CNN 联合多尺度 Transformer 的高光谱与多光谱图像融合

徐光宪，周伟杰，马 飞
（辽宁工程技术大学 电子与信息工程学院，辽宁 葫芦岛 125105）

摘要：高光谱图像具有丰富的光谱信息，多光谱图像具有精妙的几何特征，融合高分辨率的多光谱图像和低分辨率的高光谱图像可以获取更为全面的遥感数据图像。然而现有的融合网络大多数基于卷积神经网络所设计，对于结构复杂的遥感类图像而言，依赖于核大小的卷积运算，容易导致特征融合阶段缺乏一些全局上下文信息。为保证图像融合的质量，本文提出了一种 CNN（Convolutional Neural Network, CNN）联合多尺度 transformer 网络来实现多光谱和高光谱图像融合，结合了 CNN 的特征提取能力与 transformer 的全局建模优势。网络将融合任务分为了两个阶段，特征提取阶段和融合阶段。特征提取阶段，针对图像特性，基于卷积神经网络分别设计了不同模块用于特征提取。融合阶段，通过多尺度 transformer 模块从局部到全局建立信息间长距离关联，最后通过多层卷积层将特征映射为高分辨率的高光谱图像。经过在 CAVE 和 Harvard 数据集的实验结果表明，本文所提算法与其他经典算法相比，能更好地提升融合图像的质量。

关键词：高光谱图像；多光谱图像；卷积神经网络；transformer；图像融合
中图分类号：TP391 **文献标识码**：A **文章编号**：1001-8891(2025)01-0052-11

Fusion of Hyperspectral and Multispectral Images Using a CNN Joint Multi-Scale Transformer

XU Guangxian，ZHOU Weijie，MA Fei
(*School of Electronic and Information Engineering, Liaoning Technical University, Huludao* 125105, *China*)

Abstract: Hyperspectral images contain rich spectral information, and multispectral images have exquisite geometric features. More comprehensive remote sensing images can be obtained by merging high-resolution multispectral and low-resolution hyperspectral images. However, most existing fusion networks are based on convolutional neural networks. For remote sensing images with complex structures, convolution operations dependent on the kernel size tend to lead to a lack of global context information in the feature fusion stage. To ensure the quality of image fusion, this study proposes a convolutional neural network (CNN) combined with a multi-scale transformer network to realize multispectral and hyperspectral image fusion, combining the feature extraction capability of the CNN and the global modeling advantage of the transformer. The network divides the fusion task into two stages: feature extraction and fusion. In the feature extraction stage, different modules are designed for feature extraction based on the CNN. In the fusion stage, a multi-scale transformer module is used to establish a long-distance correlation between local and global information, and the features are mapped into high-resolution hyperspectral images through multilayer convolution layers. Experimental results on the CAVE and Harvard datasets show that the proposed algorithm can improve the quality of fused images better than other classical algorithms.

Key words: hyperspectral image, multispectral image, CNN, transformer, image fusion

0 引言

高光谱遥感技术是通过成像光谱仪对同一场景的不同波段进行连续地遥感成像，获取多个波段的图像数据。相较于其他遥感图像，高光谱图像具有更高的光谱分辨率和更宽的光谱范围，能够捕捉到物体更

收稿日期：2023-07-21；**修订日期**：2023-11-07.
作者简介：徐光宪（1977-），男，教授，研究方向为网络编码与信息处理。E-mail：5261009@qq.com。
基金项目：辽宁省科技厅应用基础研究项目（101300274）；辽宁省教育厅研究项目（LJKZ0357）。

丰富的光谱信息[1]。因此在地质勘察[2]、医疗诊断[3]、人脸识别[4]等领域被广泛应用。然而受物理成像系统的限制，高光谱图像在空间分辨率、光谱分辨率之间需要进行一定的权衡，以较低的空间分辨率确保较高的信噪比[5]。相反，对比于高光谱图像，多光谱图像的波段有限，但空间分辨率较高[6]。因此，为了获取更为准确和全面的遥感数据，融合高分辨率的多光谱图像（High-resolution multispectral image，HR-MSI）和低分辨率的高光谱图像（Low-resolution hyperspectral image，LR-HSI）是一种切实可行的办法。

目前实现高光谱图像融合，主要分为 3 大类：基于细节注入，基于优化和基于深度学习。基于细节注入的方法，通过图像全色锐化方法实现 HSI 和 MSI 图像融合。典型的有 Chen 的方法[7]和 Selva 的方法[8]。虽然它们的计算效率较高，却不能保证融合图像的质量，容易导致光谱失真。基于优化的方式有矩阵分解、张量分解和贝叶斯表示。矩阵分解将遥感数据分解为端元矩阵和丰度矩阵，通过对 HSI 端元矩阵和 MSI 的丰度矩阵重建来进行图像融合[9]。典型的算法有耦合非负矩阵分解[10]和 Lanara 的方法[11]。基于贝叶斯的方法，以先验约束和后验概率密度最大化为目标来生成最终的融合图像。典型的有 Sylvester 方程[12]和 Akhtar 的方法[13]。基于张量分解的方法通常将 HSI 图像看作一个三维张量，高空间分辨率的高光谱图像被分割成若干图像块，对图像块聚类，划分为对应的图形块集合[14]。典型的有 Tucker 分解[15-17]和 CP 分解[18]。虽然基于优化的方法在性能方面有着不错的表现，但它将融合问题视为了一个逆问题，依赖于相关传感器特性的知识，通过设计适当的手工先验来获取所需结果，这在实际应用中很难实现，存在着一定的局限性。

基于深度学习的方法，通过构建可学习的深度神经网络自动提取高光谱与多光谱图像特征，利用其非线性拟合能力，建立 HR-MSI、LR-HSI 和对应的 HR-HSI 之间输入到输出的映射关系，从而实现图像融合。相比于依赖手工设置先验特征的优化方法，此类方法利用学习到的映射关系作为先验知识来重建 MSI 和 HSI 中缺失的光谱与空间信息。因此这类方法能获得更好的性能。

随着深度学习的不断发展，尤其是卷积神经网络，近年来遥感领域出现了许多基于 CNN 的融合方法。Yang 等人[1]提出了一种结合 CNN 和空间注意力的 HSI 和 MSI 融合方法，在特征表达过程中专注于关键信息，增强融合图像的空间纹理细节。Li 等人[19]通过 CNN 网络，将 PSF 和 SRF 通过退化模型视为可学习的参数，帮助获得更准确的融合结果。虽然现有的基于卷积神经网络的融合算法[20-22]，通过学习带有卷积核的局部线性映射来提高泛化能力，但它们缺乏对空间位置信息的有效利用，无法对特征图像建立长程依赖关系，在特征融合阶段容易缺乏一些全局上下文信息，融合表现也因此受限。

近年来 Transformer 模型因其自注意力机制在许多视觉任务中的出色表现而受到研究界的关注。通过对图像的长距离依赖关系进行建模以编码特征。在融合特征方面相比于依赖感受野的卷积，全局操作能更好地表达图像特征。然而对图像融合而言，这种全局聚合会牺牲局部特征的表达能力[23]。同时，现有的 transformer 网络大多针对 RGB 图像设计，难以适配结构复杂度更高的遥感类图像。因此，需要进一步研究和设计基于 transformer 网络的特征融合方法，以更好地利用特征全局上下文建模和长程依赖关系，实现特征融合，从而提高融合图像的质量和效果。

