DOI:10.11918/202507016

动态门控扩散去噪与跨层注意力的多模态图像融合网络

邸 敬[1],霍婧婧[1],王鹤然[1],刘冀钊[2],廉 敬[1]

(1.兰州交通大学 电子与信息工程学院,兰州 730070;2.兰州大学 信息科学与工程学院,兰州 730000)

摘 要:针对去噪扩散模型在图像融合任务中难以适应不同噪声水平、普通残差块对特征的筛选能力有限的问题,本文构建了一种动态门控扩散去噪与跨层注意力的多模态图像融合网络。首先,设计并引入4组专家卷积核至动态特征提取器模块,根据输入内容动态组合出最优卷积核,对输入特征实现自适应处理;其次,提出了一种改进的门控特征选择模块来生成门控信号抑制无关信息,提升模型在不同噪声水平下的扩散去噪能力,实现对特征的精准控制;最后,使用R-Transformer块进行特征调整,通过构建的全局-局部空间注意力模块实现跨层特征融合,以生成纹理信息丰富、色彩保真度高的融合图像。在MSRS、RoadScene和Harvard三个数据集上的实验结果表明,与近年来图像融合领域中9种具有代表性的重要方法相比,本文方法的7种客观评价指标平均提升了5.11%~15.93%。本文方法在纹理细节保持及解剖结构完整性保留等方面均优于其他方法,符合人眼视觉特性,能够很好地处理各种光照环境场景和医学影像诊断场景下的多模态图像融合任务。

关键词:多模态图像融合;扩散模型;门控特征选择模块;跨层注意力融合模块;专家卷积核

中图分类号:TP391 **文献标志码**:A **文章编号**:0367-6234(2026)05-0033-12

Dynamic gating diffusion denoising and cross-layer attention-based multimodal image fusion network

DI Jing[1], HUO Jingjing[1], WANG Heran[1], LIU Jizhao[2], LIAN Jing[1]

(1. School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China;
2. School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China)

Abstract: To address the challenges that denoising diffusion models struggle to adapt to varying noise levels and conventional residual blocks have limited feature selection capability in image fusion tasks, this paper constructs a multimodal image fusion network integrating dynamic gating diffusion denoising and cross-layer attention. Firstly, four groups of expert convolution kernels are designed and incorporated into the dynamic feature extractor module. The optimal convolution kernels are dynamically assembled based on input content, enabling adaptive processing of input features. Secondly, an improved gated feature selection module is proposed to generate gating signals that suppress irrelevant information, enhance the model's diffusion denoising capability under different noise levels, and achieve precise feature control. Finally, R-Transformer blocks are adopted for feature adjustment. A global-local spatial attention module is constructed to realize cross-layer feature fusion, thereby generating fused images with rich texture information and high color fidelity. Experimental results on the MSRS, RoadScene, and Harvard datasets demonstrate that compared with 9 representative state-of-the-art methods in the field of image fusion in recent years, the proposed method achieves an average improvement of 5.11% to 15.93% across 7 objective evaluation metrics. The proposed method outperforms other counterparts in texture detail preservation and anatomical structure integrity maintenance, conforms to human visual perception characteristics, and can effectively handle multimodal image fusion tasks in scenarios such as various lighting environments and medical image diagnosis.

Keywords: multimodal image fusion; diffusion models; gated feature selection module; cross-layer attention fusion module; expert convolutional kernels

图像融合技术旨在通过特定算法将多幅来源不同、模态不同的图像信息进行结合,生成一幅包含更多互补信息、更符合人类视觉感知的新图像。其核心目标是整合源图像中的互补信息与重复出现的相

收稿日期:2025-07-08;**录用日期**:2025-08-19;**网络首发日期**:2025-11-06
网络首发地址:https://link.cnki.net/urlid/23.1235.T.20251106.1008.004
基金项目:甘肃省自然科学基金(24JRRA231);国家自然科学基金(62061023);甘肃省科技计划重点研发计划(24YFFA024)
作者简介:邸 敬(1979—),女,副教授,硕士生导师
通信作者:霍婧婧,Hbingcheng@126.com

似信息,提升图像的清晰度、语义丰富度和诊断分析价值[1-5]。图像融合的典型应用场景分为红外与可见光图像融合[6-8]和医学图像融合[9-10]。其中,红外与可见光图像融合的目标是融合红外图像的热辐射信息与可见光图像的纹理细节,为安防、军事侦察、道路交通等提供便利;医学图像融合的目标为融合医学影像的内部结构信息与功能代谢信息,以辅助肿瘤定位、病情诊断。

基于深度学习的图像融合算法因其强大的特征提取能力获得了广泛应用,算法模型架构主要包括卷积神经网络(CNN)模型、基于 Transformer 结构的模型和生成对抗网络(GAN)模型等。CNN 通过局部感受野机制实现图像层级特征的高效提取,这一特性也使其拥有了良好的空间建模能力。Prabhakar 等[11]提出深度特征融合网络 DeepFuse,该网络是利用深度卷积网络从源图像的亮度通道中提取信息,因未充分提取源图像信息,融合图像效果不佳。Transformer 类的深度学习算法擅长建模长距离依赖关系,通过自注意力机制捕捉全局上下文信息。Ma 等[12]提出 SwinFusion 网络,其跨域注意力设计可捕获整合同模态长距离依赖、感知连接跨模态语义关联,以提升多模态融合效果,但在模态差异极大的场景中融合图像会出现细节模糊等问题。GAN 通过生成器和判别器的对抗训练,实现高质量数据生成。Xi 等[13]提出了一种结合多尺度注意力网络与期望最大化算法的红外-可见光图像融合生成对抗网络(generative adversarial network with multiscale attention network and expectation maximization algorithm, EMA-GAN)用于图像融合,EM 算法有助于解决红外与可见光图像融合中标签缺失的问题,但 GAN 面临训练不稳定的挑战,并且生成效果易受数据分布影响。

近年来,去噪扩散概率模型备受关注,其具有良好的数学可解释性,且图片生成过程稳定可控,在图像融合领域得到广泛应用。Zhao 等[14]提出了多模态图像融合去噪扩散模型(denoising diffusion model for multi-modality image fusion, DDFM),该模型借助去噪扩散概率模型(denoising diffusion probabilistic models, DDPM)的生成先验和 EM 算法推理,解决了 GAN 训练不稳定和可解释性不足的问题,但此模型依赖预训练生成模型,导致模型泛化能力差。为解决现有红外与可见光图像融合方法无法直接处理多通道数据的问题,Yue 等[15]提出了基于扩散模型的红外-可见光高色彩保真度图像融合模型(toward high color fidelity with diffusion models, Dif-Fusion),通过隐空间多通道扩散过程直接聚合红外与可见光的多源特征,显著提升色彩保真度与纹理强度的保留效果,但在低光照可见光图像或者高噪声红外图像中,扩散过程可能出现过度平滑细节的问题。为解决扩散模型在图像融合中缺乏真实标签、难以直接应用的问题,Yi 等[16]提出了基于融合知识先验扩散模型的多模态图像融合模型(multi-modality image fusion via diffusion model with fusion knowledge prior, Diff-IF),该模型可有效保留视觉信息与弱纹理细节,但其依赖目标搜索确定的先验分布,对复杂或未见模态组合泛化性不足。现有多模态图像融合方法无法处理源图像中的曝光异常以及目标显著性不足,Zhang 等[17]提出了基于文本调制扩散模型的交互式多模态图像融合框架(interactive multi-modal image fusion framework based on text-modulated diffusion model, Text-DiFuse),通过扩散过程集成特征级信息融合,但该模型依赖文本输入的准确性与零样本模型的泛化能力,对无文本描述或语义模糊的融合场景处理效果不佳。Gao 等[18]提出了基于相位转移扩散模型的光学错觉隐藏图像生成模型(free lunch for generating optical illusion hidden pictures with phase-transferred diffusion model, PTDiffusion),其通过在去噪过程中移植扩散特征的相位谱,生成具有隐藏视觉线索的视错觉图像,但由于免训练特性,对复杂图像的处理受限。Wei 等[19]提出了基于一步扩散模型的多模态图像配准模型(multimodal image registration based one step diffusion model, OSDM-MReg),通过一步扩散以及多尺度配准网络来融合特征,但在极端场景适应性上存在局限。

