

引用格式: LI Rong, CHANG Xia, MA Ning. A Pan-sharpening Method Based on Spectral Swin Transformer and Frequency Enhancement[J]. Acta Photonica Sinica, 2027, 56(1):0110001

李荣,常霞,马宁. 基于光谱 Swin Transformer 与频域增强的全色锐化方法[J]. 光子学报, 2027, 56(1):0110001

基于光谱 Swin Transformer 与频域增强的全色锐化方法

李荣^{1,2}, 常霞^{1,2}, 马宁^{2,3}

(1 北方民族大学 数学与信息科学学院, 银川 750021)

(2 宁夏智能信息与大数据处理重点实验室, 银川 750021)

(3 北方民族大学 计算机科学与工程学院, 银川 750021)

摘 要: 针对全色锐化过程中空间细节恢复与光谱信息保持难以兼顾的问题, 提出一种基于光谱 Swin Transformer 与频域增强的全色锐化方法。该方法以改进的双分支 U 型网络为骨干结构, 实现了全色图像与多光谱图像特征的差异化提取与多尺度建模。所设计的光谱 Swin Transformer 模块分别从空间上下文关系和光谱相关通道关系两个方面增强空间-光谱联合表征能力; 同时, 在重建阶段设计了频域增强模块, 采用自适应频率分离与加权融合策略, 实现高频细节与低频结构信息的协同优化, 从而提升细节恢复能力并减少光谱失真现象。基于 WorldView-2、WorldView-3、GeoEye-1 和 Pléiades-1 四个公开数据集开展实验, 结果表明, 所提方法在多项定量评价指标以及视觉质量方面均取得较好表现, 能够有效提升空间结构恢复能力并保持光谱一致性。

关键词: 遥感图像; 全色锐化; 深度学习; 多光谱图像; Swin Transformer; 频域增强

中图分类号: TP751

文献标识码: A

doi: 10.3788/gzxb20275601.0110001

0 引言

受成像条件限制, 遥感卫星难以同时获取高空间分辨率和丰富光谱信息。因此, 高分辨率遥感卫星通常采用全色图像(Panchromatic Image, PAN)与多光谱图像(Multispectral Image, MS)协同成像模式, 获取具有空间细节信息和光谱信息互补特征的遥感数据。全色锐化(Pan-sharpening)通过融合 PAN 图像的空间细节信息与 MS 图像的光谱信息, 实现低分辨率多光谱图像空间增强, 生成兼具空间细节和光谱保真度的高分辨率多光谱图像(High-resolution Multispectral Image, HRMS)。该技术在资源勘探^[1]、军事侦察^[2]和城市规划^[3]等领域具有重要应用价值。

传统全色锐化方法主要包括分量替代(Component Substitution, CS)方法^[4]、多分辨率分析(Multiresolution Analysis, MRA)方法^[5]以及变分优化(Variational Optimization, VO)方法^[6]。CS 方法能够增强空间细节但易造成光谱失真; MRA 方法依赖分解策略; VO 方法具有较好的优化能力, 但通常依赖迭代求解, 计算复杂度较高。因此, 传统方法难以兼顾空间细节恢复和光谱保持。

随着深度学习的发展, 研究者将卷积神经网络(Convolutional Neural Network, CNN)引入全色锐化任务, 通过端到端学习空间与光谱信息映射关系提升融合质量。MASI 等^[7]提出 PNN 模型, 验证了 CNN 在全色锐化中的有效性, 但其特征表达能力受网络结构限制。随后, ADEEL 等^[8]通过金字塔残差增强细节但光谱保持不足; LI 等^[9]利用交叉注意力提升交互但纹理恢复有限; GUO 等^[10]提出 DMF-Net 增强特征表达但

基金项目: 宁夏自然科学基金(2025AAC030002), 宁夏高等教育一流学科建设基金(NXYLXK2017B09), 北方民族大学研究生创新项目(CYX25132)

第一作者: 李荣(2001—), 女, 硕士研究生, 研究方向为图像融合。Email: lirong1232025@163.com

通讯作者: 常霞(1982—), 女, 博士, 副教授, 研究方向为计算智能、图像处理与理解。Email: changxia0104@163.com

收稿日期: 2026-06-05; **录用日期:** 2026-08-06

<http://www.photon.ac.cn>

易过度锐化。

Transformer凭借全局建模能力逐渐应用于全色锐化领域。ZHOU等^[11]提出PanFormer模型,通过注意力机制实现空间与光谱特征交互融合,后续研究围绕双域特征交互和注意力机制展开,提升了融合性能^[12-14]。但现有方法仍面临以下局限: PAN与MS图像在空间分辨率、通道维度及信息属性上存在差异,难以平衡空间细节增强与光谱保持;部分Transformer方法侧重空间上下文建模,对光谱特征通道关联关系挖掘不足;同时,高低频信息的显式分离与自适应调控能力仍需提升。因此,实现异构信息有效表征、空间-光谱协同建模以及高低频特征平衡恢复,是提升融合性能的关键。

针对上述问题,本文提出一种基于光谱Swin Transformer与频域增强的全色锐化方法。该方法构建改进的U型骨干网络(Enhanced U-Net Backbone, EUB),采用双分支编码结构分别提取PAN图像与MS图像的异构特征;设计光谱Swin Transformer模块(Spectral Swin Transformer Module, SSM),通过空间注意力与光谱关系建模分支,增强空间-光谱联合特征表达能力;进一步引入频域增强模块(Frequency Enhancement Module, FEM),利用高低频信息自适应分离与融合机制,实现空间细节恢复与光谱保持之间的有效平衡。

1 相关原理

1.1 基于U-Net的全色锐化

随着深度学习在遥感图像处理领域的快速发展,U型网络(U-Net)凭借其编码器-解码器结构及跳跃连接机制,在全色锐化任务中得到广泛应用^[15]。U-Net通过编码器逐层降低特征图空间尺寸并提取深层语义信息,再通过解码器逐步恢复空间分辨率,同时利用跳跃连接将浅层细节特征传递至解码阶段,从而实现多尺度特征重建。其映射过程可表示为:

$$\hat{I} = \mathcal{F}_\theta(I) \quad (1)$$

其中, I 表示输入图像, $\hat{I}$ 表示网络输出结果, $\mathcal{F}_\theta$ 表示由参数 θ 控制的非线性映射函数。

然而,将U-Net等单流架构直接应用于全色锐化任务仍存在一定局限。全色锐化的目标是融合PAN图像与MS图像,生成HRMS图像。其过程可表示为:

$$\hat{I}_{\text{HRMS}} = \mathcal{F}_\theta(I_{\text{MS}}, I_{\text{PAN}}) \quad (2)$$

其中, I_{MS} 表示低分辨率多光谱图像, I_{PAN} 表示高分辨率全色图像, $\hat{I}_{\text{HRMS}}$ 表示重建得到的高分辨率多光谱图像。由于MS图像主要包含丰富的光谱信息,而PAN图像主要提供边缘、纹理和空间结构等高频细节信息,二者在空间分辨率、通道维度和信息属性上存在显著差异。

在单流U-Net结构中,常见做法是先将MS图像上采样至与PAN图像相同的空间尺寸,然后将二者直接拼接作为网络输入,即:

$$I_{\text{in}} = \text{Concat}(I_{\text{MS}}^\uparrow, I_{\text{PAN}}) \quad (3)$$

其中, $I_{\text{MS}}^\uparrow$ 表示经过上采样后的多光谱图像。该方式虽然实现了输入维度上的统一,但实际上隐含了不同模态数据可以通过同一编码路径进行统一表征的假设,因而难以同时兼顾两类信息的差异性,容易造成PAN图像中的空间细节提取不足,或MS图像中的光谱结构受到干扰。

从信号处理角度看,全色锐化本质上是在保持MS图像低频光谱信息的基础上,引入PAN图像中的高频空间细节,其基本形式可表示为:

$$\hat{I}_{\text{HRMS}} = I_{\text{MS}}^\uparrow + \Delta I_{\text{PAN}} \quad (4)$$

$$\Delta I_{\text{PAN}} = \mathcal{G}_\theta(I_{\text{PAN}}) \quad (5)$$

其中, ΔI_{PAN} 表示由PAN图像提取的空间细节项, $\mathcal{G}_\theta(\cdot)$ 表示从PAN图像中提取空间细节的映射函数。

一般而言,图像可以进一步分解为低频分量与高频分量。在全色锐化任务中,低频部分主要用于保持MS图像的光谱结构信息,而高频部分则主要来源于PAN图像中的边缘与纹理等细节信息,其表达形式如下:

$$\hat{I} = I_{\text{low}} + I_{\text{high}} \quad (6)$$

其中, I_{low} 主要反映整体结构与光谱分布, I_{high} 主要对应边缘、纹理等局部细节。

传统单流 U-Net 通常未对高低频成分进行显式建模,而是依赖卷积特征的隐式学习完成信息融合,容易导致细节恢复不足或光谱失真。因此,针对 PAN 图像与 MS 图像的异构特性,构建具有多分支特征提取与高、低频建模能力的网络结构,是提升全色锐化性能的重要研究方向。

1.2 基于 Transformer 的全色锐化

Transformer 架构自 2017 年提出以来,因其良好的长距离依赖建模能力,已广泛应用于自然语言处理和计算机视觉等任务^[16]。其核心在于自注意力机制,通过计算查询(Query)、键(Key)和值(Value)之间的相关性,实现对全局上下文信息的自适应建模,其表达形式为:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (7)$$

其中, Q, K, V 分别代表查询、键和值矩阵, d_k 是缩放因子。

随着 Transformer 在视觉任务中的应用不断拓展, ZHOU 等^[11]将 Transformer 引入全色锐化领域,提出了 PanFormer 模型。该方法分别对 PAN 图像与 MS 图像的补丁序列进行建模,利用自注意力提取模态特征,并通过交叉注意力实现多模态信息融合,其表达为:

$$\text{CA}(X_{\text{PAN}}, X_{\text{MS}}) = \text{Attention}(Q(X_{\text{MS}}), K(X_{\text{PAN}}), V(X_{\text{PAN}})) \quad (8)$$

其中, X_{PAN} 和 X_{MS} 分别代表 PAN 和 MS 特征图, CA 表示交叉注意力(Cross Attention, CA)操作。该方法能够有效挖掘不同模态之间的互补信息,从而提升特征融合质量。然而,标准 Transformer 在处理高分辨率遥感图像时常面临计算复杂度较高的问题。为此, LIU 等^[17]提出 Swin Transformer,通过引入移位窗口机制,将全局自注意力转化为局部窗口注意力计算,其表达形式为:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V \quad (9)$$

其中, B 表示相对位置编码。Swin Transformer 通过在不同层之间移动窗口位置,实现局部信息与跨窗口信息的有效交互,在降低计算复杂度的同时增强了特征的层次化表达能力。此外, XING 等^[18]通过频率混合注意力建模频率特征,提升了全色锐化中的空间细节恢复与光谱保持能力。