针对以上问题，本文设计了一种 CNN 联合多尺度 transformer 的高光谱与多光谱图像融合网络，结合了 CNN 网络的特征提取优势以及 transformer 网络的长距离建模能力。网络将图像融合分为了两个阶段，特征提取阶段和特征融合阶段。具体而言，在特征提取阶段基于 CNN 网络针对图像特性设计了不同模块，有效地提取图像的特征信息，最后在特征融合阶段通过多尺度 transformer 网络从局部到全局对特征图像进行分析和处理，建立特征间长距离依赖关系，从而实现图像特征的融合。

1 问题表述

给定 LR-HSI 图像、HR-MSI 图像和 HR-HSI 图像，以张量的形式分别表示为 $\boldsymbol{X}^{N\times w\times h}$，$\boldsymbol{Y}^{m\times W\times H}$，$\boldsymbol{Z}^{N\times W\times H}$。其中第一维表示光谱带数，第二维和第三维分别表示图像的宽度和高度，并且 $n\ll N$，$w\ll W$，$h\ll H$。LR-HSI 图像是从 HR-HSI 图像退化得来。首先，通过高斯滤波器对 HR-HSI 图像进行模糊处理，然后对图像进行下采样操作。这一过程可表示为：

$$\boldsymbol{X}=\boldsymbol{ZGD} \tag{1}$$

式中：$\boldsymbol{G}$ 代表高斯滤波器；$\boldsymbol{D}$ 代表下采样算子。HR-MSI 是通过光谱响应矩阵对 HR-HSI 图像进行下采样所得。这一过程可以表示为：

$$\boldsymbol{Y}=\boldsymbol{RZ} \tag{2}$$

式中：$\boldsymbol{R}$ 表示光谱响应函数。

假设将网络整体描述为一个端到端的映射函数 $f(\cdot)$，那么这个函数通过给定输入的值生成目标值。表示如下：

$$\boldsymbol{F}=f(\boldsymbol{X},\boldsymbol{Y}) \tag{3}$$

式中：***X***、***Y*** 分别表示输入的 LR-HSI 图像和 HR-MSI 图像；***F*** 则表示将要重建的 HR-HSI 图像。

2 网络及实现

网络分为特征提取和特征融合两个阶段，每个阶段设计了不同的模块，完成阶段任务。特征提取阶段通过残差通道注意力模块（Residual channel attention module，RCAM）和连续的多层次卷积模块（Multi-level convolution module，MCD）完成对高光谱图像和多光谱图像的特征提取。特征融合阶段通过多尺度 transformer 模块（Multi-scale transformer module，MST）实现特征融合。

2.1 整体网络

本文提出的网络结构如图 1 所示。给定输入图像 LR-HSI X 和 HR-MSI Y。由于 X 的尺寸小于 Y，为了保持融合图像的空间大小一致，首先通过双三次插值对 X 进行空间上采样操作。之后将输入的 HR-MSI Y 和上采样后得到的 HR-HSI X^{U} 通过相应模块进行特征提取。多层次卷积模块注重于提取多光谱的空间纹理细节特征，通过连续的多层次卷积模块实现对多光谱空间信息的提取。残差通道注意力模块则注重于对光谱信息的提取与增强。然后将两个模块提取的特征进行初步融合，输入到多尺度 transformer 模块，通过对不同尺度的特征进行长距离相关性建模，融合多光谱的空间信息和高光谱的光谱信息。最后，输出特征通过一系列卷积操作将融合特征映射为 HR-HSI Z，其中为了防止光谱失真，网络最后引入了残差连接。

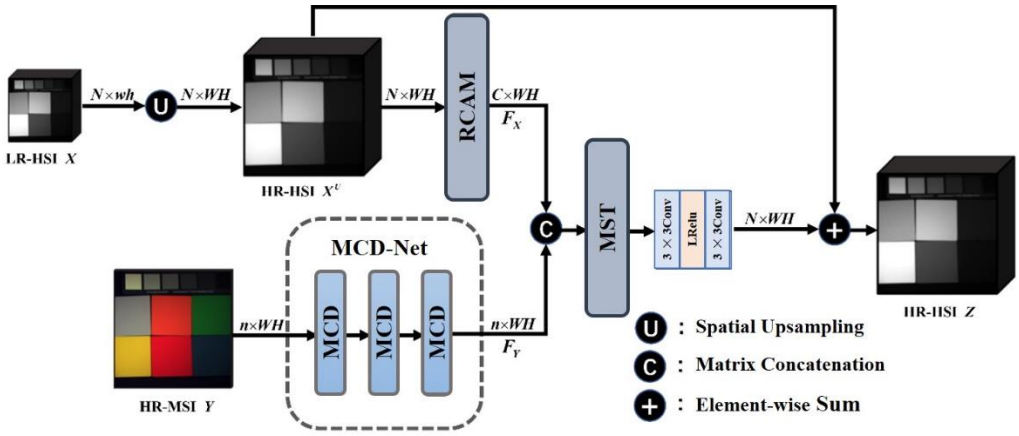


图 1 CNN 联合多尺度 transformer 高光谱融合超分辨网络示意图

Fig.1 Joint CNN and multi-scale transformer hyperspectral fusion super-resolution network

2.2 多层次卷积模块

多光谱图像与高光谱图像相比拥有着更多的空间信息和几何特征。如何对其低级细节纹理特征中的空间信息实现有效提取，是保证融合质量的关键之一。受 Res2Net 网络[24]分层思想的启发。本文设计一种多层次卷积模块，对图像进行各层次的划分，在不同层面进行特征表达，提取图像的空间纹理信息。具体结构如图 2 所示，首先对输入特征通过 1×1 卷积层将低维图像转化为高维特征。随后将所得特征图分成 s 个子组（Channel Splitting，CS），每个子组具有相同的空间大小，分别用 y_i 表示，其中 $i\in\{1,2,\cdots,s\}$。这样每个子组内部、通道之间的关联性更加紧密，可以更好地捕捉到不同层次的特征。然后从第一个子组开始通过独立的 3×3×C 卷积和 LeakyReLU 激活函数实现特征提取后，将所得特征向下传递。同时为保持原有特征图通道之间的关联性，将其与下一子组进行相加合并。重复以上操作，直到最后一组。期间由于组合爆炸的效应，每层输出包含着不同数量和不同组合的感受野大小，获得不同细粒度的特征表达。这一过程可表达为：

$$f_i=\begin{cases}L_i\left(C_i\left(y_i\right)\right) & i=1\\ L_i\left(C_i\left(y_i+f_{i\text{-}1}\right)\right) & 1<i\leqslant s\end{cases} \tag{4}$$

式中：C_i 表示 3×3 卷积；L_i 表示激活函数；f_i 表示提取的特征，其中本文将特征图分为 4 个子组。之后，将所有子组中所得特征图重新合并（Channel Merging，CM），输入 1×1 的卷积层实现信息交互和通道还原。同时为防止网络梯度弥散引入了残差连接，最后将所得特征与空间注意力图相乘实现特征增强后进行输出，完成对空间信息的提取。其中空间注意力图由平均池化对同一像素点不同通道求均值所得，并通过一个卷积层和一个激活函数来优化空间注意图中的空间权重。