针对以上扩散模型在图像融合任务中难以适应不同噪声水平及普通残差块对特征的筛选能力有限的问题,本文提出一种动态门控扩散去噪与跨层注意力的多模态图像融合网络。通过动态门控扩散去噪模块提取图像特征,对输入特征进行自适应处理,以增强这些特征的表达并保留细节信息。同时,本文采用残差连接的方式缓解梯度消失问题,从而提取更全面的全局语义信息和局部细节特征。通过跨层注意力模块进行多阶段的特征融合和重建,逐步增强图像细节并抑制噪声,生成既包含丰富纹理信息又突出热辐射目标的高质量融合图像。

1 图像融合方法

1.1 网络总体框架

在 UNet 结构中,传统卷积层参数固定,难以适应不同噪声水平和复杂特征分布,且普通残差块对特征的选择能力有限,对信息流的控制效果不佳。因此,本文提出一种动态门控扩散去噪与跨层注意

力的多模态图像融合网络，网络总体架构见图 1。首先，将 4 组专家卷积核引入动态特征提取器模块，通过 1×1 卷积门控网络生成专家权重，根据输入内容动态组合最优卷积核，对输入特征实现自适应处理。其次，利用门控特征选择模块生成的门控信号过滤无关特征，增强关键信息的提取与保留能力。最后，通过跨层注意力模块中的 R-Transformer 块进行特征调整，并利用全局–局部空间注意力机制实现全局与局部特征的有效融合，从而避免高频信息丢失，生成兼具底层细节与高层语义特征的融合图像。

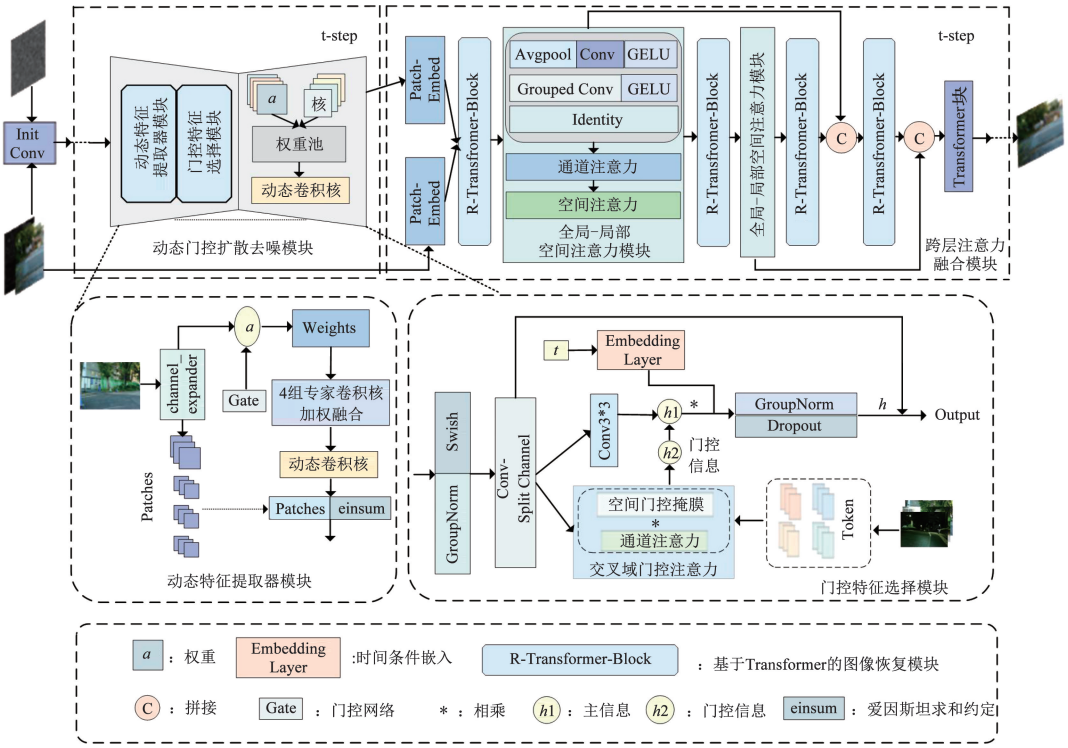


图 1　网络总体架构

Fig. 1　General network framework diagram

1.2　动态门控扩散去噪模块

在扩散模型中，数据生成过程被建模为逐步加噪的反向去噪过程。前向过程通过高斯扰动将干净图像 x_0 扩散为不同噪声水平下的中间状态 x_t，其计算公式为

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon,\ \varepsilon \sim N(0,1) \tag{1}$$

式中 $\bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$ 控制噪声随时间的积累。反向过程目标是通过神经网络 $\varepsilon_\theta(x_t,t)$ 预测噪声，并逐步恢复数据，其计算公式为

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_t(x_t,t)\right) + \sigma_t z,\ z \sim N(0,I) \tag{2}$$

由于 x_t 在不同时间步 t 的特征分布变化不同，且包含不同程度的噪声扰动，因此网络 ε_θ 需具备动态适应性，以结合当前时间步信息实现更精准的噪声去除。

为此，本文提出了动态门控扩散去噪模块，图 2 为动态门控扩散去噪模块内部结构。首先，输入图像进入动态特征提取器模块，其使用权重融合后的卷积核动态适应不同时间下的特征处理需求，以选择重要特征通道；然后，进入门控特征选择模块，该模块通过门控信号对特征通道进行加权选择，强调关键信息并过滤无关特征。如图 2(b)动态门控扩散去噪模块所示，上采样和下采样过程各包含 4 个阶段，每个阶段均由一个动态特征提取器模块和一个门控特征选择模块顺序组成，在编码路径的最深层，包含一个门控特征选择模块，进一步处理下采样的最深层次特征，为后续上采样提供基础。通过动态特征提取器与门控特征选择模块，该结构自适应地增强了网络从局部细节到全局结构的特征提取与建模能力，提升了扩散过程中不同噪声水平下的去噪能力以及特征提取能力。