尽管如此,现有基于 Transformer 的全色锐化方法仍存在不足,多数方法侧重空间注意力建模,对多光谱特性的利用不充分;局部窗口注意力虽然降低了计算复杂度,但限制了跨区域全局依赖关系建模。因此,如何兼顾计算效率与空间-光谱协同建模能力,仍是当前全色锐化研究的重要问题。

2 本文方法

2.1 整体架构

本文提出一种端到端的全色锐化深度学习框架,旨在充分利用空间、光谱和频域特征实现高质量的遥感图像融合。网络的整体架构如图 1 所示,主要由三个核心组件构成:改进的 U 型骨干网络(EUB)、光谱 Swin Transformer 模块(SSM)以及频域增强模块(FEM)。其中, EUB 用于提取 PAN 图像与 MS 图像的多尺度特征; SSM 用于增强空间上下文信息与光谱特征关联关系的表达能力; FEM 则通过对频率成分的自适应调控,实现高频细节增强与低频结构保持之间的平衡。

从整体流程来看,所提方法主要包括特征提取、特征增强与图像重建三个阶段。在特征提取阶段,首先采用双三次插值将 MS 图像上采样至与 PAN 图像一致的空间分辨率,随后,将上采样后的 MS 图像与 PAN 图像分别输入 EUB 的两个独立编码分支,提取由浅层到深层的多尺度特征表示,以充分保留空间细节信息与光谱信息。在特征增强阶段,将编码器输出的各层特征输入 SSM 模块,通过空间注意力分支和光谱关系建模分支来增强空间上下文信息以及光谱相关特征通道之间的联系。在图像重建阶段,将增强后的特征输入解码器,通过多尺度特征融合与逐层上采样恢复空间分辨率,并在重建过程中逐步整合不同层级的语义信息。最后,将解码得到的重建特征输入 FEM 模块,通过对不同频率成分进行自适应调节,进一步增强空间细节并保持低频光谱结构一致性,得到最终融合结果 HRMS 图像。

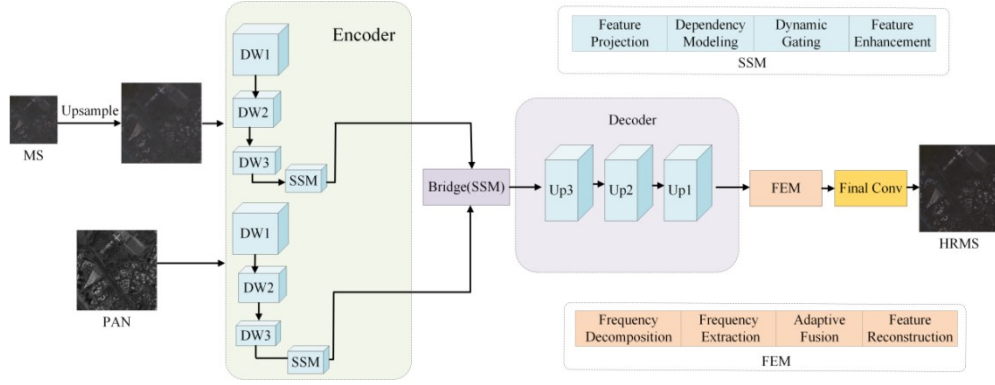


图1 所提方法整体架构图

Fig.1 Overall architecture diagram of the proposed method

2.2 改进的U型网络骨干

传统U-Net在图像分割等任务中表现出良好的多尺度特征提取能力,但单一编码路径难以同时刻画PAN图像与MS图像的异构特征。同时,普通局部卷积主要侧重邻域特征提取,缺少对较大范围空间上下文及关键通道响应的专门建模机制。针对上述问题,本文构建了改进的U型网络骨干(EUB),其整体结构如图2所示。

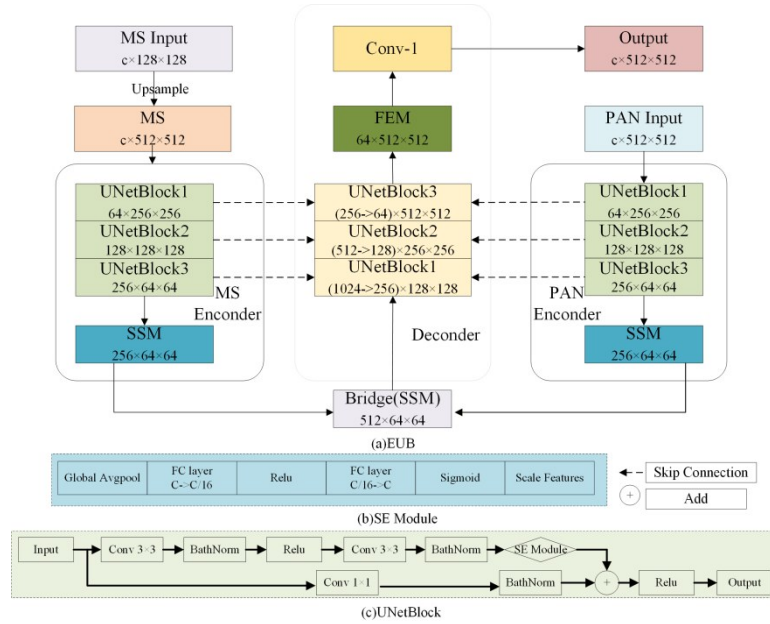


图2 改进的U-Net骨干网络架构(EUB)

Fig.2 Improved U-Net backbone network architecture (EUB)

在编码阶段,设计双分支结构,分别对PAN图像与MS图像进行特征提取,第 i 层特征表示为:

$$F_{MS}^i = E_{MS}(F_{MS}^{i-1}) \quad (10)$$

$$F_{PAN}^i = E_{PAN}(F_{PAN}^{i-1}) \quad (11)$$

其中, E_{MS} 和 E_{PAN} 分别表示MS图像和PAN图像的编码器, F_{MS}^i 和 F_{PAN}^i 为对应分支在第 i 层提取的特征。通过双分支设计,网络能够针对不同模态执行特定特征提取,从而更有效地保留MS图像的光谱信息与PAN图像的空间细节。

在深层特征提取阶段,采用光谱Swin Transformer模块(SSM)对编码器输出特征进行增强,其表达为:

$$F_{enh} = SSM(F_{deep}) \quad (12)$$

其中, F_{deep} 表示深层编码器特征, F_{enh} 为经SSM处理后的增强特征。

在特征提取过程中,本文在 UNetBlock 中引入通道注意力机制(Squeeze-and-Excitation, SE)^[19],并结合残差连接以增强关键特征表达能力,其形式为:

$$F_{\text{output}} = \text{UNetBlock}(F_{\text{input}}) = \text{SE}(f_{\text{conv}}(F_{\text{input}})) + f_{\text{res}}(F_{\text{input}}) \quad (13)$$

其中, F_{input} 是输入特征, F_{output} 是输出特征, $f_{\text{conv}(\cdot)}$ 表示卷积操作, $f_{\text{res}}(\cdot)$ 是残差连接。UNetBlock 表示 U-Net 网络的基本构建模块, SE 模块定义为:

$$\text{SE}(F) = F \cdot \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot \text{GAP}(F))) \quad (14)$$

其中, GAP 表示全局平均池化, W_1 和 W_2 为可学习参数, σ 表示 sigmoid 激活函数。SE 模块首先利用全局平均池化提取各通道的全局响应,再生成相应的通道权重,对原始特征进行重新加权。该过程主要依据全局统计信息调整已有通道响应的强弱,其作用属于通道重要性的自适应重标定,并不直接生成新的跨通道组合特征。

在解码阶段,采用多路径特征融合策略,将编码阶段提取的 MS 特征、PAN 特征与上采样特征进行融合,其表达为:

$$F_i^{\text{dec}} = \text{UNetBlock}(\text{Cat}[F_i^{\text{up}}, F_i^{\text{MS}}, F_i^{\text{PAN}}]) \quad (15)$$

$$F_i^{\text{up}} = \text{UpConv}(F_{i+1}^{\text{dec}}) \quad (16)$$

其中, F_i^{dec} 表示解码器第 i 层输出的特征, Cat 表示拼接操作, F_i^{up} 表示通过上采样得到的第 i 层特征, F_{i+1}^{dec} 表示解码器在第 $i+1$ 层输出的特征。通过多路径融合,能够有效整合不同尺度与不同模态的信息。

EUB 通过双分支编码结构、通道注意力增强以及多路径特征融合策略,实现 PAN 空间信息与 MS 光谱信息的差异化提取和多尺度融合。其中, SE 模块主要用于编码阶段的关键通道增强,为后续 SSM 模块进一步开展空间上下文与光谱关系联合建模提供更具判别性的特征表示。

2.3 光谱 Swin Transformer 模块

尽管改进的 U 型网络骨干(EUB)在多尺度特征提取方面具有较强能力,但受限于卷积操作的局部感受野,对较大范围空间上下文关系及深层光谱相关特征通道之间的信息交互建模能力有限。为进一步增强空间结构信息与光谱相关特征之间的联合表达能力,本文提出光谱 Swin Transformer 模块(Spectral Swin Transformer Module, SSM)。

SSM 基于 Swin Transformer 窗口注意力机制构建空间关系建模分支,并进一步引入光谱关系建模分支,对深层光谱相关特征通道之间的关联关系进行学习。由于网络中的特征通道由深层特征学习得到,并不严格对应输入图像中的物理光谱波段,因此本文主要建模深层特征空间中的光谱相关通道依赖关系。该模块通过分别建模空间位置依赖关系和通道间关联关系,实现空间-光谱互补特征表示,其过程表示为:

$$F_{\text{SSM}} = \text{SSM}(F_{\text{deep}}) \quad (17)$$

其中, F_{deep} 表示由编码器提取的深层特征, F_{SSM} 表示经 SSM 增强后的特征表示。

不同于标准 Swin Transformer 主要侧重空间维度建模,本文提出的 SSM 模块构建了空间-光谱双分支结构,空间注意力分支用于建模空间位置之间的上下文依赖,而光谱关系建模分支则在保持空间位置对应关系的基础上,通过跨通道映射学习光谱相关特征通道之间的信息交互关系。两个分支分别增强空间结构信息和通道关联信息,并通过融合形成联合表示:

$$X_{\text{spatial}} = W - \text{MSA}(X) \quad (18)$$

$$X_{\text{spectral}} = \text{SpectralRel}(X) \quad (19)$$

$$X_{\text{fused}} = \text{Fusion}(X_{\text{spatial}}, X_{\text{spectral}}) \quad (20)$$

其中, X 为输入特征, X_{spatial} 和 X_{spectral} 分别表示空间分支和光谱分支的输出特征。 $W - \text{MSA}(\cdot)$ 表示窗口多头自注意力操作; $\text{SpectralRel}(\cdot)$ 表示光谱关系建模分支; $\text{Fusion}(\cdot)$ 表示特征融合操作。

在空间注意力分支中,本文采用移位窗口多头自注意力机制对空间上下文信息进行建模。具体而言,将输入特征 $X \in \mathbb{R}^{B \times C \times H \times W}$ 划分为 8×8 的非重叠窗口,在每个窗口内执行自注意力计算,为实现跨窗口信息交互,本文在相邻层中进一步引入移位窗口策略,在连续层中交替使用两种窗口划分方式。该设计不仅能够增强局部区域内的特征表达能力,而且能够通过窗口偏移实现不同区域之间的信息传递。此外,本文将注意力头数设为 4,以便模型从多个表示子空间学习特征关系,从而提升对复杂空间结构的建模能力。

