引用格式：LI Zitong, ZHAO Jiankang, XU Jingran, et al. Remote Sensing Image Fusion Method Based on Improved Swin Transformer[J]. Acta Photonica Sinica, 2023, 52(11):1110001
李紫桐,赵健康,徐静冉,等.基于改进Swin transformer的遥感图像融合方法[J].光子学报,2023,52(11):1110001

基于改进Swin transformer的遥感图像融合方法

李紫桐,赵健康,徐静冉,龙海辉,刘传奇
(上海交通大学 电子信息与电气工程学院 感知科学与工程学院,上海 200240)

摘　要:针对现有基于transformer的方法未能充分融合遥感图像多尺度特征的问题,提出一种多光谱-全色融合网络。在融合网络中嵌入一个基于改进Swin transformer的多尺度窗口自注意力模块,在关注全局空间特征的同时,充分融合不同尺寸的特征信息,从而最大程度地保留光谱和空间结构信息。通过不同层级特征的跳跃连接,解码网络预测出原始多光谱图像缺失的纹理部分,最终使用细节注入模型恢复出目标图像。为了提升融合效果,在损失函数中加入了光谱损失和空间结构损失。与其他方法相比,本文提出的方法在WorldView-4、QuickBird和WorldView-2三种卫星数据集的主观视觉效果最好,相比于性能第二的方法,本文方法在三种数据集的相对全局误差指标分别减小了11.99%、0.4%和3.43%。

关键词:遥感;图像融合;多光谱图像;全色图像;Swin transformer
中图分类号:TP751　　**文献标识码**:A　　**doi**:10.3788/gzxb20235211.1110001

0　引言

遥感图像广泛应用于土地监测、环境感知、灾害预测和城市分析等工作。大部分商用卫星都同时搭载可以获取全色图像和多光谱图像的传感器[1],全色(Panchromatic,PAN)图像是二维的灰度图像,具有较高的空间分辨率,多光谱(Multispectral Image,MS)图像的波段数一般大于等于四个,但是由于设备的带宽限制,成像的空间分辨率较低[2]。将星载成像系统捕获到的MS图像和PAN图像进行融合生成高空间分辨率多光谱(High Resolution Multispectral,HRMS)图像,这一过程即多光谱-全色图像融合,也称全色锐化[3]。二者融合之后得到同时具有高空间分辨率和高光谱分辨率的遥感影像,可以获取被测对象更准确的细节,从而推动农业、环保等各个领域的进步。

多光谱-全色图像的融合方法可以分为传统方法和深度学习方法两类。传统的多光谱-全色图像融合方法可以分为多分辨率分析方法(Multi-Resolution Analysis, MRA),成分替换方法(Component Substitution, CS)和变分模型方法(Variational Optimization,VO)[4]。MRA方法把理想的融合图像看成金字塔的最顶层,通过建立两幅源图像的金字塔不同层级之间的关系,从PAN图像推导出多光谱图像缺失的细节信息并生成HRMS图像。根据尺度变换函数的不同,MRA方法包含调制传递函数广义拉普拉斯金字塔(Modulation Transfer Function Generalized Laplacian Pyramid,MTF-GLP)[5]、小波变换(Wavelet)[6]等方法,MRA方法可以对图像光谱信息进行很好地保留,但经常会出现空间变形的情况。CS方法先将多光谱图像映射到某一空间之后,再用全色图像来替换其中的空间分量,通过反向变换即可得到锐化的波段。典型的成分替换方法包含主成分分析(Principal Component Analysis,PCA)[7]、亮度-色度-饱和度变换(Intensity-Hue Saturation, IHS)[8]等方法,CS方法没有对全色图像进行空间变换,所以空间结构保留得比较完整,但是相应地,光谱扭曲的现象很严重。BALLESTER C等[9]提出一种多光谱-全色变分优化方法,该方法依赖于遥感图像的先验知识,将图像融合问题转化为最优化的求解问题,但是该方法对观测模型正则化参数的计算十分复杂。

基金项目:国家自然科学基金(No. 62171283),上海商用飞机系统工程联合研究基金(No. CASEF-2022-MQ01)
第一作者:李紫桐,lizitong@sjtu.edu.cn
通讯作者:赵健康,zhaojiankang@sjtu.edu.cn
收稿日期:2023-05-08;**录用日期**:2023-06-26

相比于传统方法，深度学习方法具有特征提取能力强、识别精度高等优点，因此被广泛应用在多光谱和全色图像的融合中。MASI G等[10]最早提出基于卷积神经网络的全色锐化方法(Pansharpening by Convolutional Neural Networks，PNN)，它的整体网络结构比较简单，只包含三层的卷积结构。PanNet[11]增加了网络的深度，为了实现光谱信息的保留，直接将学习到的高频信息注入上采样后的多光谱图像中。多尺度多深度卷积神经网络(A Multiscale and Multidepth Convolutional Neural Network，MSDCNN)[12]将多尺度特征提取和残差学习引入卷积神经网络中，使用不同大小的卷积核来充分提取图像空间信息。

上述方法都是基于卷积神经网络实现，卷积核主要关注图像的局部特征，忽略了全局特征。而遥感图像通常由大量相似地物组成，这些相似地物在进行高分辨率重建时可以互相弥补缺失的信息，增强图像的特征表示，因此充分利用遥感图像的全局特征是十分必要的。基于自注意力机制的transformer可以捕获上下文之间的全局交互，DOSOVITSKIY A等[13]提出的Vision Transformers最先将transformer应用于视觉领域，将输入图像分成固定的块，再将每个块投影为固定长度的向量，计算这些块的全局相关性。LIU Z等[14]提出的Swin transformer在Vision Transformers的基础上进行改进，将全局注意力机制转换为局部注意力机制，并通过窗口偏移的操作建立长距离依赖，不仅减少了计算量，还在视觉任务中取得更好的性能。ZHOU H等[15]最先将Swin transformer结构作为全色锐化网络的主干模块，提出的Panformer网络表现出了比基于卷积神经网络的模型更好的性能，FAN Wensheng等[16]也提出了基于Swin transformer的双分支U形融合网络，充分地利用了图像的全局上下文特征，但这两种方法只是对Swin transformer的简单使用，并没有调整其结构以更好地适应复杂地面场景的遥感图像融合任务。