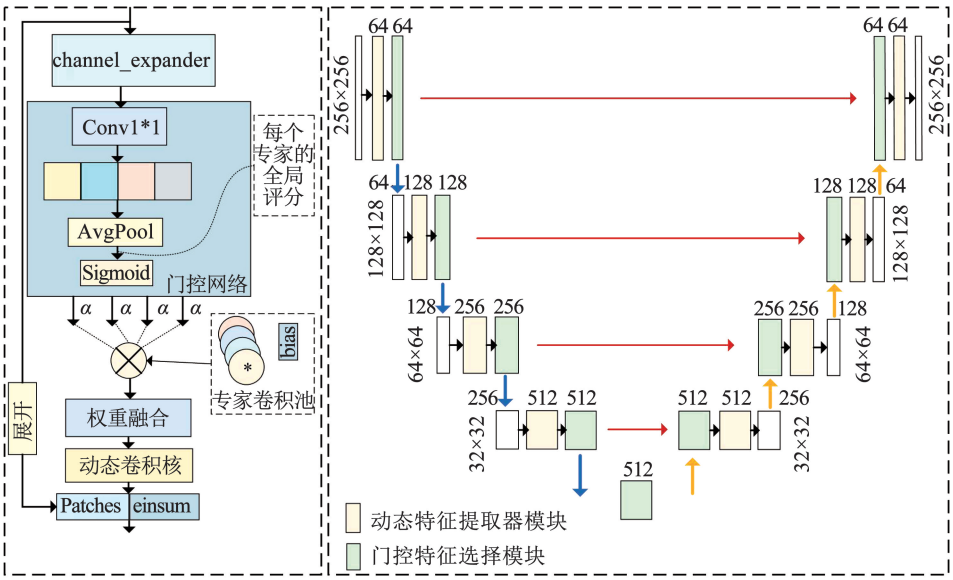
1.2.1　动态特征提取器模块

普通卷积的通道变换依赖固定设计且卷积核权重固定，难以处理复杂噪声分布和多尺度特征，对图像特征提取效果不佳。动态特征提取器的核心思想是引入多个专家卷积核，根据输入内容为每个样本自适应生成一组专属的卷积权重，从而增强模型的特征提取能力。

如图 2(a)动态特征提取器模块所示,首先,通过通道变换层将输入特征图统一投影到特定通道,形成统一维度便于后续不同专家共享处理。随后,门控模块(Gate)通过卷积层和全局平均池化对各专家进行评分以确定权重,结合 softmax 函数得到每个样本在不同专家上的概率分布。在此基础上,利用已知的专家权重对多个专家卷积核实施加权融合,最终得到适配每个样本的动态卷积核,其计算公式为

$$W^{(b)} = \sum_{m=1}^{M} \alpha_{b,m} \cdot W^{(m)} \tag{3}$$

式中:$W^{(b)}$ 为第 b 个样本所使用的最终动态卷积核;M 为专家块数量,该网络使用的专家块数量为 $M=4$,每个专家拥有一套独立的卷积核;$\alpha_{b,m}$ 为第 b 个样本对第 m 个专家的权重系数,表示该样本在融合卷积核时,对第 m 个专家的信任度;$W^{(m)}$ 为第 m 个专家的卷积核参数,维度与 $W^{(b)}$ 相同。



(a) 动态特征提取器模块 (b) 动态门控扩散去噪模块

图 2 动态门控扩散去噪模块内部结构

Fig. 2 Internal architecture of the dynamic gated diffusion denoising module

得到动态卷积核后,将输入特征图展开为一系列小块,即局部感受野区域,而后使用 einsum 将每个样本与其专属的卷积核实现逐位置滑窗卷积,其计算公式为

$$y_l^{(b)} = \sum_{c,i,j}^{e} (W_{o,c,i,j}^{(b)} \cdot x_{l,c,i,j}^{(b)}) + b_o^{(b)} \tag{4}$$

式中:$y_l^{(b)}$ 为第 b 个样本的第 l 位置的输出,表示输出特征图上一个像素点的某个通道值,$\sum^{e}$ 表示对输入通道 c 及卷积核空间位置 (i,j) 进行求和,$b_o^{(b)}$ 为第 b 个样本的第 o 个输出通道的偏置项。而后使用式(5)和式(6)恢复卷积输出格式:

$$H_{\text{out}} = \frac{H + 2 \cdot \text{pad} - K}{\text{stride}} + 1 \tag{5}$$

$$W_{\text{out}} = \frac{W + 2 \cdot \text{pad} - K}{\text{stride}} + 1 \tag{6}$$

式中:H,W 为输入特征图的高、宽,K 为卷积核的尺寸,pad 为对输入边缘的补零像素数(pad = $K//2$),保持输入输出大小一致。

最后将所有局部结果拼接得到输出特征图。通过以上步骤,每个样本根据内容选择专家,对多个专家卷积核线性加权融合,再选择重要特征通道,增强模型对不同输入的适应能力与表达力,提高特征提取能力。

1.2.2 门控特征选择模块

输入特征经动态特征提取器模块处理后进入门控特征选择模块,通过门控机制抑制无关特征,每个 Token 都有一个对应的门控信号,用于控制这个 Token 的输出是否被放大、抑制或归零。门控特征选择模块如图 3 所示,该模块首先对输入张量进行分组归一化,提升训练稳定性;而后通过卷积层将通道数扩展为 2 倍,并沿通道维度将特征拆分为两部分,第 1 部分经 3×3 卷积后作为主信息 h_1,第 2 部分使用深度可分离卷积生成空间门控掩膜,以此来关注重要区域,同时使用全局平均池化和全连接层生成通道注意力,而后将空间信息与通道信息相乘,再经 Sigmoid 函数压缩生成门控权重,得到门控信息 h_2,h_1 和 h_2 逐元素相乘实现门控单元的特征筛选。将噪声时间嵌入 t_{emb} 通过仿射变换调整特征,对门控后的特征进行缩放和平移,注入噪声条件信息,如公式(7)所示

$$h=\gamma(t_{\text{emb}})\cdot h+\beta(t_{\text{emb}}) \tag{7}$$

然后对结果再次分组归一化并应用 Dropout 正则化防止过拟合，通过卷积层恢复原始输出通道数，完成特征提取。最后将处理后的特征与输入进行残差相加，解决梯度消失问题，输出计算公式为