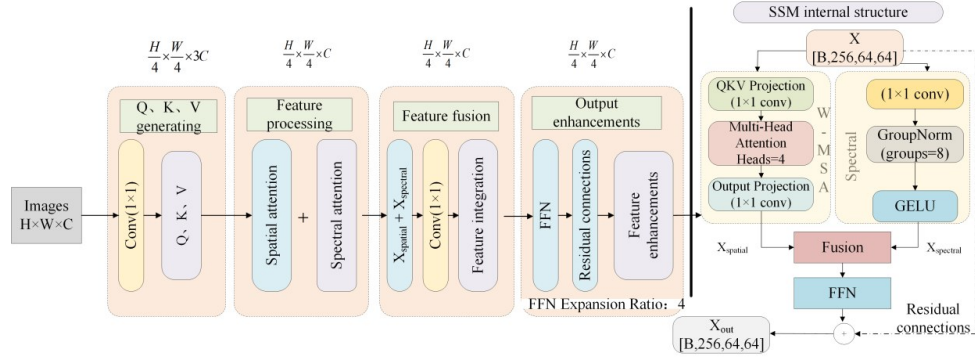


图3 SSM模块结构图

Fig.3 Structure diagram of the SSM module

在光谱关系建模分支中,本文采用轻量化跨通道映射结构,在保持空间位置对应关系不变的情况下,通过跨通道映射学习深层光谱相关特征通道之间的信息交互关系。该分支利用 1×1 卷积对每个空间位置处的通道特征向量进行共享映射:

$$\text{SpectralRel}(X) = \text{GELU}\left(\text{GN}\left(\text{Conv}_{1 \times 1}(X)\right)\right) \quad (21)$$

其中, $\text{Conv}_{1 \times 1}$ 是点卷积, GN 是归一化, GELU 是激活函数。对于任意空间位置 (h, w) ,其对应的通道特征向量记为 $x_{h,w} \in \mathbb{R}^C$,则映射过程表示为:

$$Z_{h,w} = W_s x_{h,w} + b_s \quad (22)$$

其中, W_s 为可学习的跨通道映射矩阵。该矩阵刻画输入通道与输出通道之间的映射关系,使多个输入通道的信息能够参与目标通道特征的构建。对于第 i 个输出通道,其具体计算过程可表示为:

$$z_{h,w}^i = \sum_{j=1}^C w_{ij} x_{h,w}^j + b_i \quad (23)$$

其中, $x_{h,w}^j$ 表示空间位置 (h, w) 处第 j 个输入特征通道, $z_{h,w}^i$ 表示对应位置处第 i 个输出特征通道, w_{ij} 表示输入通道 j 到输出通道 i 之间的可学习映射关系。由式(23)可知,每个输出特征通道均可以由多个输入特征通道的加权组合生成。由于 1×1 卷积权重矩阵中的非对角元素允许不同输入通道参与目标输出通道的计算,因此输出通道并非由单一输入通道决定,而是由多个输入通道共同映射得到,从而实现通道间的信息交互。该可学习映射过程参数化描述了不同特征通道之间的信息组合关系,使网络能够学习深层特征空间中的光谱相关通道关联表示。

与第2.2节所述SE通道注意力机制不同,光谱关系建模分支不通过全局平均池化聚合空间信息,也不生成通道权重对特征响应进行重新标定。SE模块利用全局统计信息调整通道响应强度,而本文分支通过逐位置跨通道映射学习不同特征通道之间的信息关联,实现通道间特征交互与重组。因此,二者分别侧重通道响应调整和通道关系建模。在获得空间分支和光谱分支的输出后,本文采用特征拼接与卷积映射相结合的方式融合,以实现空间信息与光谱信息的协同增强,其表达式为:

$$X_{\text{fused}} = \text{Conv}_{1 \times 1}(\text{Concat}[X_{\text{spatial}}, X_{\text{spectral}}]) \quad (24)$$

$$X_{\text{out}} = \text{FFN}(X_{\text{fused}}) + X \quad (25)$$

其中, $\text{Concat}[\cdot]$ 表示特征拼接操作, $\text{FFN}(\cdot)$ 表示由两层卷积组成的前馈网络,残差连接用于保留原始特征信息并稳定训练过程。

综上,SSM分别从空间位置关系和通道关联关系两个方面增强特征表达,并通过融合操作获得空间-光谱联合特征,相较于Swin Transformer的空间窗口建模和SE模块的通道重标定,SSM进一步实现了逐位置跨通道信息交互与特征重组,为后续频域优化和图像重建提供了更充分的空间-光谱联合特征表示。

2.4 频域增强模块

尽管SSM通过空间-光谱双分支机制增强了特征表达能力,但其主要在空间位置关系和深层特征通道关系上进行建模,缺乏对图像空间频率特征的显式描述。对于遥感图像而言,边缘与纹理等细节信息主要

集中于高频分量,而主体结构及整体光谱分布则主要体现在低频分量中。在多尺度特征提取与重建过程中,高频信息容易发生衰减或丢失,从而影响全色锐化结果的细节恢复能力。为此,本文提出频域增强模块(FEM),通过自适应频率分离与重构机制,实现高频细节与低频结构信息的协同增强,其结构如图4所示。

设输入特征为 F_{in} , 输出特征为 F_{out} , 则FEM的整体表达为:

$$F_{out} = F_{in} + \alpha \cdot R(F_{enhanced}) + \beta \cdot F_{enhanced} \quad (26)$$

其中, $F_{enhanced}$ 表示频域增强后的特征, $R(\cdot)$ 表示重构操作, $\alpha \cdot R(F_{enhanced})$ 是重构增强项, $\beta \cdot F_{enhanced}$ 是直接增强项, α 和 β 是可学习的平衡参数。

频率增强特征通过高频和低频分量的自适应融合得到:

$$F_{enhanced} = F_{high} \otimes W_{high} \oplus F_{low} \otimes W_{low} \quad (27)$$

其中, F_{high} 和 F_{low} 分别表示提取的高频特征和低频特征, W_{high} 和 W_{low} 表示对应的权重系数, $\otimes$ 表示逐元素乘法, $\oplus$ 表示逐元素加法。

为实现不同频率分量的动态平衡,本文由输入特征自适应生成高频权重和低频权重,其表达式为:

$$[W_{high}, W_{low}] = \text{Softmax}(\text{Conv}_{1 \times 1}(F_{in})) \quad (28)$$

其中, $\text{Conv}_{1 \times 1}$ 是 1×1 卷积操作,用于将输入特征映射为频率权重, Softmax 为激活函数,确保高频和低频权重之和为1,从而实现频率分量的动态平衡。

在频率分量提取方面,本文分别构建高通分支与低通分支以提取不同频率特征,其表达为:

$$F_{high} = \text{HighPass}(F_{in}) \quad (29)$$

$$F_{low} = \text{LowPass}(F_{in}) \quad (30)$$

其中, $\text{HighPass}(\cdot)$ 表示高通分支,用于提取边缘、纹理等高频细节信息; $\text{LowPass}(\cdot)$ 表示低通分支,用于提取平滑区域和主体结构等低频信息。具体而言,高频提取分支采用深度可分离卷积^[20]结构,以增强局部细节建模能力,其表达为:

$$\text{HighPass}(F) = \text{BN}(\text{Conv}1 \times 1(\text{GELU}(\text{BN}(\text{DWConv}3 \times 3(F)))) \quad (31)$$

其中, $\text{DWConv}_{3 \times 3}$ 表示 3×3 深度可分离卷积, BN 表示批量归一化, GELU 表示高斯误差线性单元激活函数。低频分支采用平均池化操作:

$$\text{LowPass}(F) = \text{GELU}(\text{BN}(\text{Conv}1 \times 1(\text{AvgPool}_{3 \times 3}(F)))) \quad (32)$$

其中, $\text{AvgPool}_{3 \times 3}$ 表示平均池化操作,用于抑制高频噪声并保留低频结构信息。

通过上述设计, FEM模块在特征层面对高频细节与低频结构进行显式分离与自适应融合,并通过可学习权重动态调节不同频率分量的贡献。相比传统基于固定滤波的频率处理方法,该模块具备更强的自适应能力,能够在增强空间细节恢复能力的同时有效保持光谱一致性,从而实现空间分辨率与光谱保真度之间

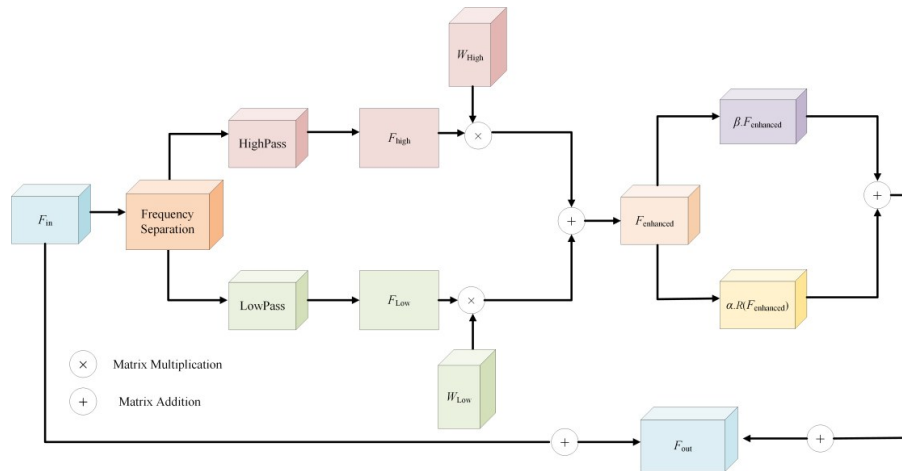


图4 FEM模块架构图

Fig.4 Architecture diagram of the FEM module

的协同优化。

2.5 多组件损失函数

传统全色锐化方法通常采用像素级 L1 损失对融合结果进行约束,但该类损失主要作用于空间域,难以同时兼顾频率特征恢复与光谱一致性约束。为此,本文构建多组件复合损失函数,从空间域、频域和光谱域三个维度对融合结果进行联合优化,以提升空间细节恢复能力与光谱保真度。

首先,在空间域采用像素级 L1 损失,其定义为:

$$L_1 = \|F_{\text{fused}} - F_{\text{gt}}\|_1 \quad (33)$$

其中, F_{fused} 表示网络生成的融合图像, F_{gt} 表示参考真值图像, $\|\cdot\|_1$ 表示 L_1 范数。该损失用于约束融合结果在像素层面的重建精度,能够有效保证整体结构与空间分布的一致性。

为进一步约束频率信息的重建质量,本文借鉴 YADAV 等^[21]的研究思路,采用多尺度频域损失以评估融合结果在不同尺度下的频率重建质量,其表达式为:

$$L_{DCT} = \sum_{s \in \{1, 2, 4\}} \|DCT(F_{\text{fused}}^s) - DCT(F_{\text{gt}}^s)\|_1 \quad (34)$$