2.3 残差通道注意力模块

高光谱图像具有丰富的光谱信息，融合任务中如何充分提取其中的特征和保留光谱信息是我们所关注的重点。为此本文设计了一种残差通道注意力模块，旨在通过学习优化通道注意力的权重实现对高光谱图像光谱信息的关注提取。

如图 3 所示，残差通道注意力模块，首先对上采样（Spatial Upsampling，SU）后的高光谱图像通过 1×1 卷积层进行通道调整和特征表达，然后将所得特征图像通过平均池化层得到各个通道间的权重，采用卷积层和激活函数优化各个通道间的权重。将其权重作为加权系数，对特征图像的每个通道特征进行加权汇聚，动态调整光谱波段的重要性。之后通过一系列卷积操作进一步提高特征的表达能力和准确性，同时通过残差连接以便更好地保留和传递光谱信息，最后通过 1×1 卷积层调整特征映射的通道数进行输出。

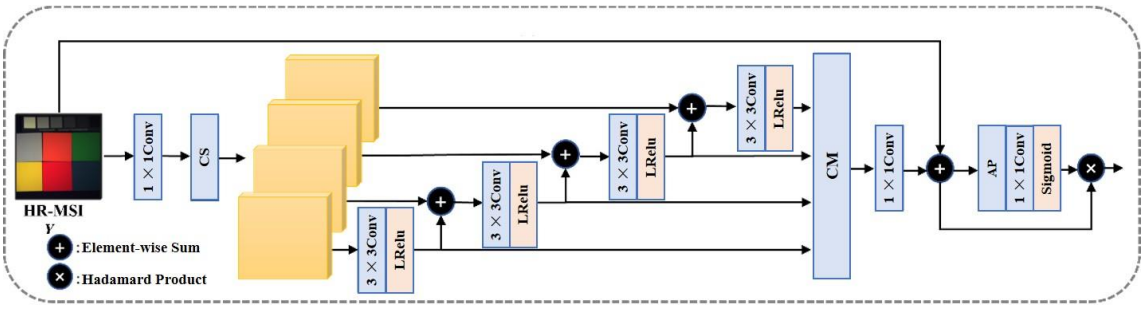


图 2 多层次卷积模块

Fig.2 The multilevel convolutional module

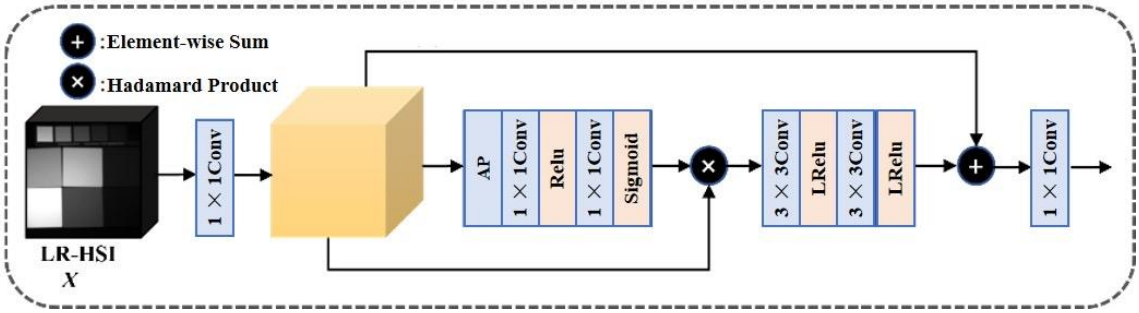


图 3 残差通道注意力模块

Fig.3 The residual channel attention module

2.4 多尺度 transformer 融合模块

如何实现多光谱和高光谱之间的特征融合，是高光谱超分辨率的关键。现有的深度融合模型，通常基于卷积神经网络设计，缺少对特征间全局关系的探索。近年来 transformer 网络因其在许多视觉任务中的出色表现而受到研究界的广泛关注，通过自注意力机制对图像之间的全局关系进行建模从而更好地表达图像特征。

受文献[25]的启发，本文提出了一种多尺度 transformer 融合模块，以 Swin Transformer layer（STL）为网络主干，通过其窗口的自注意力机制探索信息间的依赖交互，同时结合下采样（Spatial Downsampling，SD）金字塔结构，对融合特征从局部到全局实现信息间的长距离相关性建模。

多尺度 transformer 模块如图 4 所示。首先通过下采样操作将融合特征分为不同的尺度，然后将各尺度的特征输入连续的 STL 网络，探索各特征间的相关性，之后通过 1×1 卷积层实现不同通道之间的信息交互，最后通过上采样操作将各尺度还原为原来大小，融合各尺度特征，实现特征长距离相关性建模。具体来说 STL 网络中滑动窗口注意力机制，建立特征间的信息关联时只与窗口内部信息相关。为实现特征间的全局关联引入金字塔结构，融合特征每下采样一次后，窗口内部信息范围更广，实现信息关联的距离更长。随着金字塔的逐渐加深，窗口内部的信息范围从局部区域扩展到全局，实现从细节信息到语义信息的长距离相关性建模。其中，在上面的层，彼此靠近的特征信息间贡献更多的相关性，实现对低级细节信息间关联。下面的层，窗口信息范围逐渐扩散至全局，实现语义信息间的关联。因此多尺度 transformer 模块

在探索特征间长距离相关性时，能同时关注细节与语义信息，从局部到全局建立特征间的信息关联，保证特征融合的质量。

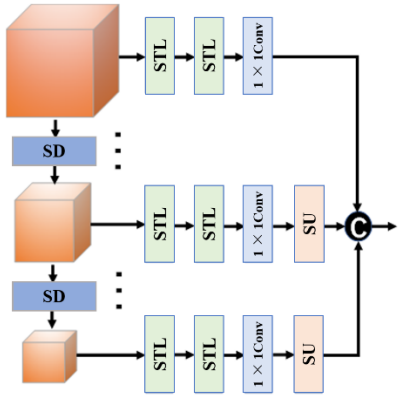


图 4 多尺度 transformer 模块

Fig.4 The multi-scale transformer module

STL 结构如图 5 所示，包含两层 transformer 层，第一层由基于窗口的多头自注意力模块（Window-based multi-head self-attention，WMSA），归一化层（Layernorm，LN）、多层感知机（Multilayer perceptron，MLP）构成。第一层 transformer 将输入特征图像从通道维度进行归一化后，分割成多个大小相等的不重叠窗口。对于每个窗口，计算局部窗口自注意力，从而捕获窗口内部的特征关系，具体步骤如公式(5)所示：

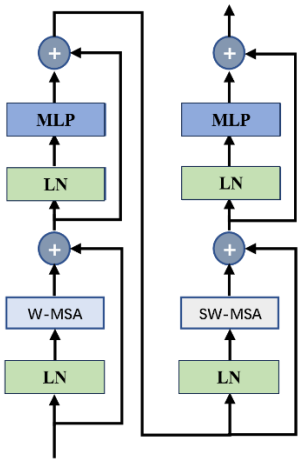


图 5 Swin transformer 模块

Fig.5 Swin transformer layer

$$\boldsymbol{Q}=\boldsymbol{\Phi}_z\boldsymbol{W}_Q,\boldsymbol{K}=\boldsymbol{\Phi}_z\boldsymbol{W}_K,\boldsymbol{V}=\boldsymbol{\Phi}_z\boldsymbol{W}_V$$