本文提出一种基于多尺度窗口自注意力的多光谱-全色融合方法(Multiscale Swin-transformer with Channel Attention NetWork，MSCANet)，在网络的融合部分集成了一种新的即插即用模块：多尺度窗口注意力单元(Multiscale Swin-transformer with Channel Attention，MSCA)。该单元将Swin transformer输出部分的多层感知器(Multi Layer Perception，MLP)替换成多尺度卷积核和通道注意力的级联模块，在利用区域之间的长程依赖的同时，更好地融合遥感图像不同尺寸地物的特征信息，从而进一步提升融合结果。MSCANet基于细节注入模型，融合网络不是直接预测HRMS图像，而是专注于预测多光谱图像丢失的高频细节，再将高频细节与原始图像相加得到高分辨率多光谱图像。

1 网络结构

网络整体结构如图1，输入图像分别为低空间分辨率多光谱(Low Resolution Multispectral，LRMS)图像和PAN图像，采用结构相同但权重不同的双流结构分别对两幅源图像进行特征提取，特征提取使用到的卷

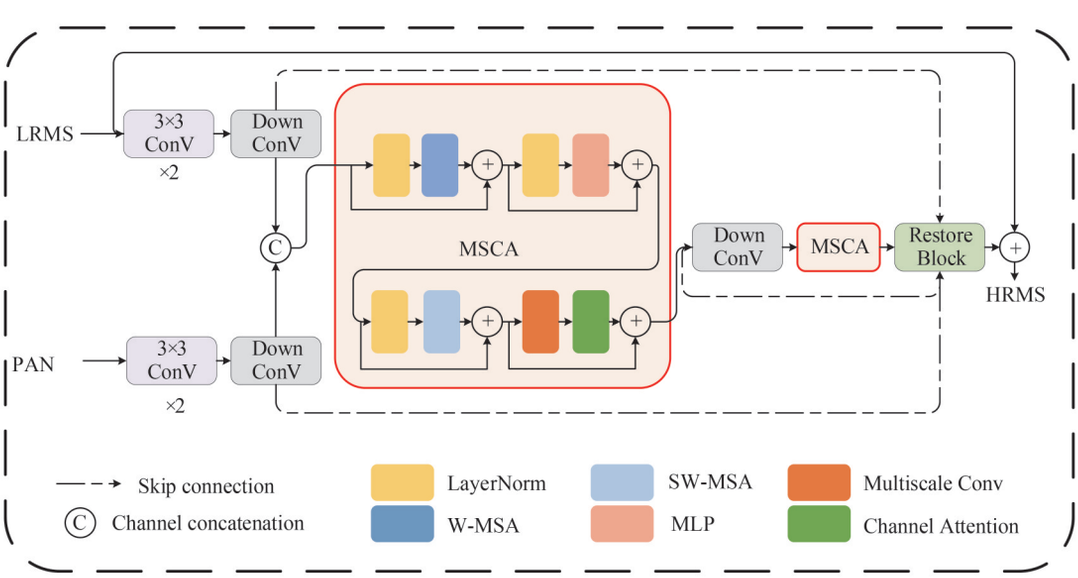

图1 网络总体结构

Fig.1 Overall network structure

积核大小为3×3。使用大小为2×2、步长为2的卷积核将特征图像的长和宽下采样为原来的1/2,下采样的特征经过通道维的拼接之后送入融合模块。在融合模块嵌入了MSCA模块,用于融合多光谱和全色图像的多尺度空间信息,并加强全局特征的交互,最后将MSCA输出的高层特征与前序的低层特征通过跳跃连接送入恢复模块(Restore Block),恢复模块输出的是多光谱图像缺失的高频信息,将该高频图像注入输入的LRMS图像中即可生成目标的HRMS图像。

1.1 细节注入模型

MRA方法采用细节注入的方法来提高多光谱图像的空间分辨率,如式(1)。原始的多光谱图像用$\boldsymbol{M}\in\boldsymbol{R}^{H\times W\times C}$表示,四倍上采样之后的多光谱图像为$\boldsymbol{M}_{\mathrm{up}}\in\boldsymbol{R}^{4H\times 4W\times C}$,全色图像用$\boldsymbol{P}\in\boldsymbol{R}^{4H\times 4W\times 1}$表示,假设要合成的高分辨率多光谱图像$\boldsymbol{M}_{\mathrm{HS}}\in\boldsymbol{R}^{4H\times 4W\times C}$包含低频分量和高频分量,第$k$个波段的多光谱图像缺少的高频细节信息可以由全色图像推断出,因此将低频分量$\boldsymbol{M}_{\mathrm{up}}^{k}$与全色图像的高频细节直接相加,$g^{k}$代表第$k$个波段的细节注入权重,$\boldsymbol{P}$和$\boldsymbol{P}_{\mathrm{L}}$分别表示全色图像和它的低通滤波部分,$\boldsymbol{P}-\boldsymbol{P}_{\mathrm{L}}$为全色图像的高频部分,向$\boldsymbol{M}_{\mathrm{up}}$注入的缺失信息得到融合图像。

$$\boldsymbol{M}_{\mathrm{HS}}^{k}=\boldsymbol{M}_{\mathrm{up}}^{k}+g^{k}\left(\boldsymbol{P}-\boldsymbol{P}_{\mathrm{L}}\right)\qquad k=1,\cdots,C \tag{1}$$