其中, F_{fused}^s 和 F_{gt}^s 分别表示在尺度因子 s 下的融合图像与参考图像, $s \in \{1, 2, 4\}$ 表示评估的尺度因子, $DCT(\cdot)$ 为二维离散余弦变换算子。该损失从频域角度衡量融合结果与参考图像之间的差异,能够增强网络对高频细节信息的感知与重建能力。

同时,为保持融合结果的光谱一致性,本文采用基于余弦相似度的光谱保持损失,其形式为:

$$L_{\text{spec}} = 1 - \cos(I_{\text{fused}}^{\text{norm}}, I_{\text{gt}}^{\text{norm}}) \quad (35)$$

其中,归一化处理的融合结果 $I_{\text{fused}}^{\text{norm}}$ 与目标图像 $I_{\text{gt}}^{\text{norm}}$ 通过余弦相似度计算光谱特征的匹配程度。该损失专注于保持波段间相对关系,防止融合过程的光谱畸变。

最终,将上述各项损失进行加权组合,得到整体优化目标函数:

$$L_{\text{total}} = \lambda_1 L_1 + \lambda_2 L_{DCT} + \lambda_3 L_{\text{spec}} \quad (36)$$

经过实验验证,权重参数 λ_1 、 λ_2 、 λ_3 的取值分别为 0.4、0.3、0.3 可实现最佳平衡效果。通过上述多组件联合约束,网络能够同时关注像素重建精度、频率细节恢复和光谱保真度,从而实现全色锐化质量的整体提升。

3 实验与分析

3.1 实验数据和设置

为验证所提方法的有效性,本文在 4 个公开遥感全色锐化数据集上开展实验^[8],包括 8 波段的 WorldView-2(WV2)、WorldView-3(WV3)以及 4 波段的 GeoEye-1(GE1)、Pléiades-1(PL1)。所有降分辨率实验数据均按照遥感全色锐化领域广泛采用的 Wald 协议^[22]进行构建,即通过对原始多光谱图像进行退化处理生成相应的低分辨率多光谱图像,从而构建配准的训练样本对。随后,将构建得到的影像块按照 8:1:1 的比例随机划分为训练集、验证集和测试集^[23-24]。各子集划分结果在后续实验中保持固定,并统一用于所有方法的训练与测试,以保证不同方法在相同数据基础上的公平比较。实验中,每组样本均由 MS 图像、PAN 图像和参考 HRMS 图像组成。其中,MS 图像尺寸为 128×128 , PAN 图像与参考 HRMS 图像尺寸均为 512×512 。各数据集的训练集(Train)、验证集(Val)和测试集(Test)划分如表 1 所示。

所提方法基于 PyTorch 框架实现,并在 NVIDIA GeForce RTX 3090 GPU 上完成训练与测试。为保证

表 1 数据集划分
Table 1 Dataset split

Satellites	(Train/Val/Test)
WV2	190/32/32
WV3	133/22/22
GE1	221/48/48
PL1	168/36/36

不同数据集之间实验设置的一致性,统一采用相同训练策略:训练轮数为 100,批大小为 4,优化器为 Adam,初始学习率设为 0.001。所有对比方法均在相同的数据划分、预处理流程和训练条件下进行训练与测试,以保证实验比较的公平性。

3.2 对比方法与评价指标

为全面评估所提方法的性能,本文选取 10 种具有代表性的全色锐化方法进行对比,包括 4 种传统方法:(a)GS^[25]、(b)Brovey^[26]、(c)GFPCA^[27]和(d)SFIM^[28]以及 6 种基于深度学习的方法:(e)PNN^[7]、(f)Pyramid^[8]、(g)PanFormer^[11]、(h)CAB^[9]、(i)DMFNet^[10]、(j)CFNet^[18]。所有对比方法均严格按照其原始文献中的参数配置实现,并在相同硬件环境与训练条件下进行测试,以确保实验结果的可靠性与公平性。

为从空间保真、光谱保真和综合视觉质量等多个角度评价融合结果,本文采用 8 项常用评价指标^[29]进行定量分析:包括相关系数(correlation coefficient, CC)、光谱角映射(Spectral Angle Mapper, SAM)、相对全局维数合成误差(Relative Dimensionless Global Error in Synthesis, ERGAS)、全局质量指数(Global Quality Index, Q),其中 8 波段数据采用 Q8,4 波段数据采用 Q4、峰值信噪比(Peak Signal-to-Noise Ratio, PSNR),以及三种无参考评价指标:无参考质量指数(Quality with No Reference, QNR)、光谱失真指数(Spectral Distortion Index, D_λ)和空间失真指数(Spatial Distortion Index, D_s)。其中,CC、Q4、Q8、QNR 和 PSNR 越大表示融合效果越好,SAM、ERGAS、 D_λ 和 D_s 越小表示光谱与空间失真越低。

3.3 降分辨率实验结果与分析

表 2-表 5 给出了不同方法在 WV2、WV3、GE1 和 PL1 数据集上的定量比较结果,其中,加粗字体表示最优结果。从评价指标来看,本文提出方法在 CC、SAM 和 PSNR 等有参考指标上取得了较好的性能,并在多个数据集上达到最优或接近最优结果,说明所提方法能够有效提升融合结果与参考 HRMS 图像之间的空间结构一致性和光谱保持能力。对于无参考评价指标,本文方法在部分数据集上未取得最优结果,这主要是由于有参考指标和无参考指标的评价目标存在差异:有参考指标直接衡量融合结果与参考 HRMS 图像之间的重建误差和光谱偏差,而无参考指标主要评价融合结果与输入 PAN 图像及 MS 图像之间的光谱和空间统计一致性。本文方法通过空间-光谱联合建模增强空间细节恢复能力,并保持光谱相关特征表达,但该增强过程可能导致融合结果与输入观测影像之间的局部统计一致性存在一定差异,从而影响部分无参考指标表现。结合有参考指标、无参考指标以及视觉评价结果可以看出,本文方法能够在空间结构恢复、光谱信息保持和整体视觉质量之间取得较好的平衡,验证了所提方法的有效性。

图 5-图 8 分别展示了 WV2、WV3、GE1 和 PL1 数据集上的可视化结果。为更加直观地比较不同方法的融合效果,本文进一步计算了融合结果与真实参考图像(Ground Truth, GT)之间的平均绝对误差(Mean Absolute Error, MAE)。从残差图来看,其误差分布更加均匀,说明模型在整体重建过程中具有较好的稳定性与一致性。此外,由于原始遥感图像包含 4 或 8 个多光谱波段,本文统一采用 RGB 三通道真彩色合成方式进行可视化展示,以便直观比较不同方法的融合效果。

3.3.1 定量分析

基于表 2-表 5 的定量结果可以看出,所提方法在各项评价指标上整体表现较好。在空间保真度方面,所提方法在 WV2、WV3 和 GE1 数据集上均取得最高的 CC 和 Q 值,在更具挑战性的 PL1 数据集上,本文方法仍然保持接近最优的性能,说明其对复杂地物结构具有较好的适应能力。在光谱保真度方面,所提方法在大多数数据集上取得最低或接近最优的 SAM 和 ERGAS 值,有效抑制了融合过程中产生的光谱失真,保证了多光谱信息的一致性。在综合重建质量方面,所提方法在 PSNR 指标上,分别在 WV2、WV3、GE1 和 PL1 数据集上达到 35.4360 dB、35.9557 dB、38.1850 dB 和 40.4912 dB,均优于对比方法,进一步验证了其在整体重建精度方面的优势。在计算效率方面,所提方法的测试时间相对较长,但其性能提升更为显著,表明通过引入光谱关系建模分支与频域增强模块,在计算复杂度与融合质量之间取得了较好的权衡。综合不同数据集实验结果可以发现,本文方法在不同传感器条件下均保持较稳定的性能,在空间细节恢复和光谱保持之间取得了较好的平衡。

3.3.2 定性分析

图 5-图 8 分别展示了各方法在 WV2、WV3、GE1 和 PL1 数据集上的全色锐化结果及对应残差图。从整

表2 WV2数据集的实验结果
Table 2 Experimental results of the WV2 dataset

	CC	SAM	ERGAS	Q8	PSNR	QNR	D_λ	D_s	Test time/s
GS ^[25]	0.8951	0.2328	28.5763	0.6430	16.7582	0.8320	0.1334	0.2645	0.1549
Brovey ^[26]	0.8917	0.0907	4.4116	0.8417	32.3718	0.8446	0.1225	0.2645	0.1195
GFPCA ^[27]	0.7803	0.1148	6.4996	0.6890	29.4725	0.8847	0.1211	0.0889	0.2023
SFIM ^[28]	0.8886	0.0926	4.3213	0.8546	32.2776	0.7946	0.1714	0.3113	0.1848
PNN ^[7]	0.8752	0.0882	4.1096	0.8317	33.0411	0.9097	0.1034	0.0352	0.0011
Pyramid ^[8]	0.9014	0.0884	3.9489	0.8595	33.3990	0.7248	0.0338	0.2497	0.0069
PanFormer ^[11]	0.8376	0.0909	4.7007	0.7860	31.8650	0.8686	0.0802	0.3074	0.0140
CAB ^[9]	0.7612	0.0939	5.8466	0.6833	30.1935	0.9296	0.0777	0.0395	0.0014
DMFNet ^[10]	0.7756	0.0866	5.5051	0.7068	30.7604	0.9461	0.0646	0.0087	0.0087
CFNet ^[18]	0.8867	0.0863	3.9291	0.8390	33.5210	0.8040	0.1064	0.4631	0.0060
Ours	0.9255	0.0729	3.0956	0.9044	35.4360	0.8300	0.0715	0.4552	0.2050
Ideal value	1	0	0	1	$+\infty$	1	0	0	0

表3 WV3数据集的实验结果
Table 3 Experimental results for the WV3 dataset

	CC	SAM	ERGAS	Q8	PSNR	QNR	D_λ	D_s	Test time/s
GS ^[25]	0.9298	0.2302	25.8112	0.6933	18.2178	0.8985	0.0817	0.1664	0.1585
Brovey ^[26]	0.9167	0.1027	4.5826	0.8784	32.0048	0.8684	0.1162	0.1662	0.1186
GFPCA ^[27]	0.8703	0.1494	6.7451	0.7846	29.5279	0.8713	0.1348	0.0940	0.2039
SFIM ^[28]	0.9194	0.1028	4.6134	0.8781	32.2459	0.5821	0.0590	0.3804	0.1846
PNN ^[7]	0.9413	0.0981	3.7322	0.9018	34.3537	0.8699	0.0301	0.1092	0.0014
Pyramid ^[8]	0.9330	0.0986	4.1494	0.8825	33.5032	0.6516	0.0162	0.3368	0.0069
PanFormer ^[11]	0.9198	0.1011	4.4196	0.8651	32.8420	0.8927	0.0669	0.2513	0.0080
CAB ^[9]	0.8622	0.1240	6.4000	0.7476	29.6324	0.9445	0.0647	0.0174	0.0018
DMFNet ^[10]	0.9201	0.0892	4.7811	0.8322	32.6285	0.8617	0.0164	0.1243	0.0086
CFNet ^[18]	0.9413	0.0962	3.7613	0.8911	34.3972	0.8351	0.0804	0.3589	0.0020
Ours	0.9567	0.0827	3.1156	0.9325	35.9557	0.8745	0.0579	0.3475	0.2110
Ideal value	1	0	0	1	$+\infty$	1	0	0	0