$$\text{Output}=h+W_{\text{residual}}(x) \tag{8}$$

动态门控扩散去噪模块动态选择重要特征通道，过滤无关特征，有效缓解了高层语义建模中特征选择粗糙、关键信息易被噪声干扰的问题，提升了特征表达的针对性与有效性。

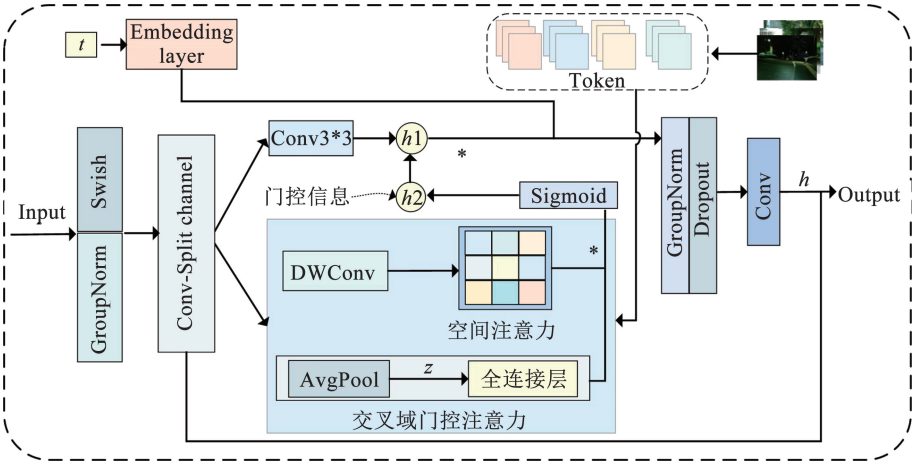


图 3　门控特征选择模块

Fig. 3　Gated feature selection module

1.3　跨层注意力融合模块

为提取更全面的全局语义信息和局部细节特征，本文设计了跨层注意力融合模块，如图 4 所示。双分支输入图像使用带有步长小于卷积核大小的卷积层（Patch-Embed 层），使得提取的图像块之间存在重叠区域，从而更好地捕捉局部信息。模型采用双阶段的逐步增强策略，第 1 阶段中，输入图像通过带有条件调制的 R-Transformer 块进行初步特征提取与调整，以捕获多模态图像间的浅层结构信息，而后通过全局–局部空间注意力模块实现跨分支的特征交互与融合，动态分配不同模态的重要性权重。在第 2 阶段，融合后的特征被进一步送入新的 R-Transformer 块中，强化深层语义信息的建模能力，而后再次通过全局–局部空间注意力模块进行特征整合，从更大的感受野中融合结构与上下文信息，使模型从浅层到深层逐步提升对多模态图像的理解与表达能力。同时，通过位置编码与多层感知机处理时间步长信息可得到噪声嵌入向量，该向量通过双阶段逐步增强中 R-Transformer 块后的条件特征变换模块（FeatureWiseAffine），在时间步嵌入向量 $\boldsymbol{T}$ 对特征图缩放与偏移的控制下，实现对噪声条件的感知。两个R-Transformer块通过自注意力，增强跨层拼接的多尺度特征表达。第 1 次拼接将上采样后的高级语义特征与第 1 阶段的低级细节特征拼接，使得特征兼具语义深度与细节精度；第 2 次拼接将解码器输出特征与初次融合并调整尺寸的特征拼接，强化特征表达并保留中间信息，以恢复图像细节。最后，通过 Transformer 块对融合特征进一步优化，输出同时保留多模态细节信息与热辐射特征的高质量融合图像。

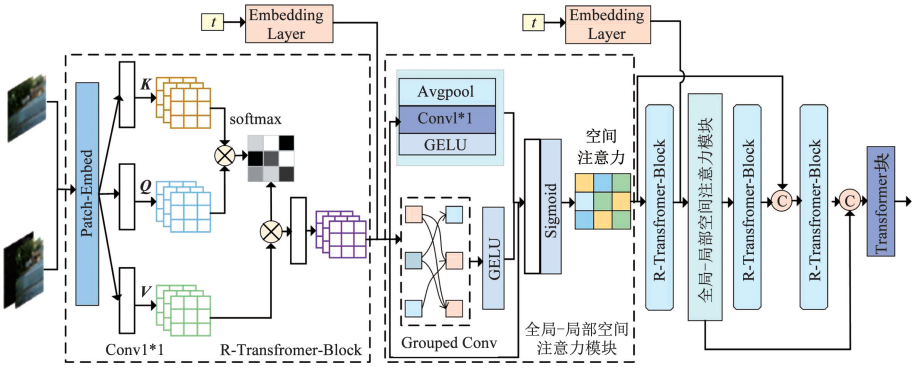


图 4　跨层注意力融合模块

Fig. 4　Cross-layer attention fusion module

传统方法对不同层次的特征直接拼接或相加，未考虑通道和空间维度的重要性差异。跨层注意力融合模块实现轻量级多尺度特征融合，通过全局–局部双路特征提取、通道注意力校准和空间注意力

增强三阶段处理,融合来自不同层的特征图。阶段1通过全局平均池化压缩空间信息,1 × 1 卷积降维,上采样提取全局信息,而后通过分组卷积和 GELU 激活函数增强非线性以提取局部特征,具体如式(9) ~ (11)所示:

$$F_{\text{global}} = \text{Upsample}(\text{Conv}_{1\times1}(\text{GAP}(F_t))) \quad (9)$$

$$F_{\text{local}} = \text{GELU}(\text{GroupConv}(F_t)) \quad (10)$$

$$F_{\text{concat}} = \text{Concat}(F_t, F_{\text{global}}, F_{\text{local}}) \quad (11)$$

式中 F_t 为原始输入特征。拼接原始、全局、局部特征后,阶段 2 通过 1 × 1 卷积和 Sigmoid 生成通道注意力图,对原始特征进行通道注意力加权,来完成通道校准,如式(12)和式(13)所示:

$$M_{\text{c}} = \sigma(\text{Conv}_{1\times1}(F_{\text{concat}})) \quad (12)$$

$$F_{\text{channel}} = M_{\text{c}} \odot F_t \quad (13)$$

式中:M_{c} 为注意力权重,σ 代表 Sigmoid 函数,$\odot$ 表示逐通道加权。阶段 3 在对融合特征应用多头自注意力后,与原始特征进行残差相加,实现空间注意力增强。如式(14) ~ (16)所示:

$$\text{Attn}(F_{\text{channel}}) = \text{softmax}\left(\frac{F_{\text{channel}}W_Q\,(F_{\text{channel}}W_K)^{\text{T}}}{\sqrt{d_{\text{k}}}}\right)\cdot F_{\text{channel}}W_V \quad (14)$$

$$F_{\text{spatial}} = \text{Norm}(F_{\text{channel}} + \text{Attn}(F_{\text{channel}})) \quad (15)$$

$$F_{\text{output}} = F_{\text{spatial}} \quad (16)$$

全局–局部空间注意力模块融合局部与全局信息,避免了高频信息丢失,通道注意力抑制噪声主导的通道,空间注意力强化有效区域,通过轻量级多尺度特征融合与注意力机制,在细节恢复与计算效率方面,优于传统卷积和普通 Transformer 架构。

2 实验结果与分析

2.1 实验概况与数据说明

2.1.1 实验平台及参数设置

实验硬件平台操作系统为 Windows11,硬件配置为3th Gen Intel(R) Core(TM) i9-13900HX 2.20 GHz 处理器和 RTX 4070Ti 16G GPU。实验训练阶段和测试阶段均在 Pytorch 框架下实现,软件环境为 PyTorch 1.13.1 + cu117、CUDA 12.7。训练过程中,优化器采用 Adam,初始学习率为 0.000 1,动态特征提取器模块中专家卷积块的数量为 $N = 4$。

2.1.2 数据集

在训练阶段,采用多光谱道路场景 MSRS 数据集中的 1 083 对数据作为训练集。在测试阶段,分别从 Harvard 数据集、MSRS 数据集和 RoadSence 数据集中选择93 对、361 对和 30 对图像进行实验验证。

2.1.3 对比算法及评价指标

为验证本文融合网络的性能,选择了 9 种典型的深度学习算法与其进行对比,即 SeAFusion[20]、MUFusion[21]、U2Fusion[22]、DDFM、MetaFusion[23]、GIFusion[24]、BTSFusion[25]、SDCFusion[26] 和 MLFuse[27] 融合算法。

为评价网络模型的有效性,选择了 7 种客观评价指标来衡量融合结果,包括信息熵[28](EN)、互信息[29](MI)、基于结构相似性[30](SSIM)、峰值信噪比[31](PSNR)、空间频率[32](SF)、边缘强度[33]($Q^{\text{AB/F}}$)及相关系数[34](CC)。其中,EN 是衡量图像信息量的重要指标,数值越大说明融合图像包含信息越丰富。MI 反映融合图像和源图像之间的信息重叠度,互信息越大,融合质量越好。SSIM 用于评估融合图像和源图像的结构一致性,数值越接近 1,融合效果越好。PSNR 通过衡量图像有效信息与噪声之间的比率来反映图像是否失真,数值越大,图像融合质量越好。SF 用于衡量图像灰度的变化,数值越大,图像越清晰,融合质量越好。$Q^{\text{AB/F}}$ 用于量化边缘信息传递能力,数值越大,边缘保留越完整。CC 用于衡量融合图像与源图像的空间线性相关程度,其值越接近 1 或者 –1,表示融合图像包含源图像信息越多,融合效果越好。