细节注入模型通过计算缺失的细节信息,在不影响光谱结构的同时最大程度提升空间分辨率,实现有效的光谱保真[17],可以有效地用于多光谱图像和全色图像的融合。但式(1)只是将全色图像的高频部分与细节注入权重直接相乘的结果作为每个波段需要注入的细节信息,而多光谱各个波段的光谱响应曲线常常是非线性的,且波段之间存在重叠现象。为了建立准确的融合模型,本文保留了MRA方法的细节注入模型的思想,使用深度学习建模端对端非线性模型,如图2。图中$F\left(\boldsymbol{M}_{\mathrm{up}},\boldsymbol{P};\theta\right)$代表以$\theta$为参数,输入为$\boldsymbol{M}_{\mathrm{up}}$和$\boldsymbol{P}$的深度学习融合模型,模型可以更好地拟合多光谱图像丢失的高频细节,最终输出的高频细节和原始图像$\boldsymbol{M}_{\mathrm{up}}^{k}$直接相加,其计算表达式为

$$\boldsymbol{M}_{\mathrm{HS}}^{k}=\boldsymbol{M}_{\mathrm{up}}^{k}+F\left(\boldsymbol{M}_{\mathrm{up}},\boldsymbol{P};\theta\right)^{k}\qquad k=1,\cdots,C \tag{2}$$

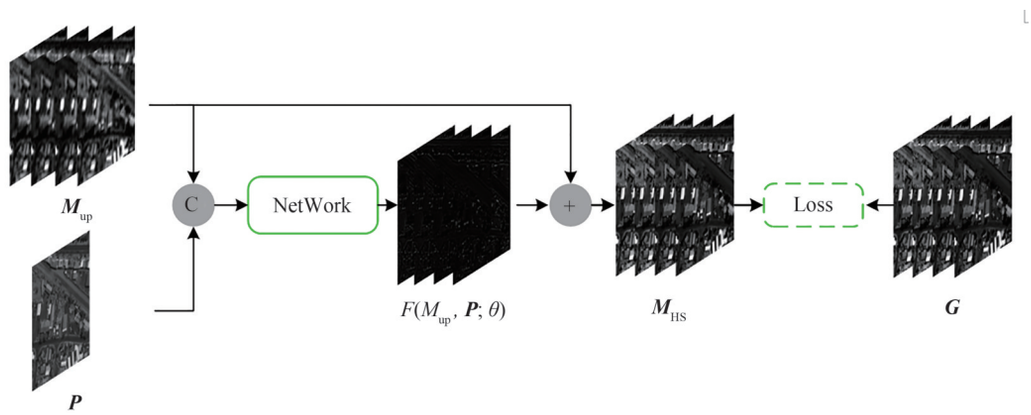

图2 细节注入模型

Fig.2 Detail injection model

1.2 多尺度窗口注意力单元MSCA

在融合阶段,将全色图像和多光谱图像的特征在通道维进行拼接,从而实现光谱和空间信息的整合。融合网络由多尺度窗口注意力单元MSCA和特征下采样结构组成。MSCA是对Swin transformer的改进,通常一个Swin transformer由两个级联的滑动窗口transformer块(Swin Transformer Block,STB)组成,第一个STB由窗口多头自注意力模块(Window Multihead Self Attention,W-MSA)和MLP级联构成,第二个STB则由滑动窗口多头自注意力模块(Shift Window Multihead Self Attention,SW-MSA)和MLP组成。MSCA保留了第一个STB结构,并将第二个STB中的MLP替换成多尺度通道注意力模块,在减小计算量的同时捕获了不同尺度的图像特征。

1.2.1 W-MSA

对于尺寸为$H\times W\times C$的输入图像,W-MSA模块将输入划分成尺寸为$M\times M$的互不重叠的窗口,然

后在每个窗口中计算自注意力，窗口数量为$\frac{HW}{M^2}$。对于局部特征窗口$\boldsymbol{X}\in\boldsymbol{R}^{M^2\times C}$，将图像块通过线性嵌入(Linear Embedding)转换成长度为$C\times 1$的序列，序列个数为M^2，对序列进行相关性的运算，得到$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$向量，分别代表查询矩阵、匹配矩阵和信息矩阵。可学习的相对位置编码表示为$\boldsymbol{B}$，利用多头注意力机制学习到不同子空间的信息，自注意力的计算表达式为

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{SoftMax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{\boldsymbol{C}}}+\boldsymbol{B}\right)\boldsymbol{V} \tag{3}$$

接下来使用MLP层来对特征$\boldsymbol{X}$进行特征变换，为了易于训练，在MSA和MLP模块之前加入层归一化(LayerNorm，LN)层和残差连接结构，如式(4)，最终输出的特征图尺寸为$\frac{HW}{M^2}\times M^2\times C$，将特征重新排列成$H\times W\times C$，并送入第二个STB模块。

$$\begin{aligned}\boldsymbol{X}&=\mathrm{WMSA}\big(\mathrm{LN}(\boldsymbol{X})\big)+\boldsymbol{X}\\ \boldsymbol{X}&=\mathrm{MLP}\big(\mathrm{LN}(\boldsymbol{X})\big)+\boldsymbol{X}\end{aligned} \tag{4}$$

1.2.2 多尺度通道注意力模块

Swin transformer的第二个STB模块包含SW-MSA和MLP部分，其中SW-MSA通过窗口的移动解决了不同窗口无法进行信息交流的问题，MLP以像素点为单位进行计算，用全连接的方式进行全局特征交互，却忽略了对不同大小地物(如建筑物、植被、河流、车道等)的特征融合。受到LI B等[18]在人体行为识别网络中将transformer中的MLP模块替换成多尺度卷积这一方法的启发，本文将第二个STB的MLP部分替换成多尺度通道注意力机制模块，从而整合和利用不同尺度的全局信息，如图3。不同大小的卷积核可以获取不同大小感受野的特征，在多尺度卷积之后建立特征通道间的依赖关系，增强高频空间特征和光谱分布特征的通道权重，削弱其他不重要的通道权重。