表4 GE1数据集的实验结果
Table 4 Experimental results of the GE1 dataset

	CC	SAM	ERGAS	Q4	PSNR	QNR	D_λ	D_s	Test time/s
GS ^[25]	0.9147	0.3145	32.0081	0.6650	19.3441	0.7453	0.1392	0.5581	0.0260
Brovey ^[26]	0.9002	0.1245	5.9613	0.8774	33.5317	0.7891	0.0958	0.5299	0.0278
GFPCA ^[27]	0.7105	0.1738	9.9865	0.7075	29.4498	0.7468	0.2365	0.3119	0.0262
SFIM ^[28]	0.9128	0.1245	5.2552	0.8905	34.2883	0.7447	0.1618	0.5243	0.0211
PNN ^[7]	0.8863	0.0999	5.8553	0.8834	33.6624	0.8282	0.0788	0.4433	0.0009
Pyramid ^[8]	0.9194	0.0875	4.4751	0.9327	35.8118	0.8420	0.0683	0.4286	0.0066
PanFormer ^[11]	0.9083	0.0793	5.0263	0.9123	34.9168	0.8522	0.0627	0.4083	0.0140
CAB ^[9]	0.7386	0.1314	9.0404	0.7118	30.2343	0.9532	0.0522	0.0314	0.0014
DMFNet ^[10]	0.8099	0.1035	7.6762	0.7878	31.5897	0.9061	0.0729	0.1677	0.0085
CFNet ^[18]	0.9408	0.0773	3.8631	0.9456	37.2606	0.8196	0.0783	0.4736	0.0050
Ours	0.9445	0.0671	3.3553	0.9536	38.1850	0.8398	0.0665	0.4374	0.0630
Ideal value	1	0	0	1	$+\infty$	1	0	0	0

表 5 PL1 数据集的实验结果
Table 5 Experimental results for the PL1 dataset

	CC	SAM	ERGAS	Q4	PSNR	QNR	D_{λ}	D_s	Test time/s
GS ^[25]	0.9670	0.1954	27.3021	0.7080	17.9134	0.8262	0.0723	0.4705	0.0278
Brovey ^[26]	0.9625	0.0882	3.1265	0.9287	36.8145	0.8534	0.0479	0.4356	0.0270
GFPCA ^[27]	0.8733	0.1429	5.5483	0.8228	32.0156	0.7251	0.2449	0.3812	0.0262
SFIM ^[28]	0.9528	0.0885	3.1274	0.9277	36.5750	0.8262	0.0925	0.4218	0.0289
PNN ^[7]	0.9167	0.0903	3.1801	0.9217	36.6217	0.8736	0.0679	0.3218	0.0012
Pyramid ^[8]	0.9424	0.0844	2.9520	0.9277	37.1183	0.9016	0.0363	0.3020	0.0067
PanFormer ^[11]	0.9646	0.0656	2.1740	0.9576	39.6807	0.8864	0.0452	0.3404	0.0090
CAB ^[9]	0.8616	0.0986	4.9638	0.8059	33.0550	0.9344	0.0689	0.0500	0.0014
DMFNet ^[10]	0.9347	0.0641	3.3986	0.8901	35.9291	0.9532	0.0163	0.1587	0.0084
CFNet ^[18]	0.9704	0.0730	2.0978	0.9563	40.2569	0.8399	0.0783	0.4183	0.0020
Ours	0.9673	0.0676	1.9877	0.9583	40.4912	0.8825	0.0391	0.3705	0.0690
Ideal value	1	0	0	1	$+\infty$	1	0	0	0

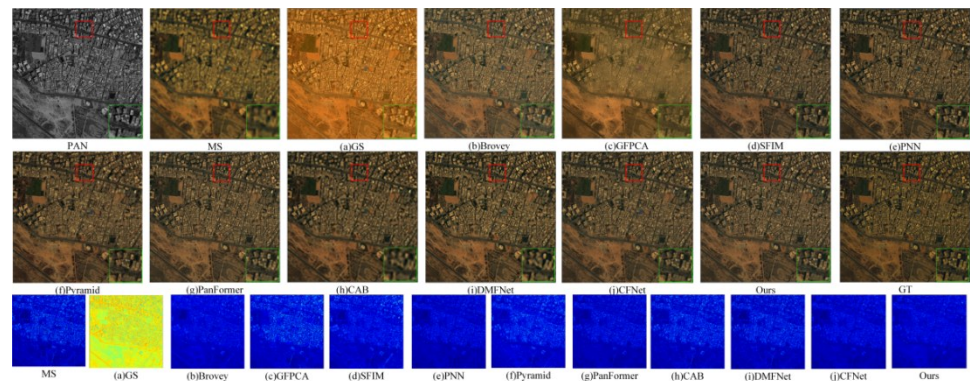


图 5 WV2 数据集不同方法融合结果及残差图
Fig.5 Fusion results and residual maps of different methods on WV2 dataset

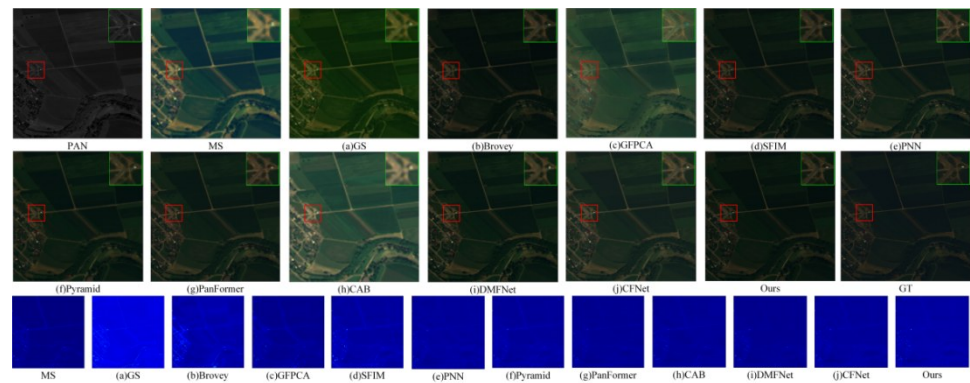


图 6 WV3 数据集不同方法融合结果及残差图
Fig.6 Fusion results and residual maps of different methods on WV3 dataset

体可视化结果来看,不同方法在各数据集上的表现趋势基本一致,但在细节恢复能力和光谱保持能力方面存在明显差异。传统方法如 Brovey 和 GS 在处理复杂地物区域时普遍存在较为明显的光谱失真问题,主要表现为颜色偏移和局部区域亮度不一致;而早期深度学习方法 PNN 和 Pyramid 虽然在一定程度上改善了光谱保真度,但由于对空间细节建模能力有限,仍存在结构纹理模糊、边缘不清晰等问题。近年来提出的 DMFNet 和 CFNet 方法在空间结构保持方面已有明显提升,但在复杂纹理区域仍存在细节不够锐利或局部信息丢失的现象。相比之下,本文方法在各数据集上均表现出更优的视觉效果。在 WV2 和 WV3 数据集

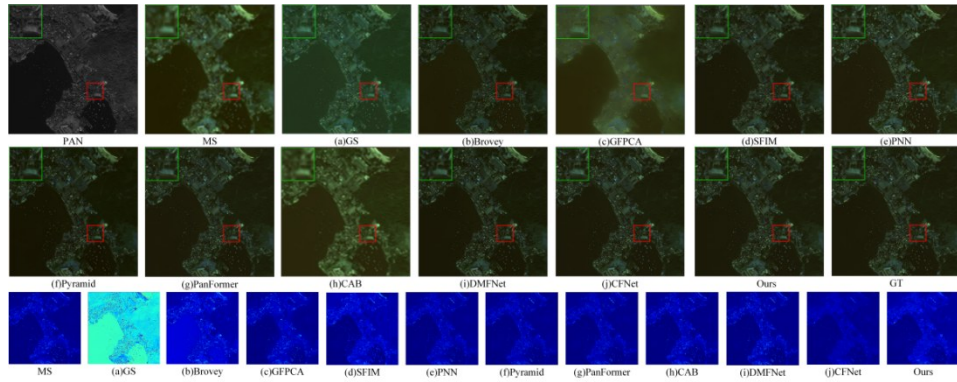


图7 GE1数据集不同方法融合结果及残差图

Fig.7 Fusion results and residual maps of different methods on GE1 dataset

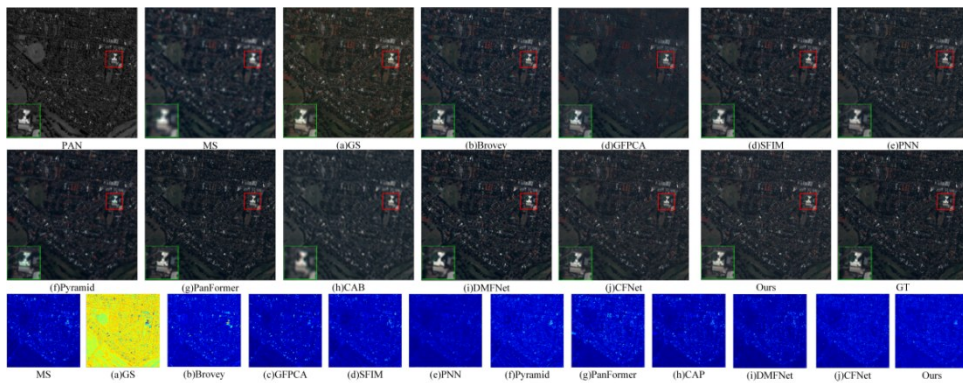


图8 PL1数据集不同方法融合结果及残差图

Fig.8 Fusion results and residual maps of different methods on PL1 dataset

中,所提方法能够更好地保持城市建筑区域的结构完整性与边缘连续性,尤其在图中标记的放大区域,复杂建筑轮廓更加清晰,纹理细节更加丰富;在GE1数据集中,由于地物类型更加复杂,不同方法在空间细节恢复和光谱保持方面表现出一定差异,部分方法在局部纹理区域存在细节表达不足或光谱偏移现象,而本文方法通过空间-光谱联合建模,在保持光谱一致性的同时,能够较好地恢复局部结构信息,对复杂地物场景具有一定的适应能力;在PL1数据集中,所提方法在不同传感器条件下依然能够稳定恢复空间细节,同时有效抑制光谱失真。从各残差图可以进一步观察到,本文方法的误差分布整体较小且更加均匀,说明其在不同区域的重建误差较为一致,具有较好的稳定性。