2.2 图像融合对比实验

2.2.1 MSRS 数据集实验验证

在 MSRS 数据集中选取白天和夜间共 6 组具有代表性的红外与可见光图像场景,对 10 种方法进行主观和客观比较。图 5 为 MSRS 数据集 6 组场景的融合结果,场景一、场景二、场景三、场景四为白天景象,场景五、场景六为夜间景象。通过图 5 可以看到,SeAFusion 算法引入边缘注意力机制,更好地保留了可见光图像中的边缘和纹理细节信息,但当图像中不同区域之间的亮度、颜色差异较小时,易导致目标与背景难以区分,如场景二的树木部分。U2Fusion 算法可以捕捉图像细节信息,但存在纹理细节模糊的问题,使图像色彩保真度严重降低。MUFusion 算法虽然信息融合全面,但对边缘细节纹理的保护效果不佳,如场景五灯的边缘。DDFM 算法融合后图像具有目标凸显性,但合成图像视觉上较暗,导致融合图像缺失部分信息。MetaFusion 算法对图像中的小目标或存在遮挡的场景处理存在不足,融合结果的目标显著性下降,如场景三中的广告牌。GIFusion 算法使用的对抗训练容易陷入模式崩溃,导致融合结果出现色彩失真,如场景四中的楼房。BTSFusion 算法未能有效保留红外图像中较微弱的热辐射信息,对场景六中的指示牌边缘处理不够清晰。SDCFusion 算法对分割任务依赖严重,分割网络的滞后可能导致融合结果与实际场景的时间错位,场景四中对移动的车辆融合效果不佳。MLFusion 算法聚焦于像素层面的特征融合,却对物体类别等高层语义信息的利用不够充分。本文算法无论在白天还是夜间场景,融合图像均能自适应保

留有效信息，目标与背景的边界过渡平滑，无明显伪影或色彩断层，关键目标如车辆、行人的热信号被精准强化，即使处于树林、阴影等复杂背景中仍能够快速被识别。

图 5　MSRS 数据集 6 组场景的融合结果

Fig. 5　Fusion results of six scenarios from the MSRS dataset

6 组源图像的融合结果表明,本文提出的融合网络生成的图像兼具高层语义信息与底层纹理细节,色彩还原度高且贴近真实场景。为了使融合结果更具说服力,本文进一步选取 MSRS 数据集中 30 对图像客观评价指标的均值对融合结果进行分析。表 1 展示了不同融合算法在 MSRS 数据集上的定量结果比较,可以看到,本文算法在 EN、MI、SSIM、PSNR、SF、$Q^{AB/F}$ 和 CC 等客观评价指标中均取得了较好的结果,在 MI、PSNR、$Q^{AB/F}$、CC 上取得最优结果,最优的 MI 结果表明,融合算法有效提取了可见光图像的纹理细节和红外图像的热目标信息,整合了这些互补信息,提升了图像的综合可读性。最优的 PSNR 结果说明,融合算法在保留源图像细节的同时,有效抑制了噪声、模糊或伪影等失真。最优的 $Q^{AB/F}$ 表明,融合算法能够有效提取出源图像中的轮廓、纹理、边界信息。最优的 CC 结果表明,融合图像保留了源图像的整体亮度、对比度和结构特征,视觉效果更佳。因此,通过主观评估与客观指标分析,本文算法的融合图像在纹理细节、边缘分布等底层特征,结构、语义等高层特征方面优于其他对比算法。

表 1 不同融合算法在 MSRS 数据集上的定量结果比较

Tab. 1 Quantitative comparison of different fusion algorithms on the MSRS dataset

算法	EN	MI	SSIM	PSNR	SF	$Q^{AB/F}$	CC
SeAFusion[20]	7.065 4	4.358 8	0.631 0	21.205 4	13.032 7	0.655 0	0.984 2
MUFusion[21]	5.408 1	1.562 5	0.588 7	16.167 0	11.212 7	0.383 0	0.765 3
U2Fusion[22]	5.629 0	1.932 5	0.596 0	17.011 9	9.367 5	0.326 2	0.733 9
DDFM[14]	6.667 0	2.900 3	0.636 3	17.460 0	9.030 1	0.434 9	0.952 6
MetaFusion[23]	6.652 0	1.690 9	0.533 3	15.698 4	13.249 6	0.353 1	0.879 2
GIFusion[24]	6.756 0	2.760 9	0.653 7	17.617 3	12.156 6	0.602 8	0.953 5
BTSFusion[25]	6.734 7	2.372 8	0.585 5	17.128 3	**13.950 4**	0.500 1	0.926 0
SDCFusion[26]	**7.090 3**	4.237 0	0.629 7	21.013 3	13.565 7	0.645 0	0.982 3
MLFuse[27]	6.740 6	3.131 8	0.637 8	18.269 1	12.071 9	0.560 9	0.970 0
Ours	7.078 7	**4.582 1**	**0.641 1**	**21.550 2**	13.642 9	**0.655 2**	**0.984 4**

2.2.2 Harvard 数据集实验验证

为了进一步验证本文算法的优越性,从 Harvard 测试集中选取 30 组包含 MRI-CT、MRI-PET 和 MRI-SPECT 的具有代表性的脑部医学图像进行分析,医学图像融合目标是整合 CT/MRI 的解剖结构与 PET/SPECT 的代谢功能的互补信息,提升病灶辨识度、诊断准确性,为临床提供更全面的决策依据。图 6为不同融合算法在 Harvard 数据集上的定性结果比较,通过图 6 可以看出,U2Fusion、MUFusion、SwinFusion 算法淡化了微小病灶边缘、纹理,无法有效突出 CT 中高密度对比增强区域。DDFM 算法融合后细节易模糊,对不同模态特征权重分配不够合理,过度强化 CT 的轮廓信息,导致 MRI 中软组织纹理细节被弱化,无法清晰呈现病灶与周围软组织的关系。MetaFusion 算法融合后微小结构边缘模糊,与周围组织对比度降低,不利于分析病灶。GIFusion 算法无法保留原始图像的强度分布,色彩保真度欠佳,丢失了部分结构信息。TarDAL 算法[35]难以精准权衡代谢功能信息与解剖结构信息,使得 MRI 图像里软组织的纹理、边界等关键特征被弱化。SDCFusion 算法对脑部微小血管、神经纤维等精细结构处理不佳,融合后易出现细节模糊、丢失。MLFusion 算法的动态调整能力不足,对于解剖结构复杂的部位,MRI 的细节信息被 PET 的代谢信号干扰,导致解剖结构显示不清。而本文算法有效抑制了噪声与伪影,保留了关键解剖信息与代谢功能信息,在细节纹理、边缘清晰度及信息保真度上均优于其他对比算法。此外,选取了 EN、MI、SSIM、PSNR、SF、$Q^{AB/F}$ 和 CC 对这些先进方法进行定量评估,表 2 为不同融合算法在 Harvard 数据集上的定量结果比较。如表 2 所示,本文方法在所有对比算法的大多数评价指标上均取得了最优或次优结果,优于其他对比方法,且在指标 MI、SSIM、$Q^{AB/F}$、CC 上取得最优结果,表明本文方法在细节纹理、边缘信息保留以及可视化方面优于其他模型方法。