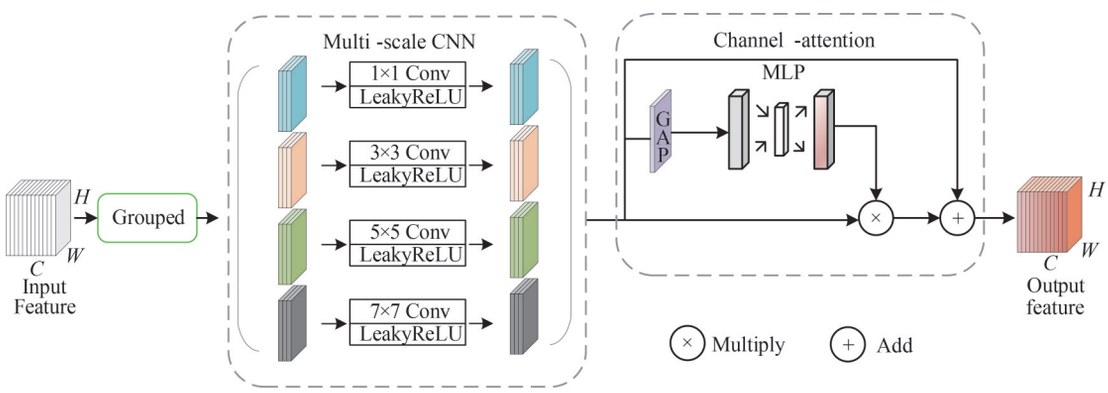

图3 多尺度通道注意力模块

Fig.3 Multi-scale CNN and channel attention module

不同于MSDCNN[12]中的多尺度卷积结构，为了节约参数，本文使用分组卷积来替代常规卷积，将尺寸为$H\times W\times C$的输入特征图按通道维平均分为4组，每组的特征尺寸为$H\times W\times\frac{C}{4}$，然后对每组分别进行卷积运算，卷积核尺寸为1×1、3×3、5×5、7×7，激活函数为LeakyReLU。为了保留有益的空间和光谱信息，MSCANet在多尺度卷积块之后加入通道注意力机制。首先将级联后尺寸为$H\times W\times C$的特征图进行全局平均池化，得到$1\times 1\times C$的权重向量；然后依次通过两个全连接层，第一次全连接的神经元数目为$\frac{C}{4}$，第二次全连接的神经元数目为C，再通过sigmoid函数将权重向量的值固定在$[0,1]$之间；最终将权重和输入特征相乘，在通道注意力的基础上加入残差连接，帮助融合网络快速收敛。

1.3 特征重建网络

获取到融合图像的特征图之后，使用转置卷积(Transpose Conv)操作对其进行逐级上采样并恢复其空

间分辨率。由于直接从高层特征恢复纹理信息是困难的，本文借鉴了U-Net[19]的思想，通过跳跃连接（Skip Connection）降低从高层特征中恢复细节纹理的难度。在每个上采样操作之后，将前序特征与当前特征图进行拼接，有利于恢复真实的空间细节。在网络的结尾使用卷积核和Tanh激活函数层来重建输出的高分辨率多光谱图像，如图4。

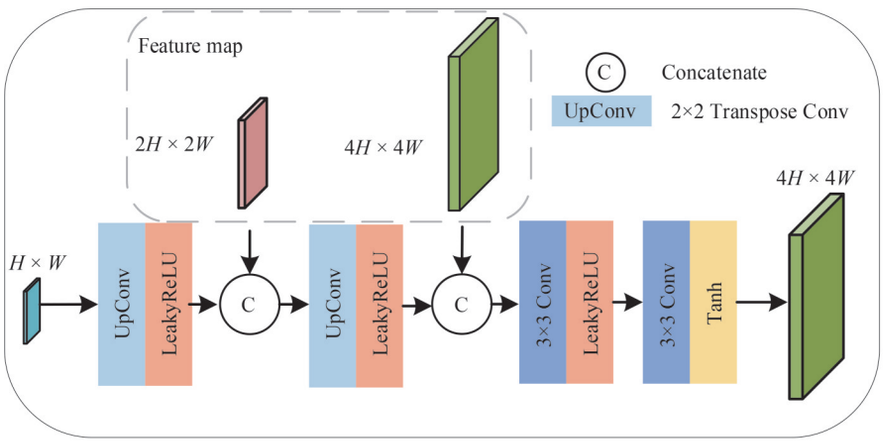

图4　特征重建网络结构

Fig.4　Structure of feature reconstruction network

2　损失函数

常见的多光谱-全色融合方法采用平均绝对误差（Mean Absolute Error，MAE）损失函数对网络参数进行优化，相比于均方误差（Mean Square Error，MSE），MAE可以更好地减少回归问题引起的平滑伪影，MAE损失表示为式（5），其中$\hat{\boldsymbol{P}}$为模型预测得到的融合图像，$\boldsymbol{G}$为参考图像，$\|\cdot\|_1$代表一范数运算。

$$L_{\mathrm{MAE}}=\|\hat{\boldsymbol{P}}-\boldsymbol{G}\|_1 \tag{5}$$

MAE损失虽然可以对融合图像进行约束，但是缺乏对光谱波段的关系运算和空间结构的保真运算。本文在MAE损失的基础上分别加入空间结构损失和光谱损失，其中光谱损失如式（6），参考了SAM的计算公式，用于约束光谱损失，$\langle\cdot,\cdot\rangle$表示内积的运算。

$$L_{\mathrm{Spectral}}=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\left|\frac{\left\langle\hat{\boldsymbol{P}}(i,j),\boldsymbol{G}(i,j)\right\rangle}{/\!/\hat{\boldsymbol{P}}(i,j)/\!/_2\times/\!/\boldsymbol{G}(i,j)/\!/_2+1\times10^{-6}}-1\right| \tag{6}$$

为了最大程度地进行空间信息的保留，空间结构损失的计算式参考SSIM公式，如式（7）、（8），其中μ_x和μ_y代表局部像素强度的平均值，σ_x和σ_y代表局部标准差，C_1和C_2是防止损失无穷大的平衡参数。

$$\mathrm{SSIM}(\hat{\boldsymbol{P}},\boldsymbol{G})=\frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)} \tag{7}$$

$$L_{\mathrm{Spatial}}=1-\mathrm{SSIM}(\hat{\boldsymbol{P}},\boldsymbol{G}) \tag{8}$$

总损失L的计算如式（9），它是MAE损失、光谱损失和空间结构损失的线性组合，光谱损失和空间结构损失的权重系数分别为α和β，通过最小化L来训练本文提出的融合模型。

$$L=\alpha L_{\mathrm{Spectral}}+\beta L_{\mathrm{Spatial}}+L_{\mathrm{MAE}} \tag{9}$$