3.3.3 光谱响应特性分析

为进一步直观分析不同融合方法的光谱信息保持能力,本文分别在WV2、WV3、GE1和PL1测试图像中选取具有典型光谱特征的地物区域,并记录对应空间位置。在相同空间位置处提取GT图像及不同融合结果的逐波段归一化光谱响应,以保证比较的一致性。同时,计算各融合结果与GT在对应波段上的响应差值,用于反映不同方法的光谱偏差。实验结果如图9所示。图9(a)–图9(d)分别展示了WV2、WV3、GE1和PL1数据集上GT与不同方法的归一化光谱响应曲线及对应的光谱误差曲线。WV2和WV3横坐标分别对应8个光谱波段,GE1和PL1横坐标对应4个光谱波段。

如图9所示,不同融合方法均能够较好恢复典型地物区域的整体光谱变化趋势,但部分方法在个别波长处仍存在一定偏差。对于WV2和WV3数据集,部分方法在724 nm、833 nm和949 nm附近波长处误差波动较大;对于GE1和PL1数据集,不同方法整体光谱趋势较为一致,但在490 nm、660 nm和830 nm处仍存在一定响应差异。相比之下,本文方法的光谱响应曲线与GT保持较高一致性,逐波长误差整体分布于零值附近,未出现明显偏离。该结果表明,所提方法在提升空间细节的同时,能够有效保持光谱响应一致性,并与表2–表5中的SAM和ERGAS指标结果相互验证。

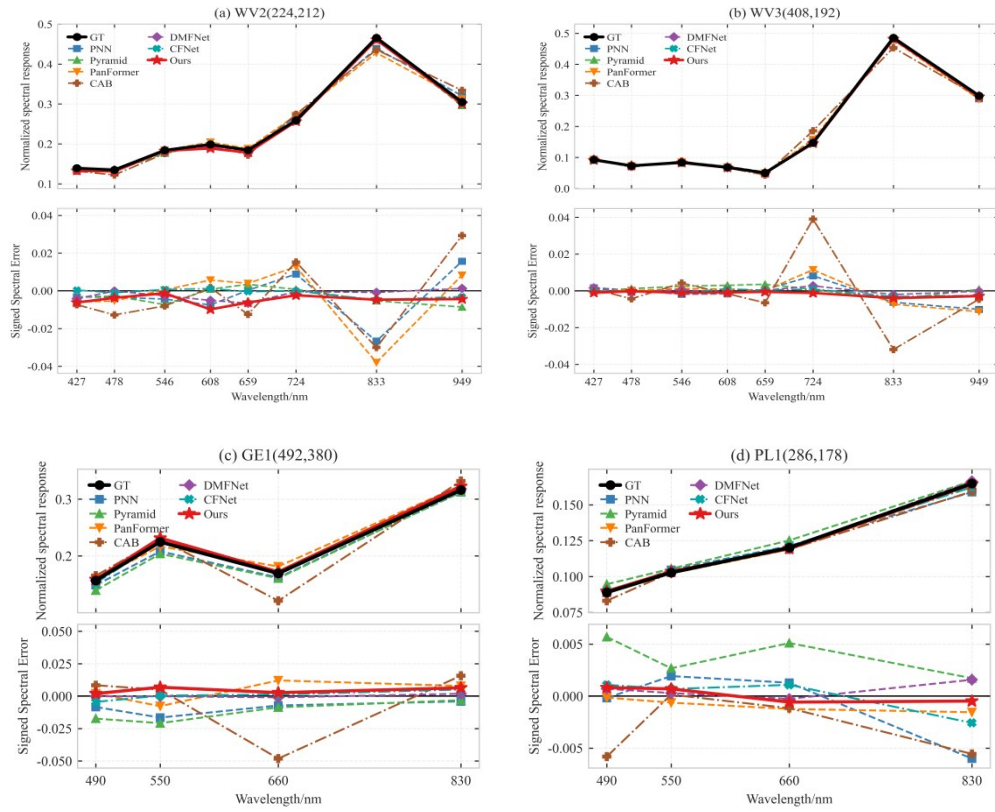


图9 四个数据集上不同方法光谱响应及光谱误差对比。(a)WV2数据集;(b)WV3数据集;(c)GE1数据集;(d)PL1数据集
Fig.9 Comparison of spectral responses and spectral errors of different methods on four datasets. (a) WV2 dataset; (b) WV3 dataset; (c) GE1 dataset; (d) PL1 dataset

3.4 全分辨率实验结果与分析

为进一步验证所提方法在真实遥感场景中的适用性,本文选取 WorldView-2 和 WorldView-3 卫星影像开展全分辨率实验。由于真实场景中缺少对应的高分辨率多光谱参考图像,本文采用无参考质量指标 QNR、光谱失真指数 D_λ 和空间失真指数 D_s 进行定量评价。

3.4.1 定量分析

表6给出了不同方法在 WV2 和 WV3 真实数据集上的无参考定量评价结果。对于 WV2 数据集,本文方法取得了较低的光谱失真指数值,表明其能够较好地保持融合结果的光谱一致性。虽然本文 QNR 和 D_s 指标未达到最优水平,但结合视觉结果可以看出,本文方法仍能够有效恢复建筑边缘、道路纹理及局部地物结构,并保持较为自然的整体色彩表现。对于 WV3 数据集,本文方法取得了最高的 QNR 值,同时保持较低的 D_λ 和 D_s 指标,说明该方法能够在光谱保持与空间细节增强之间取得较好的平衡。在真实全分辨率场景下,所提方法表现出良好的综合融合性能。

3.4.2 定性分析

图10和图11分别展示了不同方法在 WV2 和 WV3 真实全分辨率数据上的视觉融合结果。由整体图像和局部放大区域可以看出,各方法均能够提升多光谱图像的空间分辨率,但在细节恢复、结构保持和光谱一致性方面存在差异。对于 WV2 场景,局部放大区域主要包含车辆目标。相比其他方法,本文方法能够更完整地保持车辆边缘结构,并恢复更加清晰的局部纹理细节,同时有效减少目标模糊和色彩偏移现象。对于 WV3 复杂城市场景,本文方法能够较好地保持建筑边界、道路结构等空间细节,并避免过度增强导致的纹理失真。综合视觉结果可以看出,本文方法在空间细节恢复与光谱保持之间取得了较好的平衡,能够生成结构更加清晰、光谱更加自然的融合结果,验证了所提出空间-光谱联合建模策略和频域增强模块的有效性。

全分辨率实验与降分辨率实验之间的性能差异主要源于评价目标和模型优化侧重点的不同。降分辨率实验具有对应的 HRMS 参考图像,有参考指标可直接评价融合结果与参考图像之间的重建质量;本文提

表 6 WV2和WV3真实数据集的定量评价结果
Table 6 Quantitative evaluation results on the WV2 and WV3 real datasets

	WV2			WV3		
	D_λ	D_s	QNR	D_λ	D_s	QNR
GS ^[25]	0.0856	0.0496	0.8693	0.0297	0.0991	0.8741
Brovey ^[26]	0.0422	0.1251	0.8378	0.0273	0.0744	0.9006
SFIM ^[28]	0.0739	0.1357	0.8006	0.0471	0.0409	0.9144
PNN ^[7]	0.0620	0.1228	0.8233	0.0632	0.0702	0.8717
Pyramid ^[8]	0.0173	0.0828	0.9013	0.0302	0.0496	0.9220
PanFormer ^[11]	0.0527	0.1378	0.8167	0.0454	0.0603	0.8975
CAB ^[9]	0.0704	0.0625	0.8714	0.0482	0.0880	0.8684
DMFNet ^[10]	0.0551	0.0463	0.9011	0.0198	0.1098	0.8730
CFNet ^[18]	0.0079	0.0911	0.9016	0.0204	0.0592	0.9220
Ours	0.0327	0.1045	0.8665	0.0238	0.0555	0.9224
Ideal value	0	0	1	0	0	1

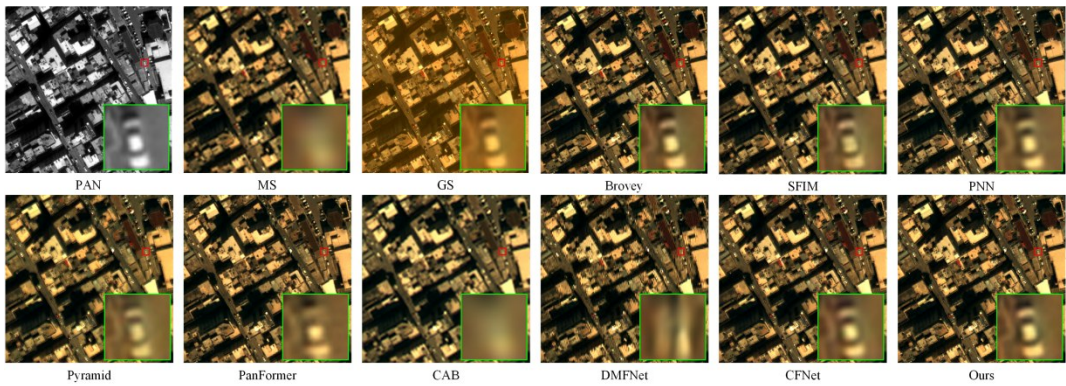


图 10 WV2 真实数据集上不同方法的融合结果
Fig.10 Fusion results of different methods on the WV2 real dataset

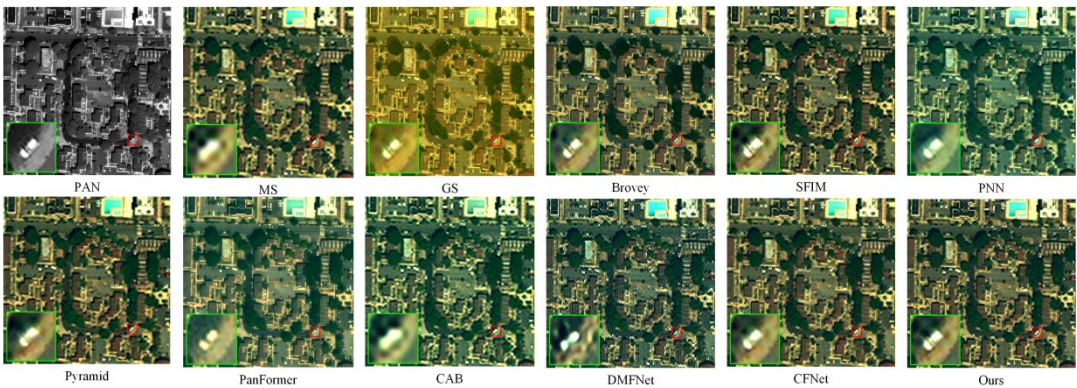


图 11 WV3 真实数据集上不同方法的融合结果
Fig.11 Fusion results of different methods on the WV3 real dataset

出的SSM和FEM模块侧重空间-光谱联合特征建模与高频细节恢复,因此在参考图像约束下表现较好。全分辨率实验缺少对应的HRMS参考图像,无参考指标主要评价融合结果与输入PAN、MS图像之间的光谱和空间统计一致性。模型在增强局部边缘和纹理细节表达的过程中,部分无参考评价指标可能对增强后的局部空间结构变化更为敏感,从而表现出与有参考评价指标不同的评价结果。结合可视化结果可以看出,本文方法仍能够较好地保持空间结构和光谱一致性。后续将进一步研究面向多传感器场景的适应性优化方法,以提升模型在复杂真实遥感场景中的泛化能力。