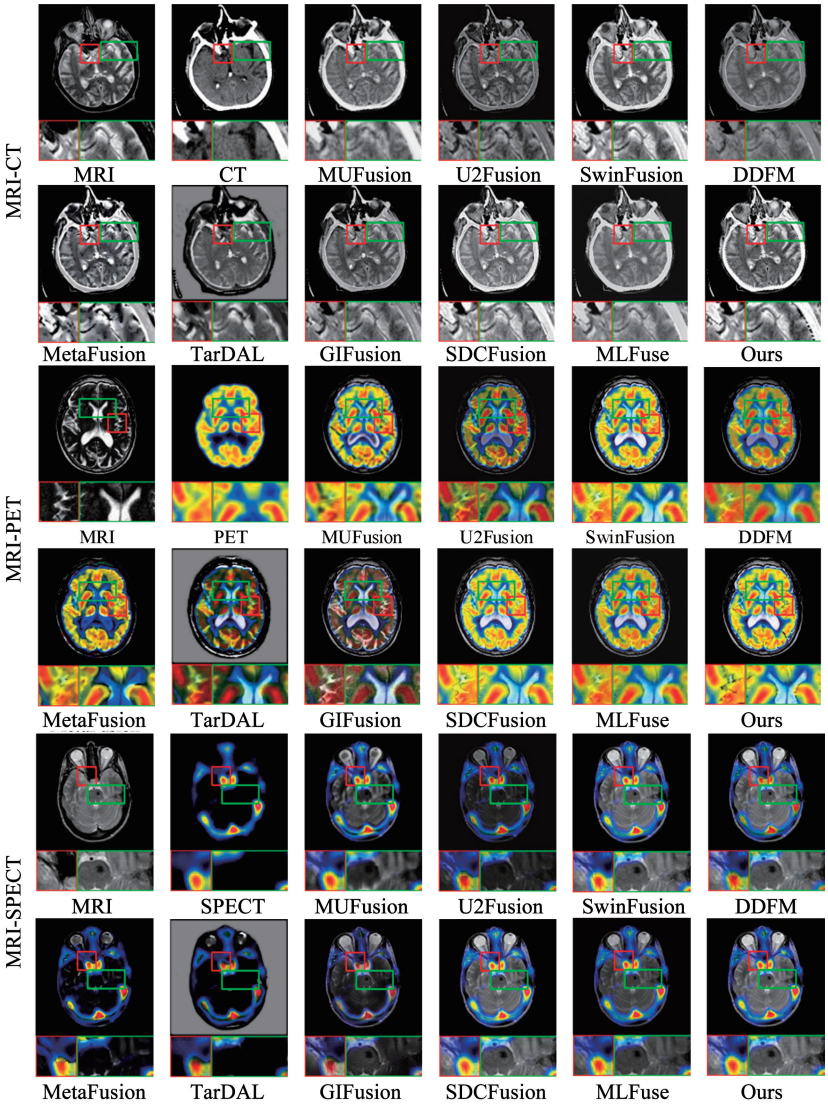


图 6　不同融合算法在 Harvard 数据集上的定性结果比较

Fig. 6　Qualitative comparison of different fusion algorithms on the Harvard dataset

表 2　不同融合算法在 Harvard 数据集上的定量结果比较

Tab. 2　Quantitative comparison of different fusion algorithms on the Harvard dataset

算法	EN	MI	SSIM	PSNR	SF	$Q^{AB/F}$	CC
MUFusion[21]	5.068 5	2.498 5	0.515 8	14.801 9	37.006 1	0.151 2	0.850 1
U2Fusion[22]	4.690 5	2.453 3	0.188 7	15.247 0	33.289 5	0.167 2	0.831 0
SwinFusion[12]	5.064 8	2.896 6	0.581 0	13.645 1	46.276 1	0.304 0	0.785 2
DDFM[14]	4.410 8	2.760 0	0.642 8	15.200 7	36.875 3	0.257 5	0.767 8
MetaFusion[23]	4.318 7	2.165 8	0.631 7	15.170 5	30.272 4	0.138 4	0.836 7
GIFusion[24]	5.283 9	2.317 0	0.552 5	15.724 7	45.651 7	0.316 6	0.711 8
TarDAL[35]	5.242 9	2.003 9	0.231 5	8.339 0	39.382 1	0.101 2	0.834 9
SDCFusion[26]	**5.343 5**	2.988 5	0.353 3	13.418 9	**56.909 5**	0.313 0	0.816 1
MLFuse[27]	4.950 4	2.672 4	0.390 6	14.320 2	40.101 5	0.205 2	0.850 1
Ours	5.268 3	**3.022 3**	**0.646 5**	**15.285 9**	56.192 3	**0.319 6**	**0.851 9**

2.2.3　RoadScene 数据集主客观评价

本文选取 RoadScene 数据集中的 20 组源图像进行泛化性测试，图 7 为不同融合算法在 RoadScene 数据集上的定性结果比较，可以看到 SeAfusion、MetaFusion 算法亮度表现良好，但其可见光图像的纹理细节、结构边缘等信息未能清晰展现，如树叶边

缘。U2Fusion、DDFM、GIFusion、MLFusion 算法过度侧重红外图像信息,可见光图像中的关键细节,如树叶、山脉的轮廓信息模糊。MUFuison 算法纹理细节模糊,汽车边缘识别不清。SDCFusion、BTSFusion 算法呈现了可见光图像的纹理细节以及红外图像的显著信息,整体效果较好,但融合图像色彩欠佳。从视觉对比结果可以看出,本文算法在清晰还原场景纹理细节的同时,有效保留了源图像的色彩信息,在整体视觉质量上展现出显著优势。表 3 为不同融合算法在 RoadScene 数据集上的定量结果比较,可以看到,本文算法在 MI、SSIM、$Q^{AB/F}$、CC 等指标上均表现优异,EN、SF、PSNR 偏低,是因为本文方法更关注红外显著信息和可见光细节纹理细节,但 EN 指标主要衡量图像的亮度信息,SF 指标则反映图像中像素灰度值的变化,而 PSNR 主要关注图像整体的像素值差异。对比分析可知,本文融合算法纹理清晰,色彩保真度高,在目标显著性和整体视觉效果方面均表现更优。

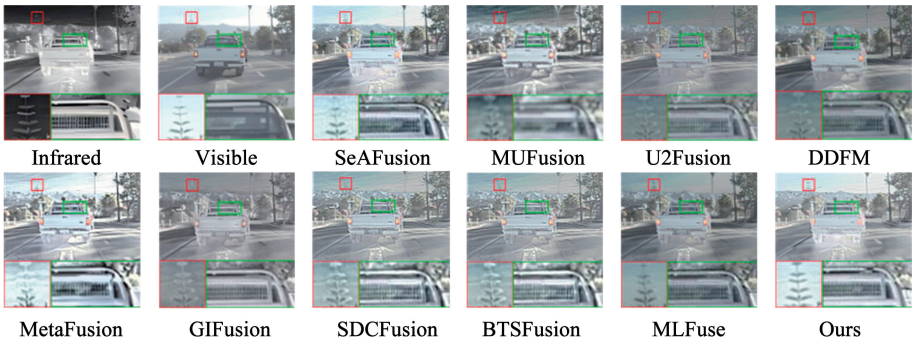


图 7 不同融合算法在 RoadScene 数据集上的定性结果比较

Fig. 7 Qualitative comparison of different fusion algorithms on the RoadScene dataset

表 3 不同融合算法在 RoadScene 数据集上的定量结果比较

Tab. 3 Quantitative comparison of different fusion algorithms on the RoadScene dataset