3　实验结果与分析

3.1　数据集

实验使用公开的标准数据集NBU_PansharpRSData[20]，选取其中的WorldView-4（WV4）、QuickBird（QB）以及WorldView-2（WV2）卫星数据。每个卫星数据包含500对多光谱-全色图像，其中原始多光谱图像的尺寸为256×256×C，WV4卫星和QB卫星包含四个波段，$C=4$，WV2卫星包含八个波段，$C=8$，原

始全色图像的尺寸为1 024×1 024×1。

按照9∶1的比例划分训练集和测试集，将其中的50张图像用于测试，450张图像用于训练。由于缺少真实的高分辨率多光谱参考图像，在训练时把原始多光谱图像作为参考的HRMS图像，并采用Walds协议，根据卫星的MTF函数分别对原始多光谱图像和原始全色图像进行高斯滤波和四倍的下采样，得到LRMS图像和PAN图像。为了扩充训练的数据集，还要对HRMS图像、LRMS图像和PAN图像进行裁剪，经过滤波之后的PAN图像的尺寸为256×256×1，对每张PAN图像按照32像素的重叠从左至右，从上至下依次进行裁剪，裁剪得到的每个图像块的尺寸为64×64×1，因此一张原始尺寸为256×256×1的PAN图像可以裁剪成49张尺寸为64×64×1的PAN训练图像。MS图像的裁剪方式与之相同，重叠像素设置为8，图像块大小为16×16×C。最终得到的训练集包含22 000对空间尺寸分别为16×16×C和64×64×1的多光谱-全色图像对，参考图像的尺寸为64×64×C。全分辨率测试集则直接对原始的LRMS图像和PAN图像进行融合，没有参考的HRMS图像，数据集的具体信息如表1。

表1 数据集的具体信息

Table 1 Specific information about the dataset

		Training dataset		Testing dataset(reduced resolution)		Testing dataset(full resolution)	
		Number	Size	Number	Size	Number	Size
WV4	LRMS	22 000	16×16×4	50	64×64×4	50	256×256×4
	PAN	22 000	64×64×1	50	256×256×1	50	1 024×1 024×1
	HRMS	22 000	64×64×4	50	256×256×4	—	—
QB	LRMS	22 000	16×16×4	50	64×64×4	50	256×256×4
	PAN	22 000	64×64×1	50	256×256×1	50	1 024×1 024×1
	HRMS	22 000	64×64×4	50	256×256×4	—	—
WV2	LRMS	22 000	16×16×8	50	64×64×8	50	256×256×8
	PAN	22 000	64×64×1	50	256×256×1	50	1 024×1 024×1
	HRMS	22 000	64×64×8	50	256×256×8	—	—

3.2 实验设置

本模型基于PyTorch框架实现，使用的显卡型号为NVIDIA GeForce RTX 3090 Ti。设置的最大训练代数(Epoch)为200，批大小(Batch Size)为32。根据式(9)进行损失函数的计算，损失函数中的α和β均设置为0.1，并采用Adam优化器对模型参数进行优化。初始学习率为1×10^{-4}，每隔3 500个训练步数，学习率衰减为原来的0.99倍，训练时将图像数据归一化到$[-1,1]$。在训练过程中，输入的LRMS和PAN的尺寸分别为16×16×C和64×64×1，C为波段数目，网络输出的是融合图像归一化到$[-1,1]$之后的结果，融合图像的尺寸为64×64×C。在真实图像的测试阶段，使用训练好的模型参数，输入的LRMS和PAN的尺寸分别为256×256×C和1 024×1 024×1，输出的HRMS的尺寸为1 024×1 024×C。MSCANet的训练过程如下：

```
for i in epochs                                  第 i 个 epoch，最大 epoch 个数设为 200
    for j in batches                             第 j 个 batch
        Select 32 patches of PAN images;         选取 PAN 数据集的 32 张图像；
        Select 32 patches of LRMS images;        选取 LRMS 数据集的 32 张图像；
        Select 32 patches of HRMS images;        选取 HRMS 数据集的 32 张图像；
        Produce the output P̂ = f(PAN, LRMS);     计算模型生成的融合图像；
        Calculate the loss L ;                   计算融合图像和参考图像的损失函数 L；
        Update parameters by AdamOptimizer;      根据 L，利用 Adam 优化器更新模型的参数；
    end
end
```

为了验证提出方法的有效性，将MSCANet与9种常见方法进行比较，包括四种传统方法MTF-GLP[5]方法、Wavelet[6]方法、IHS[8]方法和PCA[7]方法以及五种深度学习方法FusionNet[21]、Panformer[15]、MSDCNN[12]、LAGConv[22]和TFNet[23]，其中Panformer基于Swin transformer实现，FusionNet、MSDCNN、LAGConv和TFNet基于卷积神经网络实现。所有对比方法都使用了原始的设定参数和相同的测试集，基于深度学习的方法使用了相同的训练集。

3.3 客观评价指标

在图像融合过程中，目标的高分辨率多光谱图像的真实值实际上不存在，因此通常采用的评价指标体系基于两种方式，分别是降分辨率（Reduced Resolution，RR）评价指标和全分辨率（Full Resolution，FR）评价指标。

降分辨率评价指标用于有参考影像的评估，将待融合的全色和多光谱图像分别下采样到更低分辨率的训练图像，将原始的多光谱图像看成真实值，并计算融合图像和真实值的差异。本文采用四种评价指标，分别是衡量融合综合性能的相对全局误差（Erreur Relative Global Adimensionnelle Synthesis，ERGAS）、图像峰值信噪比（Peak Signal to Noise Ratio，PSNR）、计算融合图像和参考图像之间光谱相似度的光谱角度（Spectral Angle Mapper，SAM）和计算融合图像和参考图像高频空间细节相似度的空间相关系数（Spatial Correlation Coefficient，SCC）。

全分辨率指标用于无参考影像的评估，使用真实的数据，将多光谱和全色图像在原始尺度进行融合，分别计算融合图像和两幅源图像的差异。本文采用无参考质量指标（Quality with No Reference index，QNR）以及它的光谱细节损失分量D_λ和空间细节损失分量D_S来定量评价全分辨率下的融合结果。