3.5 配准误差敏感性分析

全色锐化依赖 PAN 与 MS 影像之间准确的空间对应关系。然而,受成像条件及几何校正误差等因素影响,实际遥感影像中仍可能存在残余空间错位,进而导致空间细节注入位置偏差,并降低融合结果的光谱保真度。为评价所提方法在非理想配准条件下的融合稳定性,本文借鉴 BARONTI 等^[30]关于通过人为引入空间错位评价配准误差影响的实验思想,采用人工平移 PAN 图像的方式构造不同程度的可控空间偏移。

表 7 不同 PAN 偏移量下各方法在 PL1 数据集上的融合性能
Table 7 Fusion performance of different methods on the PL1 dataset under different PAN shifts

Pyramid				PanFormer			
Shift/pixel	PSNR	CC	ERGAS	Shift/pixel	PSNR	CC	ERGAS
0	37.1183	0.9424	2.9520	0	39.6807	0.9646	2.1740
1	34.5184	0.9069	4.0019	1	35.2287	0.9211	3.7161
2	32.1153	0.8432	5.3790	2	31.7738	0.8301	5.6983
4	31.1464	0.8038	6.0609	4	29.8814	0.7339	7.2143
CFNet				Ours			
Shift/pixel	PSNR	CC	ERGAS	Shift/pixel	PSNR	CC	ERGAS
0	40.2569	0.9704	2.0978	0	40.4912	0.9673	1.9877
1	34.7619	0.9167	3.9154	1	37.7455	0.9490	2.7291
2	31.2492	0.8121	6.0441	2	34.2730	0.9030	4.1618
4	29.4082	0.7072	7.5737	4	32.4219	0.8547	5.2216

实验中保持模型参数、MS 图像及参考 HRMS 图像不变,仅对 PAN 图像施加 0、1、2 和 4 个高分辨率像素的平移扰动,并沿四个方向测试,最终取平均结果。考虑到 PL1 数据集具有较高空间分辨率和更明显的空间结构变化,本文选取 PL1 作为配准误差实验对象。为比较不同网络结构和融合机制方法在空间错位条件下的性能变化,选择 Pyramid、PanFormer 和 CFNet 作为对比方法,分别代表金字塔多尺度特征表示、Transformer 跨模态融合和频率混合注意力融合方法,并采用 PSNR、CC 和 ERGAS 指标进行评价。

由表 7 可知,随着 PAN 偏移量增加,各方法的 PSNR 和 CC 整体呈下降趋势,而 ERGAS 逐渐增大,说明空间配准误差会对融合结果造成一定影响。在不同像素偏移条件下,本文方法均保持较好的指标表现,其中 PSNR 和 CC 均优于大部分对比方法,ERGAS 也保持较低水平。当偏移量达到 4 个像素时,本文方法仍取得最高的 PSNR 和 CC 以及较低的 ERGAS。相比无偏移情况,本文方法的 CC 下降 0.1126,下降幅度小于 PanFormer 和 CFNet,与 Pyramid 相近,说明所提出方法在一定程度的全局平移误差下仍能够保持较稳定的融合性能。但本文实验目前主要针对全局平移型配准误差进行分析,对于旋转、尺度变化等更复杂的失配情况将在后续工作中进一步研究。

3.6 消融实验

为验证所提方法中各关键模块的有效性,本文在 PL1 数据集上开展消融实验。具体设置 5 种网络结构:1)基础模型(S-U),即单一 U-Net 架构;2)EUB 模型,引入双分支特征提取结构;3)EUB+SSM 模型,在 EUB 基础上加入光谱 Swin Transformer 模块(SSM);4)EUB+FEM 模型,在 EUB 基础上加入频域增强模块(FEM);5)完整模型(Ours),在 EUB+SSM 基础上进一步引入频域增强模块(FEM)。

表 8 给出了不同网络结构的定量评价结果。与基础模型 S-U 相比,引入 EUB 后,CC、ERGAS、Q4 和

表 8 不同网络架构评价结果
Table 8 Evaluation results of different network architectures

	EUB	SSM	FEM	CC	SAM	ERGAS	Q4	PSNR
S-U	×	×	×	0.9333	0.0743	2.6087	0.9558	38.7752
EUB	✓	×	×	0.9474	0.0762	2.4331	0.9565	39.0776
EUB+SSM	✓	✓	×	0.9675	0.0773	2.5765	0.9569	40.1020
EUB+FEM	✓	×	✓	0.9568	0.0729	2.4333	0.9523	39.1786
Ours	✓	✓	✓	0.9673	0.0676	1.9877	0.9583	40.4912

PSNR均有所改善,说明双分支编码结构有助于分别提取PAN图像的空间细节和MS图像的光谱特征。进一步加入SSM后,CC提升至0.9675,PSNR提升至40.1020 dB,表明该模块能够增强空间结构关系建模和整体重建能力。单独引入FEM后,CC和PSNR有所提高,SAM降低至0.0729,说明其对空间细节恢复和光谱保持具有一定促进作用。进一步结合SSM与FEM后,完整模型在SAM、ERGAS、Q4和PSNR指标上取得最优结果,且CC接近最优值,表明两种模块在空间-光谱关系建模与频率信息增强方面具有互补作用。总体来看,各模块分别从异构特征提取、空间-光谱联合建模和频率特征增强等方面改善了融合性能,验证了所提网络结构的合理性与有效性。

图12展示了不同模块组合下的可视化对比结果。从图中可以看出,相较于单一U-Net模型,EUB结构能够明显改善建筑物边缘轮廓的清晰度,使地物边界过渡更加自然。引入SSM模块后,建筑群与河流交界区域的结构细节更加完整,颜色过渡整体较为自然,表明空间-光谱联合特征表示得到增强。加入FEM模块后,融合结果在高频细节恢复和整体视觉一致性方面达到最佳效果,不仅增强了纹理细节,还有效抑制了过度锐化带来的伪影现象。这种逐步改进的效果充分验证了所提出各个模块在提升全色锐化性能方面的有效性,尤其在复杂城市场景中具有较好的适应性。

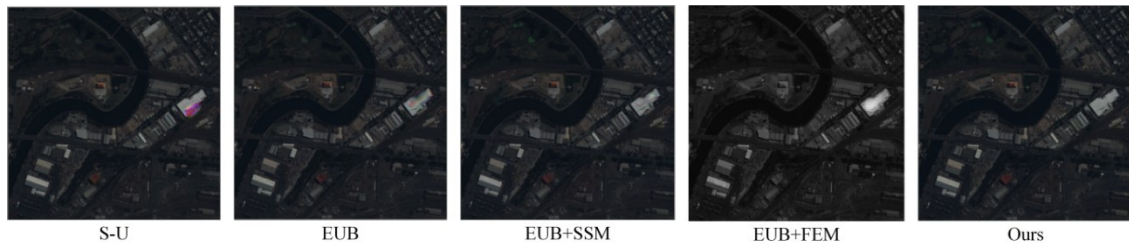


图12 消融实验对比图

Fig.12 Comparison of ablation experiments

3.7 计算复杂度分析

为进一步评估本文方法在融合性能提升过程中引入的计算开销,本文统计不同深度学习方法的参数量(Params)和浮点运算次数(FLOPs),结果如表9所示。所有方法均在PAN图像尺寸为 512×512 、MS图像尺寸为 128×128 的统一输入条件下进行理论复杂度统计,以保证比较的一致性。

表9 不同深度学习全色锐化方法的模型复杂度对比

Table 9 Comparison of model complexity of different deep learning-based pan-sharpening methods

	Params/M	FLOPs/G
PNN	0.08	21.06
Pyramid	0.32	184.10
PanFormer	1.55	48.68
CAB	0.26	31.65
DMFNet	1.02	266.16
CFNet	0.12	17.16
Ours	1.03	143.78

由表9可知,本文方法参数量为1.03 M,与DMFNet基本相当,低于PanFormer,表明本文网络在引入双分支特征编码、空间-光谱关系建模和频域增强模块后,仍保持较为合理的参数规模。在计算量方面,本文方法FLOPs为143.78 G,低于DMFNet和Pyramid,但高于其他轻量化方法。其计算开销主要来源于双分支特征提取、多尺度特征交互以及空间-光谱与频域特征建模过程。此外,Params和FLOPs主要反映理论计算复杂度,实际推理时间还受到输入尺寸、特征图规模和网络结构等因素影响,因此在处理大尺寸遥感影像时会产生额外计算开销。总体而言,本文方法通过增加一定计算成本提升了空间细节恢复和光谱保持能力,在融合性能提升与计算开销增加之间取得了较好的平衡。未来将进一步通过网络结构优化和轻量化设计降低推理时间,提高模型在大规模遥感数据处理中的应用效率。

4 结论

本文针对全色锐化过程中空间细节增强与光谱保持难以兼顾的问题,提出了一种基于光谱 Swin Transformer和频域增强的融合方法。通过构建改进的双分支 U-Net 骨干网络,实现了全色图像与多光谱图像特征的解耦建模与多尺度表达;采用光谱 Swin Transformer 模块,增强空间上下文信息与光谱相关特征通道关系的表征能力;设计频域增强模块,对高频与低频成分进行自适应分离与融合,从而改善细节恢复效果并缓解光谱失真问题。在 WV2、WV3、GE1 和 PL1 四个公开数据集上的实验表明,所提方法不仅能够有效处理 4 波段的 GE1 和 PL1 数据,而且在 8 波段的 WV2 和 WV3 数据集上也表现出较好的融合性能。消融实验进一步验证了各个模块在提升网络性能方面的有效性。未来研究将考虑引入更轻量化的网络设计,降低计算开销并提高处理效率;其次,探索跨传感器泛化能力,使模型能够适应不同卫星传感器获取的图像,减少对特定传感器数据的依赖。