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{SoftMax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}+P\right)\boldsymbol{V} \tag{5}$$

式中：$\boldsymbol{W}_Q$、$\boldsymbol{W}_K$、$\boldsymbol{W}_V$ 分别表示基于局部窗口 $\boldsymbol{\Phi}_z$ 所生成的映射矩阵，分别与 $\boldsymbol{\Phi}_z$ 相乘得到矩阵 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$，由此生成局部窗口自注意力矩阵，P 为相对位置编码，d 表示映射矩阵 $\boldsymbol{Q}$ 与 $\boldsymbol{K}$ 的向量维度，Attention 表示局部窗口自注意力。随后再次进行归一化，防止训练过程中的梯度消失和梯度爆炸问题。最后，通过多层感知机进行输出。

第二层 transformer 层采用基于偏移窗口的多头自注意力模块（Shift window-based multi-head self-attention，SW-MSA），对窗口进行滑动，计算滑动后局部窗口自注意力，这样在双层滑窗机制的作用下实现更多的信息交互。

3 实验与分析

3.1 数据集

为评估 CNN 联合多尺度 transformer 网络在遥感图像融合中的性能，本文在 CAVE、Harvard 和 Pavia University（PU）数据集上进行了实验，并与其他先进方法进行了对比。

CAVE 数据集包含 32 个高光谱图像，每张图像为 512×512 像素，31 个波段（400～700 nm）。训练时使用前 20 张图像，后 12 张图像用于测试。原始 HR-HSI 图像通过交叠裁剪为 64×64×31 的图像作为训练样本，LR-HSI 图像通过高斯滤波器模糊后下采样得到，HR-MSI 图像则通过光谱响应矩阵下采样得到，总波段为 3。

Harvard 数据集由 50 张图像组成，每张图像为 1392×1040 像素，31 个波段（420～720 nm）。训练使用前 30 张图像，后 20 张用于测试，处理方式与 CAVE 数据集相同。为了降低测试时间，测试数据从中心裁剪为 512×512×31。

PU 数据集来自意大利 Pavia University，图像为 610×340 像素，103 个波段（0.43～0.86 μm）。选择中心 192×192 区域，使用滑动窗口生成训练和测试图像。LR-HSI 图像生成方式与 CAVE 数据集相同，HR-MSI 则通过 IKONOS 光谱响应矩阵下采样生成，波段数为 4。

3.2 配置环境与网络参数

本文在单个 NVIDIA GeForce 3090 Ti GPU 中使用 Pytorch 训练网络。采用均方误差作为损失函数。同时选择具有默认参数的 Adam 优化器[26]来最小化训练的损失参数，其中学习率被设置为 1×10^{-4}。训练总迭代次数为 500 代，训练批次大小为 16。

3.3 对比算法及评价指标

为了评估所提出方法的性能，本文提出的方法与深度学习类方法和经典的传统类方法做对比。其中基于传统的方法有 NLSTF（Non-Local Sparse Tensor Factorization）[15]，NLSTF-SMBF（Non-Local Sparse Tensor Factorization - Semiblind Fusion）[16]，CSTF（Coupled Sparse Tensor Factorization）[17]，基于深度学习的方法有 SSRNET（Spatial–Spectral Reconstruction

Network）[27]，ResTFNet（Residual Tensor Fusion Network）[28]，MHF-Net（Multispectral and Hyperspectral Image Fusion Network）[29]，HSRnet（Deep Spatiospectral Attention Convolutional Neural Network）[30]。同时使用峰值信噪比（Peak signal-to-noise ratio，PSNR）、光谱角映射（Spectral angle mapper，SAM）、相对全局融合误差（Error relative globale adimensionnelle de Synthese，ERGAS）和结构相似性指数（Structural similarity index，SSIM）这 4 个指标来对本文方法和对比方法进行定量分析，PSNR 和 SSIM 测量空间域中的融合质量。SAM 测量光谱质量，计算融合图像和参考图像之间的平均光谱角度。ERGAS 测量融合结果的整体质量，包括空间和频谱。PSNR 和 SSIM 值越大，SAM 和 ERGAS 值越小表明融合质量越好。

3.4 CAVE 数据集实验结果

测试集由 CAVE 数据集中后 12 张图像组成。为了更客观地展示各种算法的融合结果，我们取 12 幅图像的客观评价指标的平均值，并标注红色最优，蓝色次之，如表 1 所示，本文算法的所有客观指标都是最优的。与表现最好的传统方法 NLSTF 相比 PSNR 指标提高了 1.88 dB，与性能第二的 HSRnet 相比 PSNR 指标提高了 0.75 dB，其中 ERGAS 值在各算法对比中也取得了最低值。表明本文方法在保留光谱信息的同时能更好地提高空间分辨率。同时为显示每种算法在测试集中每张图像上的表现，本文根据 CAVE 数据集中 12 张测试图像的 PSNR 值绘制了雷达图。雷达图的半径越大，测试图像的 PSNR 值越大，融合效果越好。外圈数字代表 12 张测试图像。如图 6 所示，从图中可以看出，本文经过了多次实验测试，该方法在大多数图像上都获得了最优值，排除了实验的偶然性，同时证明了本文方法的优越性。

为了更直观地展示不同融合算法的有效性和融合结果，本文给出了各算法融合结果与真实图像的伪彩色图像（31-15-2 波段），重建图像与真实图像第 15 个波段的误差图，其中误差图由融合图像与真值图像的误差绝对值生成，误差越小表示与原图像越接近。同时为了方便观察图像，局部细节进行了放大。如图 7 所示，可以看出传统方法 NLSTF-SMBF 的融合性能

表 1 不同方法在 CAVE 数据集上的融合结果

Table 1 Fusion results of different methods on CAVE data set

Methods	CAVE			
	PSNR	SAM	ERGAS	SSIM
CSTF	45.14	4.4	1.44	0.9873
NLSTF	47.69	3.47	1.22	0.9887
NLSTF-SMBF	42.82	6.12	2.06	0.9818
SSRNET	44.41	2.60	1.40	0.9928
ResTFNet	45.99	2.19	1.07	0.9947
MHF-Net	48.40	2.01	0.85	0.9973
HSRnet	48.82	1.86	0.79	0.9978
Ours	49.57	1.77	0.67	0.9986

Note: Red font represents the best; Blue font represents suboptimal.