3.4 降低分辨率下的实验结果分析

在四波段卫星数据集WorldView-4、QuickBird以及八波段卫星数据集WorldView-2进行降低分辨率下的多光谱-全色融合实验，分别从主观视觉结果和客观评价指标ERGAS、SAM、PSNR、SCC给出各个方法的实验结果。

3.4.1 WV4数据集实验结果

WV4数据集下主观视觉比较结果如图5，MTF-GLP方法虽然恢复出了大部分的纹理信息，但是在放大区域泛白严重，而且整体较参考图片颜色更淡；Wavelet方法出现了明显的和细节模糊；IHS方法和PCA方法的空间清晰度很高，但是整体颜色与真值差距较大，无法很好地还原地物色彩。可以看出，五种基于深度学习的方法相比传统方法的主观视觉融合效果更好。

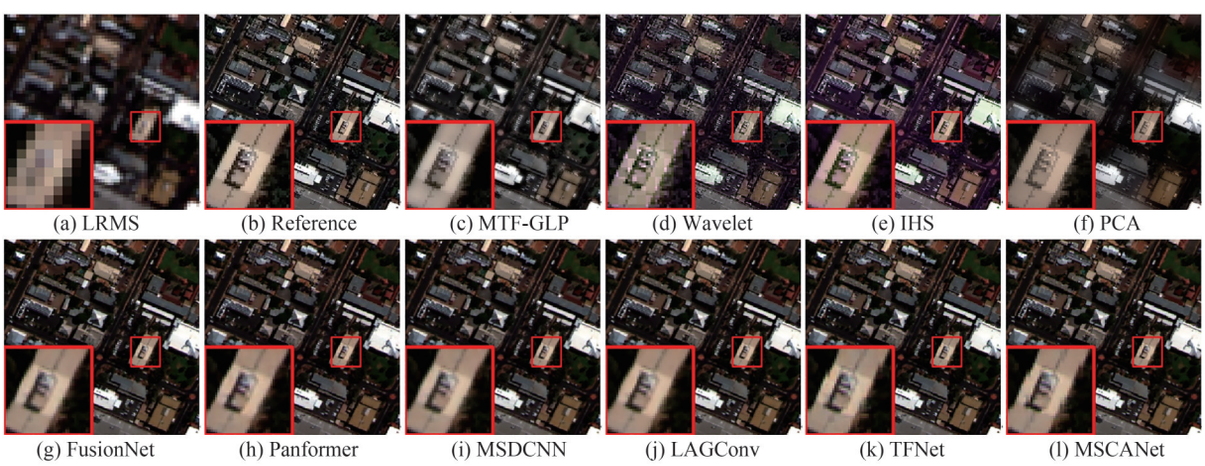

图5 WorldView-4仿真数据集融合结果

Fig.5 Fusion result of WorldView-4 simulation dataset

计算融合图像与参考图像在各个波段的的平均差值，结果如图6。FusionNet、Panformer、MSDCNN、LAGConv方法和TFNet的边缘轮廓失真比MSCANet更加明显，而MSCANet和真值图像具有更好的相似性，在主观视觉上具有最好的效果。

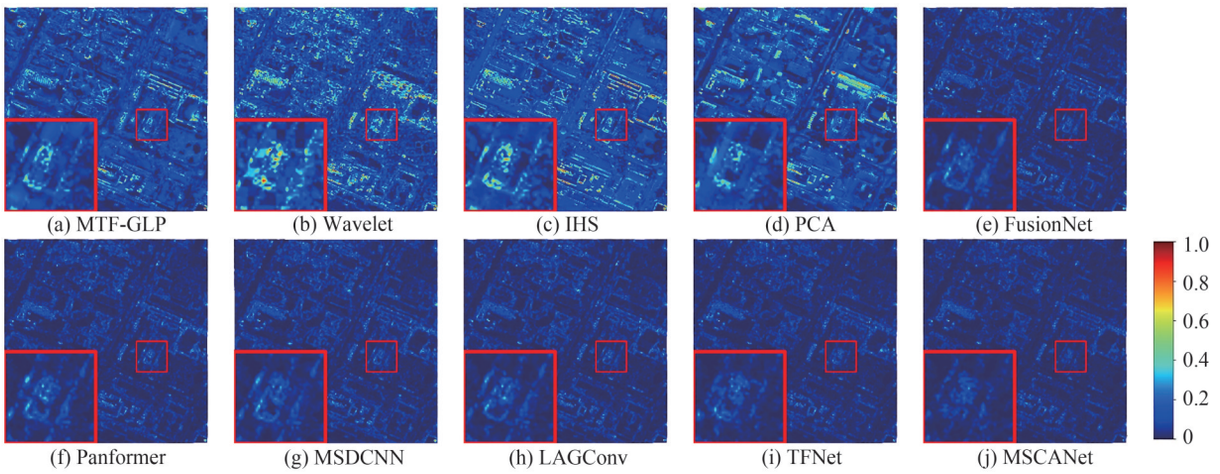

图 6 WorldView-4 仿真数据集的残差图

Fig.6 Residual graph of WorldView-4 simulation dataset

对客观评估结果进行分析，如表 2，其中传统方法中 IHS 的结果最好，MLP、Wavelet 和 PCA 方法的各项评估指标较深度学习方法有较大差距，证明传统方法无法对多光谱图像进行很好的光谱保真和空间细节增强。TFNet 在各种深度学习方法中性能第二，MSCANet 的各项指标均优于 TFNet，其中 SAM 减少了 9.12%，ERGAS 减少了 11.99%，PSNR 和 SCC 分别提升了 3.47% 和 0.41%。