参考文献

- [1] FAN Y, WAN Y, WANG H, et al. Application of an airborne hyperspectral survey system CASI/SASI in the gold-silver-lead-zinc ore district of Huaniushan, Gansu, China[J]. *Geologia Croatica*, 2021, 74(1): 73-83.
- [2] WU Y, HUANG M, LI Y, et al. A distributed fusion framework of multispectral and panchromatic images based on residual network[J]. *Remote Sensing*, 2021, 13(13): 2556.
- [3] BAKKEN S, LANGER D D, GARRETT J L, et al. Spatial resolution enhancement of hyperspectral images from miniaturized sensors[C]//Image and Signal Processing for Remote Sensing XXIX. Bellingham: SPIE, 2023: 24-38.
- [4] LIN H, DONG Y, DING X, et al. Unsupervised pansharpening via mutually guided detail restoration[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI Press, 2024: 3386-3394.
- [5] WANG P, YAO H, LI C, et al. Multiresolution analysis based on dual-scale regression for pansharpening[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2021, 60: 1-19.
- [6] VIVONE G, DALLA MURA M, GARZELLI A, et al. A new benchmark based on recent advances in multispectral pansharpening: Revisiting pansharpening with classical and emerging methods[J]. *IEEE Geoscience and Remote Sensing Magazine*, 2020, 9(1): 53-81.
- [7] MASI G, COZZOLINO D, VERDOLIVA L, et al. Pansharpening by convolutional neural networks[J]. *Remote Sensing*, 2016, 8(7): 594.
- [8] ADEEL H, ALI S S, RIAZ M M, et al. Learning deep pyramid-based representations for pansharpening[J]. *Arabian Journal for Science and Engineering*, 2022, 47(8): 10655-10666.
- [9] LI J, FAN W, LIAN T, et al. Cross-attention-based common and unique feature extraction for pansharpening[J]. *IEEE Geoscience and Remote Sensing Letters*, 2023, 20: 5002905.
- [10] GUO H, JIN X, JIANG Q, et al. DMF-Net: A dual remote sensing image fusion network based on multiscale convolutional dense connectivity with performance measure[J]. *IEEE Transactions on Instrumentation and Measurement*, 2024, 73: 1-15.
- [11] ZHOU H, LIU Q, WANG Y. PanFormer: A transformer-based model for pansharpening[C]//2022 IEEE International Conference on Multimedia and Expo (ICME). New York: IEEE, 2022: 1-6.
- [12] MA M, JIANG Y, ZHAO M, et al. Deep spatial-spectral fusion transformer for remote sensing pansharpening[J]. *Information Fusion*, 2025, 118: 102980.
- [13] SONG Q, LU H, XU C, et al. Invertible attention-guided adaptive convolution and dual-domain transformer for pansharpening[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025, 18: 5217-5231.
- [14] ZHENG Jianwei, XIA Hongyi, XU Honghui. Spatial-spectral Collaborative Unrolling Network for pansharpening[J]. *Acta Photonica Sinica*, 2025, 54(1): 0110003.
郑建伟,夏弘毅,徐宏辉.空间-光谱联合解卷的全色锐化网络[J]. *光子学报*, 2025, 54(1): 0110003.
- [15] JONNALA N S, BHEEMANA R C, PRAKASH K, et al. DSIA U-Net: Deep shallow interaction with attention mechanism UNet for remote sensing satellite images[J]. *Scientific Reports*, 2025, 15(1): 549.
- [16] HAN K, WANG Y, CHEN H, et al. A survey on vision transformer[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 45(1): 87-110.
- [17] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2021: 10012-10022.
- [18] XING Y, QU L, ZHANG K, et al. Complementary fusion network based on frequency hybrid attention for pansharpening[C]//ICASSP 2024 IEEE International Conference on Acoustics, Speech and Signal Processing. New York: IEEE, 2024: 2650-2654.

- [19] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 7132–7141.
- [20] LI T, LIU L, LI M. Multi-scale residual depthwise separable convolution for metro passenger flow prediction[J]. Applied Sciences, 2023, 13(20): 11272.
- [21] YADAV O, GHOSAL K, LUTZ S, et al. Frequency-domain loss function for deep exposure correction of dark images[J]. Signal, Image and Video Processing, 2021, 15(8): 1829–1836.
- [22] PALSSON F, SVEINSSON J R, ULFARSSON M O, et al. Quantitative quality evaluation of pansharpened imagery: Consistency versus synthesis[J]. IEEE Transactions on Geoscience and Remote Sensing, 2015, 54(3): 1247–1259.
- [23] FAN W, LIU F, LI J. A pan-sharpening network using multi-resolution transformer and two-stage feature fusion[J]. PeerJ Computer Science, 2023, 9: e1488.
- [24] XIE Y, WU W, YANG H, et al. Detail information prior net for remote sensing image pansharpening[J]. Remote Sensing, 2021, 13(14): 2800.
- [25] XU S, ZHANG J, ZHAO Z, et al. Deep gradient projection networks for pansharpening[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2021: 1366–1375.
- [26] ALCARAS E, PARENTE C, VALLARIO A. Automation of pansharpening methods for Pléiades images using GIS basic functions[J]. Remote Sensing, 2021, 13(8): 1550.
- [27] ABURAED N, ALKHATIB M Q, MARSHALL S, et al. A review of spatial enhancement of hyperspectral remote sensing imaging techniques[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023, 16: 2275–2300.
- [28] WANG Mi, HE Luxiao, CHENG Yufeng, et al. Panchromatic and multispectral fusion method combined with adaptive Gaussian filter and SFIM model[J]. Acta Geodaetica et Cartographica Sinica, 2018, 47(1): 82–90. DOI: 10.11947/j. AGCS.2018.20170421.
王密,何鲁晓,程宇峰,等. 自适应高斯滤波与SFIM模型相结合的全色多光谱影像融合方法[J]. 测绘学报, 2018, 47(1): 82–90.
- [29] JAVAN F D, SAMADZADEGAN F, MEHRAVAR S, et al. A review of image fusion techniques for pansharpening of high-resolution satellite imagery[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2021, 171: 101–117.
- [30] BARONTI S, AIAZZI B, SELVA M, et al. A theoretical analysis of the effects of aliasing and misregistration on pansharpened imagery[J]. IEEE Journal of Selected Topics in Signal Processing, 2011, 5(3): 446–453.

A Pan-sharpening Method Based on Spectral Swin Transformer and Frequency Enhancement

LI Rong^{1,2}, CHANG Xia^{1,2}, MA Ning^{2,3}

(1 School of Mathematics and Information Science, North Minzu University, Yinchuan 750021, China)

(2 Ningxia Key Laboratory of Intelligent Information and Big Data Processing, Yinchuan 750021, China)

(3 School of Computer Science and Engineering, North Minzu University, Yinchuan 750021, China)

Abstract: Pan-sharpening reconstructs high-resolution multispectral (HRMS) images by integrating the spatial details of panchromatic (PAN) images and the spectral information of multispectral (MS) images. However, achieving a balance between spatial detail enhancement and spectral fidelity remains challenging due to the complex spatial-spectral relationships and frequency characteristics of PAN and MS images. This study proposes a pan-sharpening method based on Spectral Swin Transformer and frequency enhancement to improve spatial-spectral feature representation and reconstruction quality. The proposed method enhances the interaction between heterogeneous PAN and MS features, captures spatial structures and spectral correlations, and reduces spectral distortion during detail reconstruction. Experimental evaluations demonstrate that the proposed framework improves fusion performance by jointly optimizing spatial detail recovery and spectral information preservation. An end-to-end deep learning framework is developed with an enhanced dual-branch U-Net backbone (EUB), a Spectral Swin Transformer module (SSM), and a frequency enhancement module (FEM). The proposed framework achieves effective fusion of PAN spatial information and MS spectral information through modality-specific feature extraction, spatial-spectral interaction modeling, and frequency-aware reconstruction. The EUB adopts a dual-branch encoder-decoder architecture, where PAN and MS images are processed through independent encoding

pathways. The PAN branch focuses on extracting spatial structures, edge information, and texture details, while the MS branch preserves spectral distributions and inter-band characteristics. This design enables heterogeneous inputs to be represented according to their intrinsic characteristics and reduces feature interference before fusion. During decoding, multi-scale features from different levels are progressively integrated to combine shallow spatial details with deep semantic information for HRMS image reconstruction. To enhance spatial-spectral feature representation, the SSM is introduced to separately model spatial contextual dependencies and spectral-related feature interactions. The spatial relation modeling branch adopts the shifted-window attention mechanism to capture local and cross-window spatial relationships, improving the representation of complex structures and texture variations. The spectral relation modeling branch further learns correlations among deep feature channels through cross-channel interaction, thereby enhancing spectral information exchange and maintaining spectral consistency during spatial detail enhancement. The joint modeling of spatial and spectral relationships improves reconstruction capability for complex scenes while preserving spectral information. Furthermore, the FEM is designed to optimize frequency information during the reconstruction process. The input features are adaptively decomposed into high-frequency and low-frequency components, and their contributions are dynamically adjusted according to image characteristics. High-frequency components are enhanced to recover missing spatial details, whereas low-frequency components are preserved to maintain structural information and spectral stability. The reconstructed frequency features are further integrated with spatial-spectral representations to achieve a better balance between detail enhancement and spectral preservation. A composite loss function containing spatial reconstruction constraints, frequency consistency constraints, and spectral similarity constraints is adopted to guide network optimization. The proposed model is implemented using PyTorch and evaluated on four public satellite datasets, including WorldView-2, WorldView-3, GeoEye-1, and Pléiades-1. The proposed method is evaluated on four public remote sensing datasets, including WorldView-2, WorldView-3, GeoEye-1, and Pléiades-1, and compared with representative traditional and deep learning-based pan-sharpening methods. Multiple full-reference metrics, including correlation coefficient (CC), spectral angle mapper (SAM), relative dimensionless global error in synthesis (ERGAS), quality index (Q4/Q8), and peak signal-to-noise ratio (PSNR), are used for quantitative evaluation. The experimental results demonstrate that the proposed method achieves stable and competitive performance across different satellite datasets. In terms of spatial fidelity, the proposed method achieves competitive CC and Q4/Q8 values on most datasets, indicating improved preservation of spatial structures and image details. Regarding spectral consistency, the proposed method achieves lower SAM and ERGAS values on multiple datasets, demonstrating its capability to reduce spectral distortion during detail enhancement. The PSNR results further indicate that the proposed method provides improved overall reconstruction quality compared with existing approaches. Visual comparisons show that the proposed method produces clearer edges, richer textures, and more consistent spectral representations, particularly in complex scenes with dense buildings, roads, and heterogeneous land covers. Residual analysis reveals more uniform error distributions, indicating stable reconstruction performance in different regions. Spectral response analysis further confirms that the proposed method better maintains the original spectral characteristics of multispectral images. Ablation experiments verify the effectiveness of each component, demonstrating that the enhanced dual-branch backbone improves heterogeneous feature extraction, the Spectral Swin Transformer module strengthens spatial-spectral feature interaction, and the frequency enhancement module contributes to detail recovery and spectral consistency. In addition, evaluations using both reference-based and non-reference indicators show that the proposed method maintains a favorable balance between spatial enhancement and spectral preservation under different sensor conditions. The results on multiple datasets demonstrate that the proposed framework has strong robustness when processing remote sensing images with different spatial resolutions, spectral characteristics, and scene complexities. The proposed framework demonstrates reliable generalization ability across different satellite datasets and provides an effective approach for high-quality HRMS image reconstruction in remote sensing applications. The results indicate that the proposed approach has strong adaptability for remote sensing image fusion tasks involving complex spatial structures and diverse spectral characteristics.

Key words: Remote sensing image; Pan-sharpening; Deep learning; Multispectral image; Swin Transformer; Frequency enhancement

OCIS Codes: 110.4234; 100.2980; 100.3010; 280.4991

CSTR: 32255.14.gzxb20275601.0110001

Foundation item: Natural Science Foundation of Ningxia (2025AAC030002), Construction Project of First-Class Disciplines of Ningxia Higher Education (NXYLXK2017B09), Graduate Innovation Project of North Minzu University (CYX25132).