算法	EN	MI	SSIM	PSNR	SF	$Q^{AB/F}$	CC
SeAFusion[20]	7.433 1	3.121 0	0.650 0	14.059 8	18.485 9	0.118 0	0.661 3
MUFusion[21]	**7.438 0**	2.096 0	0.609 4	15.602 8	13.211 7	0.120 0	0.416 0
U2Fusion[22]	7.009 7	2.882 5	0.694 3	**15.982 3**	13.762 4	0.106 9	0.382 5
DDFM[14]	7.299 8	1.936 2	0.327 8	13.698 3	13.071 6	0.067 5	0.484 6
MetaFusion[23]	7.179 6	2.154 4	0.564 1	12.866 0	21.212 1	0.113 1	0.666 9
GIFusion[24]	7.222 4	1.863 4	0.338 1	14.953 5	16.585 9	0.064 3	0.347 9
BTSFusion[25]	7.300 7	2.237 6	0.608 3	15.735 0	**22.758 9**	0.092 6	0.612 3
SDCFusion[26]	7.362 9	2.727 7	0.655 4	15.420 0	18.001 6	0.122 0	0.661 4
MLFuse[27]	7.274 0	3.010 2	0.685 5	15.364 7	15.270 6	0.116 5	0.629 0
Ours	7.426 6	**3.667 8**	**0.738 1**	15.801 9	17.725 6	**0.124 1**	**0.690 1**

2.3 消融实验

为验证本文算法所提各个模块的有效性,选取 Harvard 数据集中 20 组图像进行验证,设计了 6 组消融实验。去掉全局-局部空间注意力模块,其余模块不变,记为 DFE + GCFS;去掉门控特征选择模块,其余模块不变,记为 DFE + GLSA;去掉动态特征提取器模块,其余模块不变,记为 GCFS + GLSA;将动态特征提取器模块使用的专家卷积块设置为 $N=3$,其余模块不变,记为 DFE_3;将动态特征提取器模块使用的专家卷积块设置为 $N=5$,其余模块不变,记为 DFE_5。实验验证了使用专家卷积块 $N=4$ 效果最佳。使用动态特征提取器模块提取特征,门控特征选择模块增强特征表达,全局-局部空间注意力模块生成融合图像,用于观察完整模型架构下的融合图像,记为 ALL。

为直观对比不同模块对融合效果的影响,本文随机选取 Harvard 数据集两组典型场景的融合结果进行主观分析,图 8 为上述 6 种消融实验的定性结果对比。由图 8 可以看出:去掉全局-局部空间注意力模块(DFE + GCFS),融合图像纹理细节信息模糊;去掉门控特征选择模块(DFE + GLSA),图像变暗,可以看到融合图像背景细节信息丢失;去掉动态特征提取器模块(GCFS + GLSA),导致网络关注显著信息的能力下降;将动态特征提取器模块使用的

专家卷积块设置为 $N=3$（DFE_3），融合图像变暗，红外信息缺失；将动态特征提取器模块使用的专家卷积块设置为 $N=5$（DFE_5），融合图像对比度不佳。综上所述，本文所提模型有效保留了纹理细节信息，同时视觉效果更佳。此外，选取 EN、MI、SSIM、PSNR、SF、$Q^{AB/F}$ 和 CC 指标作为消融实验的客观评价指标，表 4 为消融实验中 6 种不同网络结构的客观评价指标。由表 4 可以看到，本文算法在融合图像中既很好地保留了显著目标，同时又增强了图像中的细节和纹理信息，且更符合人眼视觉。

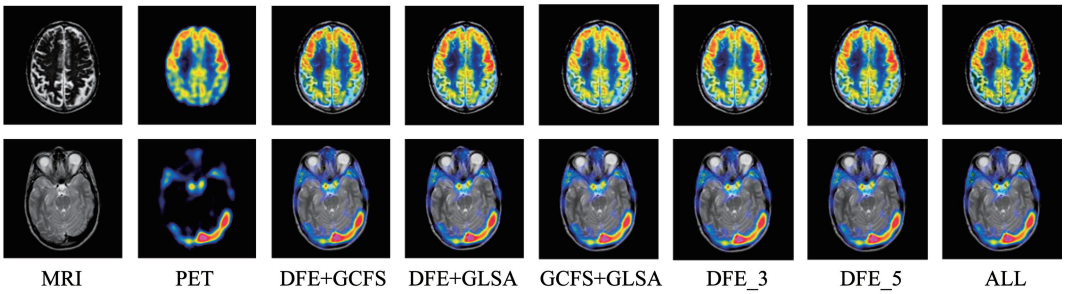


图 8　消融实验中 6 种不同网络结构的定性结果比较

Fig. 8　Qualitative comparison of six different network structures in ablation experiments

表 4　消融实验中 6 种不同网络结构的客观评价指标

Tab. 4　Objective evaluation metrics for six different network structures in ablation experiments

模型	EN	MI	SSIM	PSNR	SF	$Q^{AB/F}$	CC
DFE + GCFS	3.592 6	2.107 1	0.745 6	15.802 1	47.258 7	0.296 3	0.812 0
DFE + GLSA	3.220 5	1.970 1	0.748 2	16.451 2	44.973 5	0.275 1	0.887 8
GCFS + GLSA	3.404 5	2.073 2	0.530 2	15.468 1	46.841 0	0.267 0	0.816 1
DFE_3	3.384 5	2.068 4	0.764 0	16.113 9	51.178 4	0.251 0	0.872 5
DFE_5	3.172 7	1.985 9	0.756 8	16.344 0	45.127 8	0.294 4	0.878 7
ALL	**3.674 7**	**2.436 4**	**0.756 6**	**15.221 1**	**45.890 4**	**0.297 7**	**0.891 4**

3　结　论

1）本文设计了一种动态门控扩散去噪与跨层注意力的多模态图像融合网络，通过动态门控扩散去噪模块增强关键信息的提取能力，采用跨层注意力融合模块融合跨层信息，有效提升扩散过程中不同噪声水平下的去噪能力以及特征提取能力，实现了高质量多模态图像融合。

2）设计了动态特征提取器和门控特征选择模块，利用动态特征提取器中动态卷积核实现输入特征自适应处理，利用门控特征选择模块产生的门控信号控制信息流，增强关键信息的提取与保留能力，完成图像的特征提取任务。

3）构建了跨层注意力融合模块，通过全局-局部空间注意力模块进行跨层特征提取和融合，融合局部与全局信息，避免了高频信息丢失。

4）在 MSRS 和 RoadScene 以及 Harvard 数据集上的实验结果表明，本文方法在 EN、MI、SSIM、PSNR、SF、$Q^{AB/F}$ 和 CC 等客观评价指标上，相较于其他 9 种高水平方法平均提高了 5.11%、52.84%、16.77%、13.90%、8.58%、29.66%、15.93%，证明本文算法不仅保留了丰富的纹理细节信息以及完整的解剖结构信息，同时还拥有高清晰度，展现出较强的泛化能力，能够很好地处理各种光照环境场景和医学影像诊断场景下的图像融合任务。

参考文献

[1] ZHOU M, ZHANG Y, XU X, et al. Edge-enhanced dilated residual attention network for multimodal medical image fusion[C]//2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Lisbon, Portugal: IEEE, 2024: 4108. DOI: 10.1109/BIBM62325.2024.10821967

[2] ZHU Y, XIAO M, ROBBINS D, et al. Walking representation and simulation based on multi-source image fusion and multi-agent reinforcement learning for gait rehabilitation[J]. Artificial Intelligence in Medicine, 2024, 156: 102945. DOI: 10.1016/j.artmed.2024.102945

