数权重设为2，相比默认权重增大，促使模型更关注梯度损失优化。

本文对上述消融实验结果开展了定性与定量对比分析，其中定性结果如图10所示。

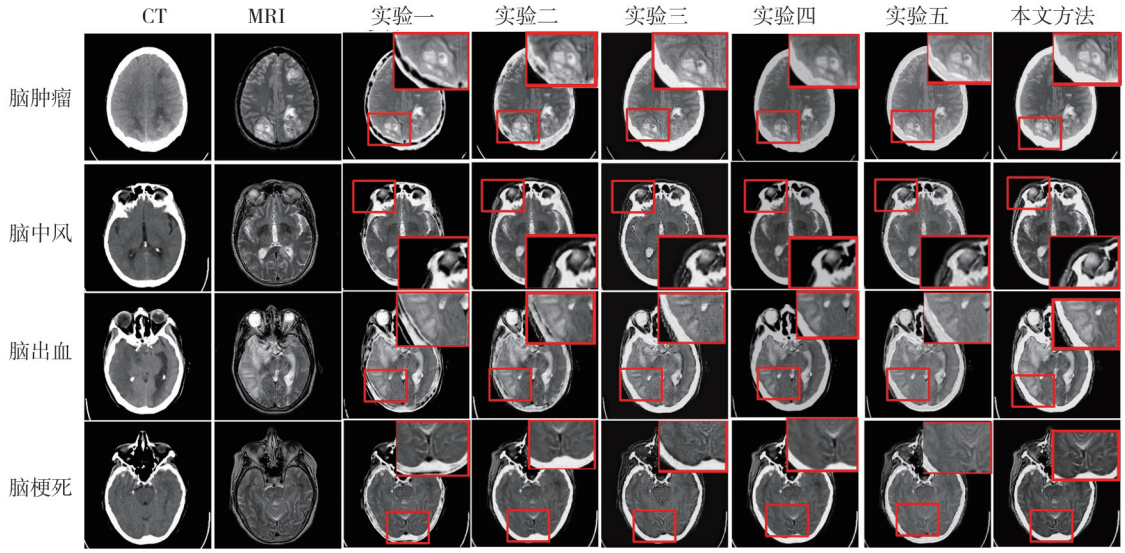

图10 消融实验的定性比较结果

Fig. 10 Qualitative comparison results of ablation experiment

由图10可以看出，相比其他方法，本文方法得到的融合图像更好地保留了骨骼结构和软组织细节，有效地融合了CT/MRI图像的低级特征，同时对脑肿瘤等病变区域的关键语义信息表征更为清晰，且具有更丰富的高级特征。定量结果见表2，黑色加粗字体表示该指标的最优值，带有下划线的为次优值。相较于表2中的每个指标的次优值，本文方法在EN值、SD值、QABF值、SF值、SCD值和MI值上分别提升了3.93%，2.32%，4.08%，3.98%，0.72%和5.06%。综合上述定量和定性比较，本文方法的融合效果优于其他方法。

表2 融合实验的客观评价指标平均值

Tab. 2 Average values of objective evaluation metrics for fusion experiment

消融实验	EN	SD	AG	QABF	SF	SCD	MI
实验一	4.89	69.56	6.74	0.45	25.87	0.88	0.74
实验二	4.93	68.83	6.11	0.44	25.01	1.21	0.73
实验三	5.35	69.50	5.48	0.49	21.92	1.37	0.72
实验四	5.25	67.50	5.94	0.42	21.27	1.27	0.79
实验五	5.19	71.17	5.28	0.46	21.29	1.29	0.72
本文方法	**5.56**	**72.82**	**6.74**	**0.51**	**26.90**	**1.38**	**0.83**

4 结 论

本文所提出的基于膨胀卷积和图注意聚合的多模态医学图像融合方法，通过采用基于CNN的低级编码器和基于GCN的高级编码器分别充分提取了低级和高级特征，有效解决了多模态医学图像融合中存在高级特征提取不足和低级特征容易丢失的问题。从实验结果上可知，本文方法得到的融合图像保留了更多的低级特征，对纹理细节表征清晰，保留了丰富的边缘特征，同时保留了更为完整的高级特征，清晰地描述了图像中病变区域的组织结构，更有助于医生的诊断，其在主观视觉评价和客观指标评价方面都有较好的表现。