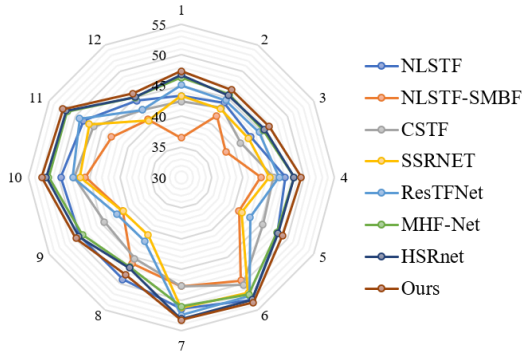
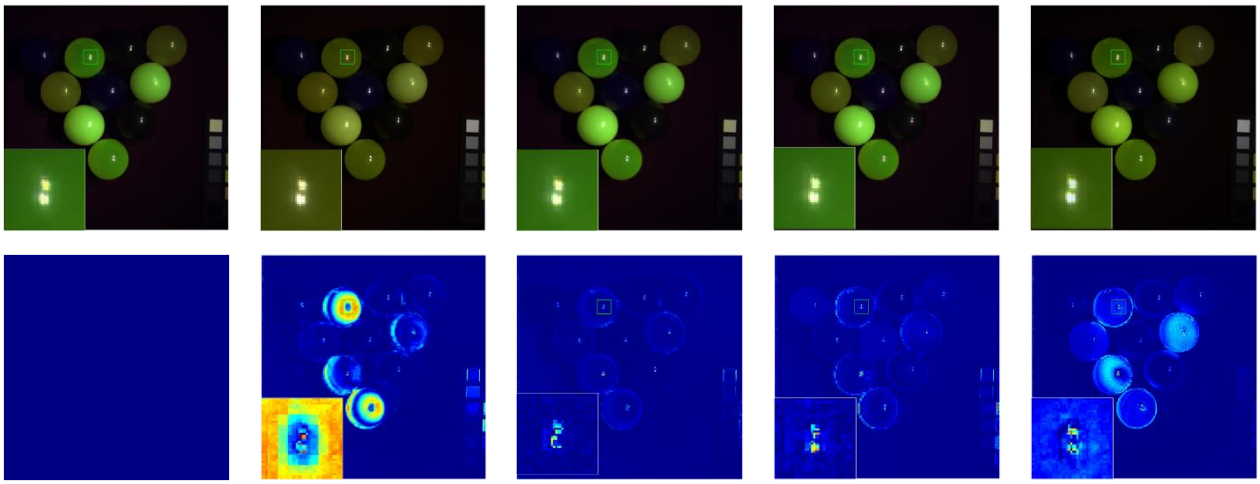
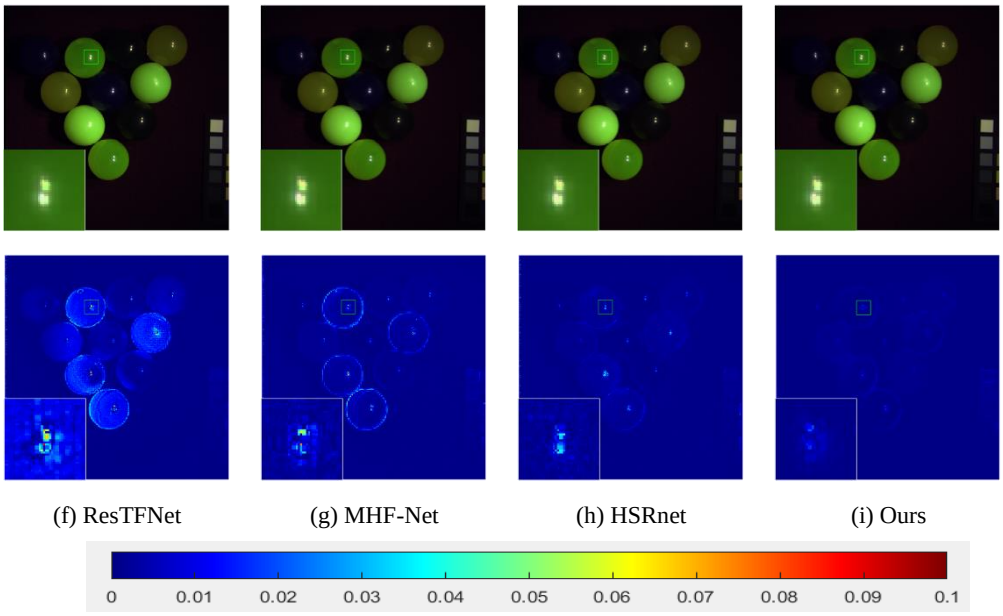


图 6 CAVE 数据集 12 张测试图片的 PSNR 值

Fig.6 PSNR values of 12 test images in the CAVE dataset



(a)GT (b) NLSTF-SMBF (c) NLSTF (d) CSTF (e) SSRNET



(f) ResTFNet (g) MHF-Net (h) HSRnet (i) Ours

图 7 CAVE 数据集重建结果与误差图

Fig.7 CAVE dataset reconstruction results and error images

最差，小球部分光谱失真明显。基于深度学习的方法中 SSRNET 小球部分失真明显，其余方法都取得了不错的重建结果。其中对图像重建结果的细节部分结合波段误差图，可以看出光亮部分本文方法的重建结果更接近真实图像。表明本文方法在图像融合中具有更好的性能，通过本文算法得到的融合图像具有最佳的视觉效果。

3.5 Harvard 数据集实验结果

本文在 Harvard 数据集中选择了后 20 张图像作为测试集，并展示了各种融合算法的客观评价指标平均值，同时红色标注最优，蓝色次之。如表 2 所示，实验结果表明，本文网络在各个方面都取得了最出色的性能。与性能第二的 HSRnet 相比，PSNR 数值提高了 0.55 dB，ERGAS 也取得了最优值。这充分说明了本文所提算法能更好地保留光谱信息以及原图像的空间细节。从侧面表明本文所提算法将 CNN 与 transformer 模型结合的有效性。其中可以观察到相较于 CAVE 数据集，基于深度学习的方法在 Harvard 数据集上表现更好，其评价指标数值较传统方法更优。这表明在融合图像的整体质量方面，基于深度学习的方法具有更大的优势。另外根据 Harvard 数据集 20 张测试图像的 PSNR 值绘制了雷达图，如图 8 所示。从图 8 中可以看出，本文经过了多次实验测试，该方法在所有的图像上都获得了最优值，再次证明了该方法的有效性。

表 2 不同方法在 Harvard 数据集上的融合结果

Table 2 Fusion results of different methods on Harvard dataset

Methods	Harvard			
	PSNR	SAM	ERGAS	SSIM
CSTF	45.43	3.20	2.14	0.9814
NLSTF	46.30	3.05	1.91	0.9826
NLSTF-SMBF	45.52	3.43	2.04	0.9819
SSRNET	46.25	2.26	1.56	0.9939
ResTFNet	46.89	1.98	1.42	0.9946
MHF-Net	47.62	1.79	1.18	0.9965
HSRnet	48.28	1.69	1.09	0.9972
Ours	48.83	1.58	0.94	0.9978

Note: Red font represents the best; Blue font represents suboptimal.

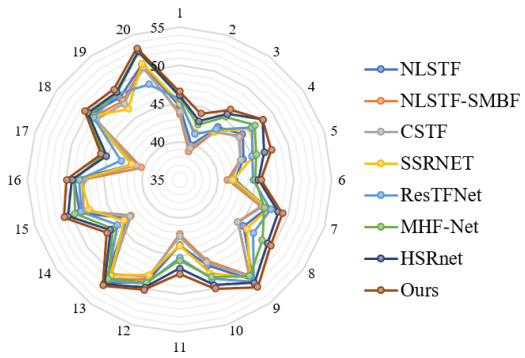


图 8 Harvard 数据集 20 张测试图片的 PSNR 值

Fig.8 PSNR values of 12 test images in the Harvard dataset

为了更直观地观察每种融合算法的重建图像的质量，主要视觉呈现了各类算法的重建图像与真实图像的 R-G-B 图像（31-15-2 波段）和细节放大部分，以及重建图像与真实图像第 15 个波段的可视化误差