[3] TANG L, XIANG X, ZHANG H, et al. DIVFusion: Darkness-free infrared and visible image fusion[J]. Information Fusion, 2023, 91: 477. DOI: 10.1016/j.inffus.2022.10.034

[4] ZHANG W, LU Y, ZHENG H, et al. MBRARN: Multibranch residual attention reconstruction network for medical image fusion[J]. Medical & Biological Engineering & Computing, 2023, 61(11): 3067. DOI: 10.1007/s11517-023-02902-2

[5] TIAN J, SUN D, GAO Q, et al. A novel infrared and visible image fusion algorithm based on global information-enhanced attention network[J]. Image and Vision Computing, 2024, 149: 105161. DOI: 10.1016/j.imavis.2024.105161

[6] SHI Y, SHI C, WENG Z, et al. CrossFuse: Learning infrared and

visible image fusion by cross-sensor top-k vision alignment and beyond[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35(8): 7579. DOI:10.1109/TCSVT.2025.3544746

[7]MA W, WANG K, LI J, et al. Infrared and visible image fusion technology and application: A review [J]. Sensors (Basel, Switzerland), 2023, 23(2): 599. DOI:10.3390/s23020599

[8]王瑾春, 马萍, 张宏立, 等. 基于语义驱动的红外与可见光图像交互融合[J]. 哈尔滨工业大学学报, 2025, 57(9): 56
WANG Jinchun, MA Ping, ZHANG Hongli, et al. Semantic-driven interactive fusion of infrared and visible light images[J]. Journal of Harbin Institute of Technology, 2025, 57(9): 56. DOI:10.11918/202406056

[9]MERGIN A A, PREMI M S G. Convolutional neural networks (CNN) with quantum-behaved particle swarm optimization (QPSO)-based medical image fusion[J]. International Journal of Image and Graphics, 2024, 24(5): 2340005. DOI:10.1142/S0219467823400053

[10]ZHOU Y, YANG X, LIU S, et al. Multimodal medical image fusion network based on target information enhancement[J]. IEEE Access, 2024, 12: 70851. DOI:10.1109/ACCESS.2024.3402965

[11]RAM PRABHAKAR K, SAI SRIKAR V, VENKATE BABU R. DeepFuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs [C]//2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 4724. DOI: 10.1109/ICCV.2017.505

[12]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. DOI:10.1109/JAS.2022.105686

[13]XI X, JIN X, JIANG Q, et al. EMA-GAN: A generative adversarial network for infrared and visible image fusion with multiscale attention network and expectation maximization algorithm [J]. Advanced Intelligent Systems, 2023, 5(11): 17. DOI: 10.1002/aisy.202300310

[14]ZHAO Z, BAI H, ZHU Y, et al. DDFM: Denoising diffusion model for multi-modality image fusion [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE Computer Society, 2023: 8082. DOI: 10.1109/ICCV51070.2023.00742

[15]YUE J, FANG L, XIA S, et al. Dif-fusion: Toward high color fidelity in infrared and visible image fusion with diffusion models[J]. IEEE Transactions on Image Processing, 2023, 32: 5705

[16]YI X, TANG L, ZHANG H, et al. Diff-IF: Multi-modality image fusion via diffusion model with fusion knowledge prior [J]. Information Fusion, 2024, 110: 102450. DOI:10.1016/j.inffus.2024.102450

[17]ZHANG H, CAO L, MA J. Text-DiFuse: An interactive multi-modal image fusion framework based on text-modulated diffusion model[J]. Advances in Neural Information Processing Systems, 2024, 37: 39552

[18]GAO X, YANG S, LIU J. PTDiffusion: Free lunch for generating optical illusion hidden pictures with phase-transferred diffusion model [C]//Proceedings of the Computer Vision and Pattern Recognition Conference. Nashville, TN, USA: Computer Vision Foundation, 2025: 18240

[19]WEI X, GUO W, YU W, et al. OSDM-MReg: Multimodal image registration based one step diffusion model [J]. arXiv: 2504.06027. DOI: 10.48550/arXiv.2504.06027

[20]TANG L, YUAN J, MA J. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network[J]. Information Fusion, 2022, 82: 28. DOI:10.1016/j.inffus.2021.12.004

[21]CHENG C, XU T, WU X-J. MUFusion: A general unsupervised image fusion network based on memory unit [J]. Information Fusion, 2023, 92: 80. DOI:10.1016/j.inffus.2022.11.010

[22]XU H, MA J, JIANG J, et al. U2Fusion: A unified unsupervised image fusion network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44(1): 502. DOI:10.1109/TPAMI.2020.3012548

[23]ZHAO W, XIE S, ZHAO F, et al. Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada: IEEE, 2023: 13955. DOI:10.1109/CVPR52729.2023.01341

[24]WANG W, DENG L J, VIVONE G. A general image fusion framework using multi-task semi-supervised learning[J]. Information Fusion, 2024, 108: 102414. DOI:10.1016/j.inffus.2024.102414

[25]QIAN Y, LIU G, TANG M C R. BTSFusion: Fusion of infrared and visible image via a mechanism of balancing texture and salience[J]. Optics and Lasers in Engineering, 2024, 173: 107925. DOI:10.1016/j.optlaseng.2023.107925

[26]LIU X, HUO H, LI J, et al. A semantic-driven coupled network for infrared and visible image fusion[J]. Information Fusion, 2024, 108: 102352. DOI: 10.1016/j.inffus.2024.102352

[27]LEI J, LI J, LIU J, et al. MLFuse: Multi-scenario feature joint learning for multi-modality image fusion[J]. IEEE Transactions on Multimedia, 2025, 27: 3880. DOI: 10.1109/TMM.2025.3535355

[28]VENKATESAN B, RAGUPATHY U S. An investigation on multimodal brain image fusion in the time-frequency domain using wavelet transforms[J]. IETE Journal of Research, 2024, 70(6): 11. DOI:10.1080/03772063.2023.2280670

[29]LIU J, WU G, LIU Z, et al. Infrared and visible image fusion: From data compatibility to task adaption[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 47(4): 2349. DOI: 10.1109/TPAMI.2024.3521416

[30]XU S, ZHAO Z, BAI H, et al. Hipandas: Hyperspectral image joint denoising and super-resolution by image fusion with the panchromatic image [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Honolulu, HI, USA: IEEE, 2025: 12002. DOI:10.1109/CVPR52729.2023.01341

[31]DU S, ZOU Y, WANG Z, et al. Unsupervised hyperspectral and multispectral image fusion via self-supervised modality decoupling[EB/OL]. (2024-12-06)[2026-04-29]. arXiv:2412.04802. DOI: 10.48550/arXiv.2412.04802

[32]ZHONG Y, HE J, LIANG Z, et al. Medical image fusion for high-level analysis: A mutual enhancement framework for unaligned PAT and MRI[EB/OL]. (2024-07-04)[2026-04-29]. arXiv:2407.03992. DOI: 10.48550/arXiv.2407.03992

[33]CHENG C, XU T, WU X J, et al. TextFusion: Unveiling the power of textual semantics for controllable image fusion [J]. Information Fusion, 2025, 117: 102790

[34]JIANG C, LIU X, ZHENG B, et al. HSFusion: A high-level vision task-driven infrared and visible image fusion network via semantic and geometric domain transformation [EB/OL]. (2024-07-13)[2026-04-29]. arXiv:2407.10047. DOI:10.48550/arXiv.2407.10047

[35]LIU J, FAN X, HUANG Z, et al. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE Computer Society, 2022: 5802. DOI: 10.1109/CVPR52688.2022.00571

(编辑 吕雪梅)