本方法采用的双分支编码器在可解释性方面还存在一定局限，同时，模型的复杂结构使得其决策过程和特征表示难以直观理解，这在一定程度上限制了其在临床实践中的广泛应用。后续可将本文方法与现有大模型相结合，借助大模型强大的语义理解和知识表示能力来改善双分支编码器的可解释性，进一步提升方法的性能和实用性，从而更好地服务于医学诊断。

参考文献：

[1] DIWAKAR M, SINGH P, RAVI V, et al. A non-conventional review on multi-modality-based medical image fusion[J]. Diagnostics, 2023, 13(5): 820.

[2] HUANG B, YANG F, YIN M, et al. A review of multimodal medical image fusion techniques [J]. Com-

putational and Mathematical Methods in Medicine, 2020, 2020: 8279342.
[3] DU J, LI W. Two - scale image decomposition based image fusion using structure tensor [J]. International Journal of Imaging Systems and Technology, 2020, 30(2): 271-284.
[4] XIA J M, CHEN Y M, CHEN A Y, et al. Medical image fusion based on sparse representation and PCNN in NSCT domain [J]. Computational and Mathematical Methods in Medicine, 2018(1): 2806047.
[5] WEI X, QIU Y, XU X, et al. ECINFusion: A novel explicit channel-wise interaction network for unified multi-modal medical image fusion[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35(5): 4011-4025.
[6] CHEN J, DING J, YU Y, et al. THFuse: An infrared and visible image fusion network using transformer and hybrid feature extractor [J]. Neurocomputing, 2023, 527: 71-82.
[7] XU H, MA J, JIANG J, et al. U2Fusion: A unified unsupervised image fusion network [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(1): 502-518.
[8] LI H, XU T, WU X J, et al. LRRNet: A novel representation learning guided fusion network for infrared and visible images [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(9): 11040-11052.
[9] SHAO D, YANG H, MA L, et al. AFPNet: An adaptive frequency-domain optimized progressive medical image fusion network[J]. Biomedical Signal Processing and Control, 2025, 103: 107357.
[10] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows [DB/OL]. (2021-08-17) [2025-01-06]. https://arxiv.org/abs/2103.14030.
[11] CHEN J, DING J, MA J. HitFusion: Infrared and visible image fusion for high-level vision tasks using transformer [J]. IEEE Transactions on Multimedia, 2024, 26: 10145-10159.
[12] LI X, HE H, SHI J. HDCCT: Hybrid densely connected CNN and transformer for infrared and visible image fusion[J]. Electronics, 2024, 13(17): 3470.
[13] LIN Z, SUN W, TANG B, et al. Semantic segmentation network with multi-path structure, attention reweighting and multi-scale encoding[J]. The Visual Computer, 2023, 39(2): 597-608.
[14] XU J, BIAN Q, LI X, et al. Contrastive graph pooling for explainable classification of brain networks[J]. IEEE Transactions on Medical Imaging, 2024, 43(9): 3292-3305.
[15] LI J, CHEN J, LIU J, et al. Learning a graph neural network with cross modality interaction for image fusion [DB/OL]. (2023-08-07) [2025-01-06]. https://arxiv.org/abs/2308.03256.
[16] LI J, BAI L, YANG B, et al. Graph representation learning for infrared and visible image fusion [DB/OL]. (2023-11-01) [2025-01-06]. https://arxiv.org/abs/2311.00291.
[17] MA J, LI X, ZHANG Y, et al. U-Convnext network for infrared small target detection[C]//IEEE International Conference on Image Processing(ICIP), 2024: 1371-1376.
[18] ZHOU M, XU X, ZHANG Y. An attention-based multi-scale feature learning network for multimodal medical image fusion [DB/OL]. (2022-12-09) [2025-01-06]. https://arxiv.org/abs/2212.04661.
[19] 曲海成, 李瑞柯, 王蒙, 等. 基于特征重用和膨胀卷积的遥感图像舰船检测[J]. 智能系统学报, 2024, 19(5): 1298-1308.
QU Haicheng, LI Ruike, WANG Meng, et al. Ship detection in remote sensing images via feature reuse and dilated convolution [J]. CAAI Transaction on Intelligent Systems, 2024, 19(5): 1298-1308. (in Chinese)
[20] 马吉权, 赵淑敏, 孔凡辉. 多尺度条形池化与通道注意力的图像语义分割[J]. 中国图象图形学报, 2022, 27(12): 3530-3541.
MA Jiquan, ZHAO Shumin, KONG Fanhui. Semantic image segmentation by using multi-scale strip pooling and channel attention [J]. Journal of Image and Graphics, 2022, 27(12): 3530-3541. (in Chinese)
[21] LIANG F, QIAN C, YU W, et al. Survey of graph neural networks and applications [J]. Wireless Communications and Mobile Computing, 2022(1): 9261537.
[22] 许智宏, 张天润, 王利琴, 等. 融合图谱重构的时序知识图谱推理[J]. 计算机工程与应用, 2024, 60(9): 181-187.
XU Zhihong, ZHANG Tianrun, WANG Liqin, et al. Temporal knowledge graph reasoning with graph reconstruction [J]. Computer Engineering and Applications, 2024, 60(9): 181-187. (in Chinese)
[23] PANCINO N, GALLEGATI C, ROMAGNOLI F, et al. Protein - protein interfaces: A graph neural network approach [J]. International Journal of Molecular