图。如图 9 所示，在重建 R-G-B 图像可以看出各类算法都有着不错的融合效果，然而与误差图结合分析可以看出基于深度学习的整体重建误差相较于传统算法性能更佳，同时本文算法的误差最小，重建图像更接近真实图像，如细节部分，本文算法更清晰地还原了高塔墙砖部分的纹理细节。表明了本文算法的融合结果具有优越的视觉效果，重建图像的质量更高，算法性能更好。

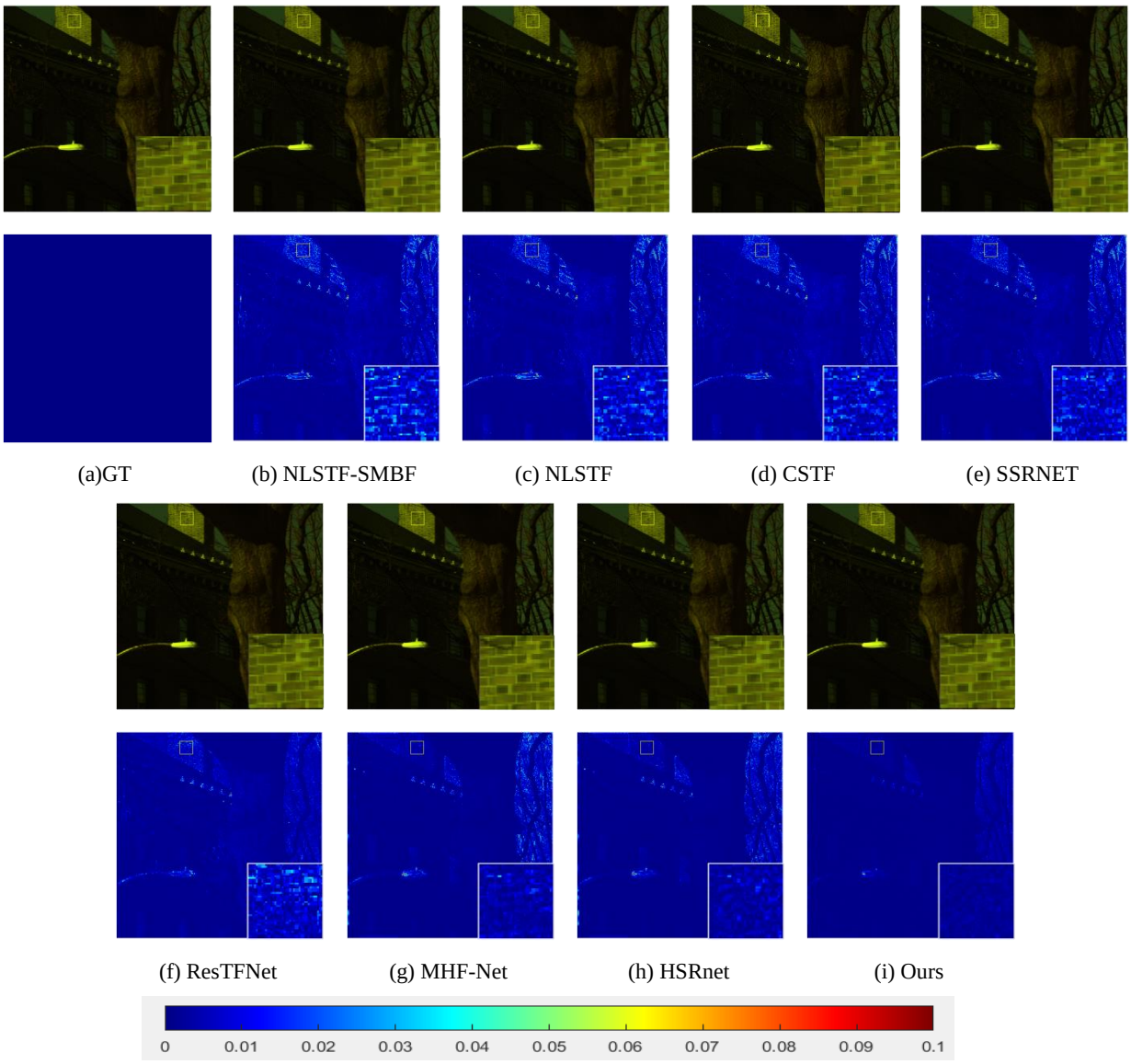


图 9 Harvard 数据集重建结果与误差图

Fig.9 Harvard dataset reconstruction results and error images

3.6 PU 数据集实验结果

PU 测试集共 9 张测试图像。各融合算法的客观评价指标平均值如表 3 所示，其中红色标注最优，蓝色次之。可以看出，本文所提算法的客观评价指标均最优，PSNR 值比次优值高 0.9 dB。这表明本文算法适用于遥感数据。SAM 和 ERGAS 与对比算法相比也取得了最优值。表明本文算法在重建图像质量方面有着绝对的优势。此外，可以看出基于深度学习算法的评价指标均高于传统类方法，这表明基于深度学习的方法相比于传统类方法更适应于结构复杂的遥感图像融合。同时本文根据 PU 测试集 9 张图像的 PSNR 值绘制雷达图。如图 10 所示，从图中可以看出本文算法和 HSRnet 都取得了不错的效果，但本文的算法比 HSRnet 更优秀。并且本文算法的雷达图接近于圆形，这表明它是稳定的，并且在遥感数据上具有泛化能力。

为了更直观地证明本文方法的有效性，本文对 PU 数据集上的融合结果进行了可视化。如图 11 所示，包含各算法融合结果与真实图像的伪彩色图像（102-60-30 波段）、重建图像与真实图像第 60 个波段的误

表 3 不同方法在 PU 数据集上的融合结果

Table 3 Fusion results of different methods on PU data set

Methods	PU (Pavia University)			
	PSNR	SAM	ERGAS	SSIM
CSTF	45.87	2.10	1.25	0.9824
NLSTF	46.41	1.97	1.20	0.9831
NLSTF-SMBF	45.23	2.12	1.34	0.9809
SSRNET	47.25	1.43	1.16	0.9918
ResTFNet	48.44	1.98	0.90	0.9940
MHF-Net	48.83	1.40	0.885	0.9943
HSRnet	49.18	1.22	0.74	0.9954
Ours	50.08	0.989	0.56	0.9963

Note: Red font represents the best; Blue font represents suboptimal.

差图。从彩色图像以及细节放大可以看出各类算法都取得了不错的重建结果，当结合误差图进行观测时可以看出，本文方法和 HSRnet 重建结果与原始图像的误差更小，更接近真实图像。同时对比误差图细节部分，可以看出本文方法相较于 HSRnet 在细节重构方面表现更优。本文方法对高光谱与多光谱图像融合任务，有着更优越的性能，充分保留了高光谱的光谱信息和多光谱的空间纹理细节。