表 2 仿真数据集的客观评价指标

Table 2 Objective evaluation index of simulation dataset

	WV4				QB				WV2			
Method	ERGAS ↓	SAM ↓	PSNR ↑	SCC ↑	ERGAS ↓	SAM ↓	PSNR ↑	SCC ↑	ERGAS ↓	SAM ↓	PSNR ↑	SCC ↑
MTF-GLP	6.340	5.772	23.524	0.914	2.698	2.334	37.271	0.857	6.338	7.699	26.891	0.878
Wavelet	6.425	6.460	23.401	0.864	4.316	2.981	32.160	0.660	6.703	8.435	26.096	0.845
PCA	6.505	7.337	23.326	0.878	2.981	3.162	36.584	0.792	7.881	8.842	25.081	0.828
IHS	5.661	5.394	24.486	0.902	2.826	2.573	36.048	0.723	6.454	7.780	26.628	0.876
MSDCNN	2.811	3.232	30.590	0.973	1.359	1.468	43.334	0.953	4.036	5.145	30.938	0.944
FusionNet	2.910	3.190	30.280	0.972	1.270	1.369	43.856	0.959	3.845	5.050	31.217	0.948
Panformer	2.820	3.170	30.677	0.975	1.251	1.362	44.077	0.961	3.888	5.013	31.229	0.948
LAGConv	2.693	3.110	30.956	0.976	1.272	1.406	43.813	0.958	3.878	5.070	31.140	0.947
TFNet	2.585	3.115	31.390	0.978	1.238	1.344	44.154	0.961	3.795	5.003	31.397	0.950
MSCANet	**2.275**	**2.831**	**32.478**	**0.982**	**1.233**	**1.310**	**44.202**	**0.962**	**3.665**	**4.869**	**31.691**	**0.953**

3.4.2 QuickBird 数据集实验结果

QuickBird 数据集下的主观效果如图 7，其中四种传统方法与真实值的颜色差距较大。图 8 为 QuickBird 数据集融合图像和参考图像在各个波段的平均残差图，FusionNet 和 MSDCNN 在建筑物的边缘轮廓与真值的偏差较大，Panformer、LAGConv 和 TFNet 在细节处的空间清晰度相对较高，但是在房屋和道路的分界区出现了一些失真，MSCANet 相比于其他方法在残差图中出现的亮点更少。对客观评估结果进行分析，其中 MTF-GLP 在传统方法中表现最好，但各项评估结果还是落后于深度学习方法，MSCANet 相比于排名第二的 TFNet 分别在 SAM 指标减少 2.53%，ERGAS 指标减少 0.4%，PSNR 和 SCC 指标分别提升了 0.11% 和 0.04%。

3.4.3 WorldView-2 数据集实验结果

WorldView-2 是八波段卫星，选取其中 RGB 波段显示主观视觉效果，如图 9。MTF-GLP 方法的图像颜色偏淡，Wavelet 和 IHS 方法对空间细节的恢复程度较好，但整体光谱扭曲较明显，蓝色建筑物和灰褐色道路的色彩和真值图像差距较大，PCA 方法的表现较差，几乎无法辨认房屋的边缘轮廓。

进一步分析波段平均残差图的结果，如图 10，可以看出 FusionNet、MSDCNN、LAGConv 在白色 U 形区域出现了明显的纹理失真，Panformer 和 TFNet 的效果相对较好但也出现了较多亮线，证明与真实图像存在一定的偏差，而 MSCANet 在各种方法中的主观视觉效果最好。