Sciences, 2024, 25(11): 5870.
[24] ACHANTA R, SHAJI A, SMITH K, et al. SLIC superpixels compared to state-of-the-art superpixel methods[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(11): 2274-2282.
[25] ZHOU H, LUO F, ZHUANG H, et al. Attention multihop graph and multiscale convolutional fusion network for hyperspectral image classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 5508614.
[26] SHI C, WU H, WANG L. CEGAT: A CNN and enhanced-GAT based on key sample selection strategy for hyperspectral image classification[J]. Neural Networks, 2023, 168: 105-122.
[27] PENG F, LU W, TAN W, et al. Multi-output network combining GNN and CNN for remote sensing scene classification[J]. Remote Sensing, 2022, 14(6): 1478.
[28] YING C, CAI T, LUO S, et al. Do transformers really perform bad for graph representation? [DB/OL]. (2021-06-09)[2025-01-06]. https://arxiv.org/abs/2106.05234.
[29] HAN K, WANG Y, GUO J, et al. Vision GNN: An image is worth graph of nodes[DB/OL]. (2022-11-04)[2025-01-06]. https://arxiv.org/abs/2206.00272.
[30] CHEN J, LIU W, HUANG Z, et al. Universal deep GNNs: Rethinking residual connection in GNNs from a path decomposition perspective for preventing the over-smoothing[DB/OL]. (2022-05-30)[2025-01-06]. https://arxiv.org/abs/2205.15127.
[31] ZHAO Y, ZHENG Q, ZHU P, et al. TUFusion: A transformer-based universal fusion algorithm for multimodal images[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(3): 1712-1725.
[32] MA J, TANG L, FAN F, et al. SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer[J]. IEEE/CAA Journal of Automatica Sinica, 2022, 9(7): 1200-1217.
[33] LI J, CHEN J, LIU J, et al. Learning a graph neural network with cross modality interaction for image fusion[DB/OL]. (2023-08-07)[2025-01-06]. https://arxiv.org/abs/2308.03256.
[34] TANG W, HE F, LIU Y. ITFuse: An interactive transformer for infrared and visible image fusion[J]. Pattern Recognition, 2024, 156: 110822.
[35] TANG W, HE F, LIU Y, et al. DATFuse: Infrared and visible image fusion via dual attention transformer[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(7): 3159-3172.
[36] AZAM M A, KHAN K B, SALAHUDDIN S, et al. A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics[J]. Computers in Biology and Medicine, 2022, 144: 105253.