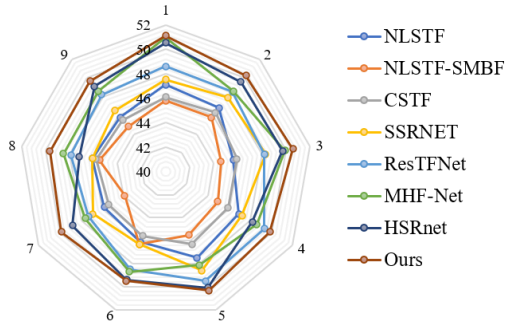
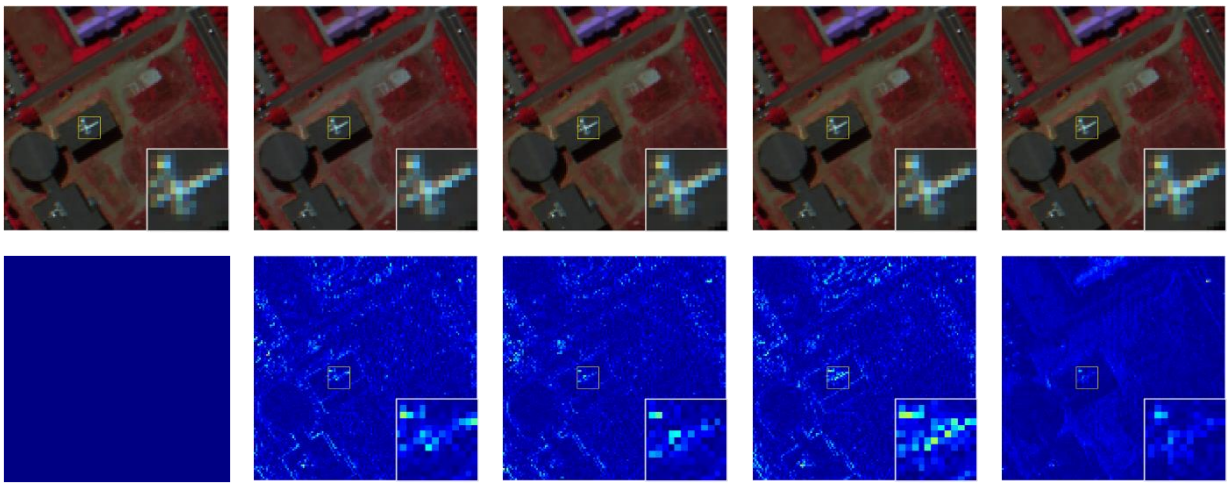
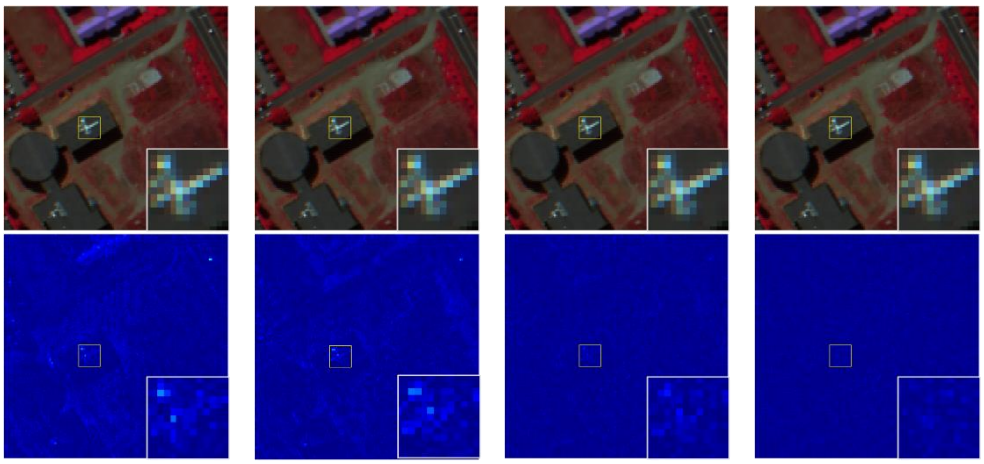


图 10 PU 数据集 9 张测试图片的 PSNR 值

Fig.10 PSNR values of 9 test images in the PU dataset



(a)GT (b) NLSTF-SMBF (c) NLSTF (d) CSTF (e) SSRNET



(f) ResTFNet (g) MHF-Net (h) HSRnet (i) Ours

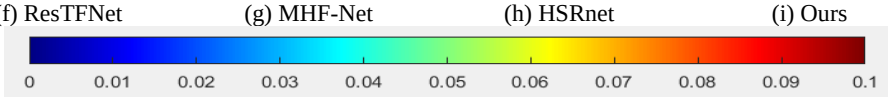


图 11 PU 数据集重建结果与误差图

Fig.11 PU dataset reconstruction results and error images

3.7 消融实验

本文方法主要分为特征提取阶段和特征融合阶段。为了验证这两个阶段中模块的有效性，我们对其进行消融实验。并在 CAVE 数据集上进行训练和测试。各项实验参数设置保持不变。

首先，为测试特征提取阶段的多层次卷积模块和残差注意力模块的有效性，对特征提取阶段单独实验，将提取特征进行通道拼接后，输入卷积层进行特征融合输出。同样地，为验证融合阶段多尺度 transformer 的有效性，对融合网络进行单独实验，将高光谱图像进行上采样保持与多光谱图像尺寸相同后，进行通道拼接，输入融合网络进行输出。如表 4 所示，经过特征提取后通过卷积层进行融合，也能取得不错的融合效果。但与仅使用多尺度 transformer 后通过卷积层相比，PSNR 低了 2.04 dB 说明了多尺度 transformer 融合网络的重要性。同时通过对比可以看出，整体网络可以进一步提高融合效果，PSNR 提高了 1.05 dB，从而验证了特征提取阶段模块的有效性，表明本文各阶段模块设计的合理性和有效性。

此外，基于 transformer 所设计的 Fusformer 融合网络[31]在高光谱与多光谱图像融合任务中有着不错的性能。为验证本文所提出的多尺度 transformer 融合网络的优势，将其作为消融实验中的对比网络。同时，两个网络采用本文所提出的特征提取网络来初始化图像特征，从实验结果可以看出，本文的多尺度 transformer 融合网络相比于基于 transformer 的 Fusformer 融合网络，通过对各尺度图像特征进行长距离相关性建模，能更好地融合多光谱的空间信息和高光谱的光谱信息。

表 4 网络模块有效性分析

Table 4 Analysis of the effectiveness of network modules

Feature Extractor	Fusion network	PSNR/dB	SAM	ERGAS	SSIM
√	×（multi-scale Transformer）	46.48	2.13	1.05	0.9959
×	√（multi-scale Transformer）	48.52	1.94	0.89	0.9978
√	√（Fusformer）	48.76	1.89	0.82	0.9981
√	√（multi-scale Transformer）	49.57	1.77	0.67	0.9986

4 结论

本文提出的网络，结合了 CNN 的特征提取能力与 transformer 长距离相关性建模优势。网络分为特征提取和特征融合两个阶段。特征提取阶段考虑到图像间所蕴含的信息特征，基于卷积神经网络分别设计了多层次卷积模块提取多光谱空间信息，残差通道注意力模块用于关注光谱信息的提取。融合阶段设计了多尺度 transformer 融合模块，通过 transformer 的自注意力机制实现对特征间的长距离相关性建模，同时防止局部特征的丢失，采用金字塔结构，在探索信息间长距离依赖关系时，能同时关注细节与语义信息，保证特征融合的质量。通过各项实验与分析表明，本文提出的高光谱与多光谱融合网络能更好地融合多光谱的空间信息和高光谱的光谱信息，获得更为全面和准确的遥感图像。由于本文实验的结果通过训练大量 HR-MSI 和 LR-HSI 模拟数据所得。因此，后续为降低对训练数据的依赖，可以进一步对方法进行改进和探索，实现无监督的高光谱与多光谱图像融合网络。

本文的贡献如下：1）特征提取阶段，针对图像特性基于 CNN 分别设计了多层次卷积模块和残差通道注意力模块用于提取图像信息，有效地提取了多光谱的空间信息和高光谱的光谱信息。2）融合阶段通过多尺度 transformer 融合模块，从局部到全局建立特征间的长程依赖关系，实现特征间的有效融合。3）通过在不同的数据集进行实验和对比分析，结果显示，无论是从指标评价还是视觉效果上，本文的方法与其他方法相比都取得了最好的结果，表明了本文方法在提高高光谱图像空间分辨率的同时有效保留了光谱信息。

参考文献：

[1] YANG Q, XU Y, WU Z, et al. Hyperspectral And Multispectral Image Fusion Based On Deep Attention Network[C]//*Proceedings of the* 2019 10th *Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing* (WHISPERS), 2019: 1-5.

[2] 童庆禧, 张兵, 张立福. 中国高光谱遥感的前沿进展[J]. **遥感学报**, 2016, **20**(5): 689-707.
TONG Qingxi, ZHANG Bing, ZHANG Lifu. Current progress of hyperspectral remote sensing in China[J]. *Journal of Remote Sensing*, 2016, **20**(5): 689-707.

[3] Akbari H, Kosugi Y, Kojima K, et al. Detection and analysis of the intestinal ischemia using visible and invisible hyperspectral imaging[J]. *IEEE Transactions on Biomedical Engineering*, 2010, **57**(8): 2011-2017.

[4] Zhihong P, Healey G, Prasad M, et al. Face recognition in hyperspectral images[J]. *IEEE Transactions on Pattern Analysis and Machine*

Intelligence, 2003, **25**(12): 1552-1560.

[5] DIAN R, LI S, SUN B, et al. Recent advances and new guidelines on hyperspectral and multispectral image fusion[J]. *Information Fusion*, 2021, **69**: 40-51.

[6] Fasbender D, Radoux J, Bogaert P. Bayesian data fusion for adaptable image pansharpening[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2008, **46**(6): 1847-1857.

[7] CHEN Z, PU H, WANG B, et al. Fusion of hyperspectral and multispectral images: a novel framework based on generalization of pan-sharpening methods[J]. *IEEE Geoscience and Remote Sensing Letters*, 2014, **11**(8): 1418-1422.

[8] SELVA M, AIAZZI B, BUTERA F, et al. Hyper-sharpening of hyperspectral data: A first approach[C]//*Proceedings of the* 2014 6th *Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing* (WHISPERS), 2014: 24-27.

[9] Zurita-Milla R, Clevers J G P W, Schaepman M E. Unmixing-based landsat TM and MERIS FR data fusion[J]. *IEEE Geoscience and Remote Sensing Letters*, 2008, **5**(3): 453-457.

[10] Yokoya N, Yairi T, Iwasaki A. Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2012, **50**(2): 528-537.

[11] Lanaras C, Baltsavias E, Schindler K. Hyperspectral super-resolution by coupled spectral unmixing[C]// *Proceedings of the* 2015 *IEEE International Conference on Computer Vision* (ICCV), 2015: 3586-3594.

[12] WEI Q, Dobigeon N, Tourneret J Y. Fast fusion of multi-band images based on solving a sylvester equation[J]. *IEEE Transactions on Image Processing*, 2015, **24**(11): 4109-4121.

[13] Akhtar N, Shafait F, Mian A. Bayesian sparse representation for hyperspectral image super resolution[C]// *Proceedings of the* 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2015: 7-12.

[14] 孙佳敏, 宋慧慧. 基于 DWT 和生成对抗网络的高光谱多光谱图像融合[J]. **无线电工程**, 2021, **51**(12): 1434-1441.
SUN Jiamin, SONG Huihui. Hyperspectral and multispectral image fusion based on discrete wavelet transform and fenerative adversarial networks[J]. *Radio Engineering*, 2021, **51**(12): 1434-1441.

[15] DIAN R, FANG L, LI S. Hyperspectral image super-resolution via non-local sparse tensor factorization[C]//*Proceedings of the* 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2017: 21-26.

[16] DIAN R, LI S, FANG L, et al. Nonlocal sparse tensor factorization for semiblind hyperspectral and multispectral image fusion[J]. *IEEE Transactions on Cybernetics*, 2020, **50**(10): 4469-4480.

[17] LI S, DIAN R, FANG L, et al. Fusing hyperspectral and multispectral images via coupled sparse tensor factorization[J]. *IEEE Transactions on Image Processing*, 2018, **27**(8): 4118-4130.

[18] Kanatsoulis C I, Fu X, Sidiropoulos N D, et al. Hyperspectral super-resolution: a coupled tensor factorization approach[J]. *IEEE Transactions on Signal Processing*, 2018, **66**(24): 6503-6517.

[19] LI J, ZHENG K, YAO J, et al. Deep unsupervised blind hyperspectral and multispectral data fusion[J]. *IEEE Geoscience and Remote Sensing Letters*, 2022, **19**: 1-5.

[20] WANG X, WANG X, ZHAO K, et al. FSL-Unet: full-scale linked unet with spatial–spectral joint perceptual attention for hyperspectral and multispectral image fusion[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, **60**: 1-14.

[21] DONG M, LI W, LIANG X, et al. MDCNN: multispectral pansharpening based on a multiscale dilated convolutional neural network[J]. *Journal of Applied Remote Sensing*, 2021, **15**: 036516.

[22] Benzenati T, Kessentini Y, Kallel A. Pansharpening approach via two-stream detail injection based on relativistic generative adversarial networks[J]. *Expert Systems with Applications*, 2022, **188**: 115996.

[23] FU Y, XU T, WU X, et al. PPT Fusion: pyramid patch transformerfor a case study in image fusion[J]. *arXiv preprint arXiv*: 2107.13967, 2021.

[24] GAO S H, CHENG M M, ZHAO K, et al. Res2Net: a new multi-scale backbone architecture[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, **43**(2): 652-662.

[25] LIU Z, LIN Y, CAO Y, et al. Swin Transformer: hierarchical vision transformer using shifted windows[C]//*Proceedings of the* 2021 *IEEE/CVF International Conference on Computer Vision* (ICCV), 2021: 10-17.

[26] Kingma D P, BA J. Adam: a method for stochastic optimization[J/OL]. arXiv:1412.6980, 2014, https://doi. org/10.48550/arXiv.1412.6980.

[27] ZHANG X, HUANG W, WANG Q, et al. SSR-NET: spatial–spectral reconstruction network for hyperspectral and multispectral image fusion[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2021, **59**(7): 5953-5965.

[28] LIU X, LIU Q, WANG Y. Remote sensing image fusion based on two-stream fusion network[J]. *Information Fusion*, 2020, **55**: 1-15.

[29] XIE Q, ZHOU M, ZHAO Q, et al. MHF-Net: an interpretable deep network for multispectral and hyperspectral image fusion[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, **44**(3): 1457-1473.

[30] HU J F, HUANG T Z, DENG L J, et al. Hyperspectral image super-resolution via deep spatiospectral attention convolutional neural networks [J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2022, **33**(12): 7251-7265.

[31] HU J F, HUANG T Z, DENG L J, et al. Fusformer: a transformer-based fusion network for hyperspectral image super-resolution[J]. *IEEE Geoscience and Remote Sensing Letters*, 2022, **19**: 1-5.