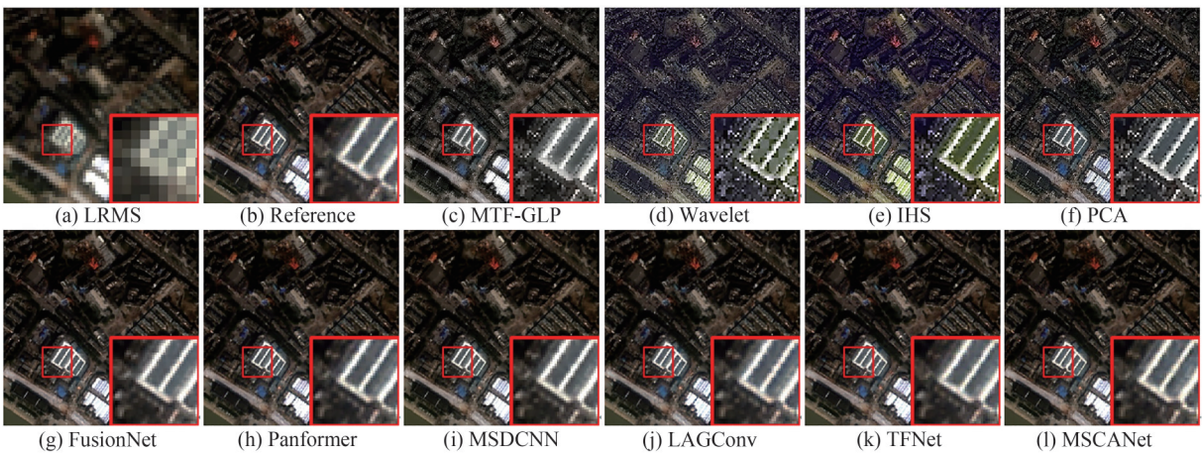

图7　QuickBird仿真数据集融合结果

Fig.7　Fusion result of QuickBird simulation dataset

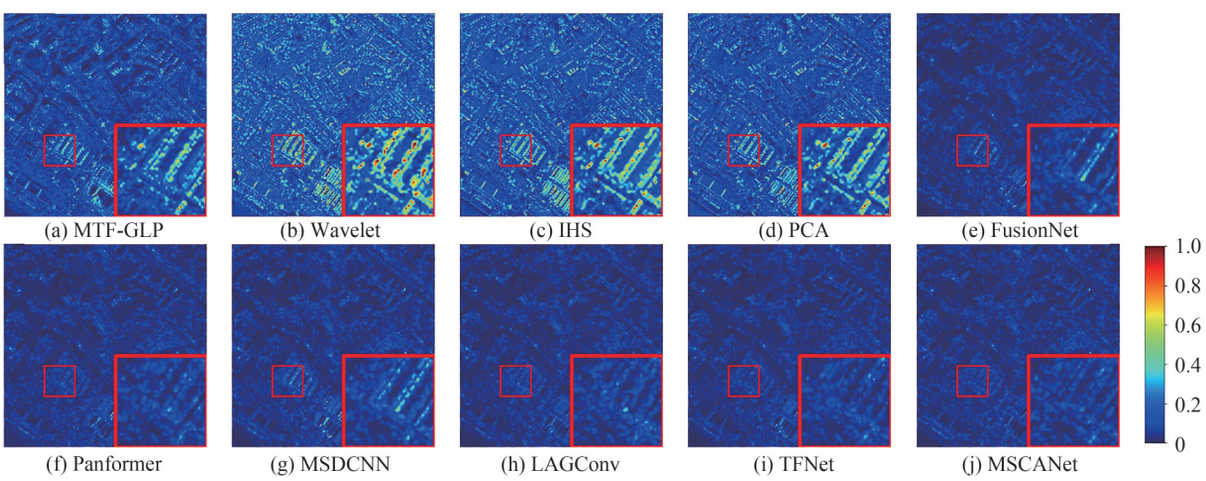

图8　QuickBird仿真数据集的残差图

Fig.8　Residual graph of QuickBird simulation dataset

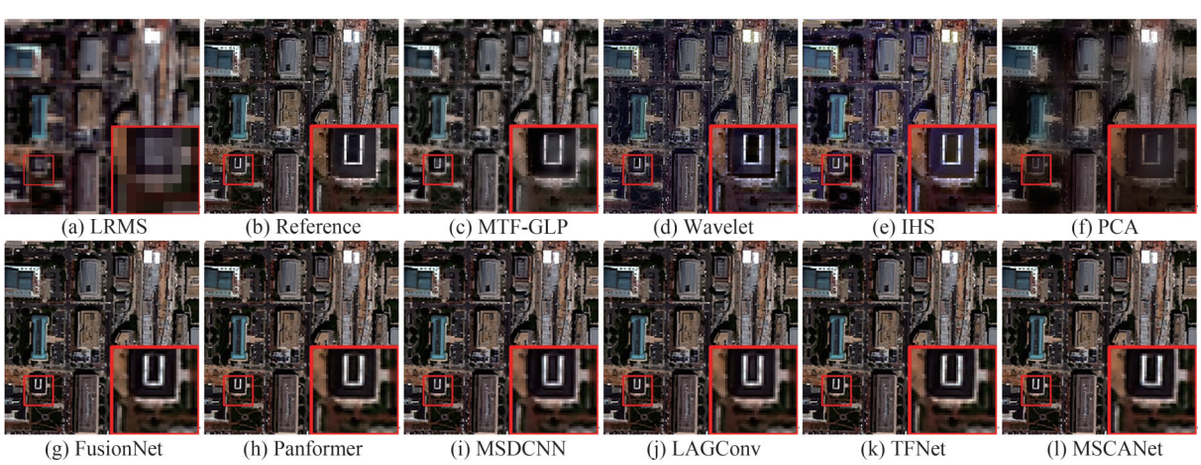

图9　WorldView-2仿真数据集融合结果

Fig.9　Fusion result of WorldView-2 simulation dataset

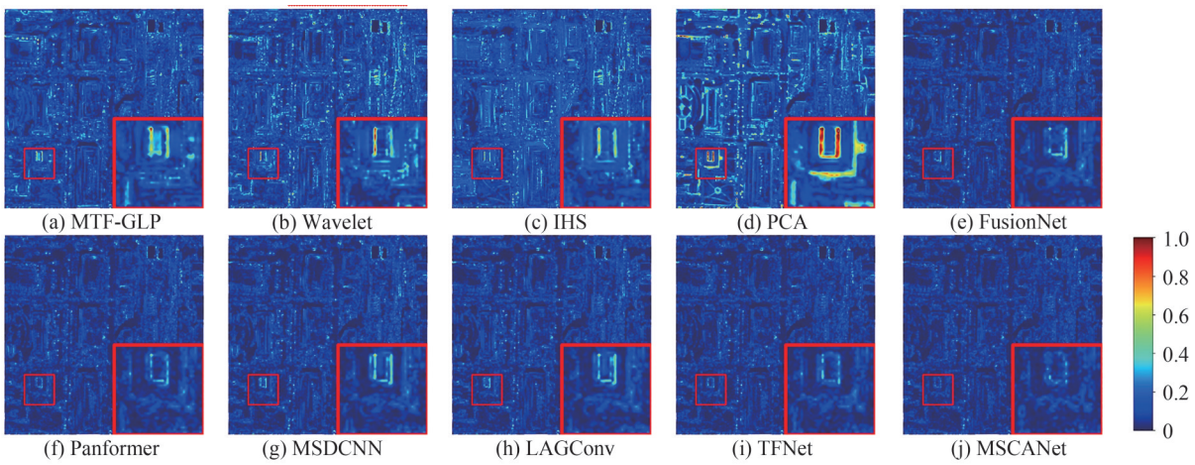

图 10 WorldView-2 仿真数据集的残差图

Fig.10 Residual graph of WorldView-2 simulation dataset

WorldView-2 卫星含有八个波段，所以融合难度要高于四波段图像，在传统方法和深度学习方法的定量评估结果相比其他两种数据集略差，但 MSCANet 依然在各种方法中表现出了最好的效果，相比于排名第二的方法，ERGAS 降低了 3.43%，SAM 降低了 2.68%，PSNR 和 SCC 分别提高了 0.94% 和 0.32%。

3.5 全分辨率下的实验结果分析

在三个数据集分别进行全分辨率实验，将真实的多光谱和全色图像融合，使用主观视觉图像效果以及无参考质量指标 QNR、QNR 的光谱细节损失分量 D_λ、QNR 的空间细节损失分量 D_S 三个指标来对比各个方法的融合效果。图 11 给出各种方法在 WV4 数据集的真实图像融合效果，MTF-GLP 和 Wavelet 方法的边缘模糊严重，IHS 方法的图像整体偏红，PCA 方法生成的图像比较暗，而且屋顶处的红色物体几乎难以辨认，存在一定的光谱扭曲。FusionNet、MSDCNN 和 LAGConv 在建筑物的交界处出现了细节丢失的现象，Panformer、TFNet 则在左下角的泳池处出现了一定程度的色彩失真，而 MSCANet 的主观视觉效果最好。

表 3 展示了各种方法的客观评价指标结果。MSCANet 在所有数据集的 QNR 数值指标均优于其他方法，综合视觉效果和数值指标结果分析，本文方法在全分辨率下的实验结果最佳。

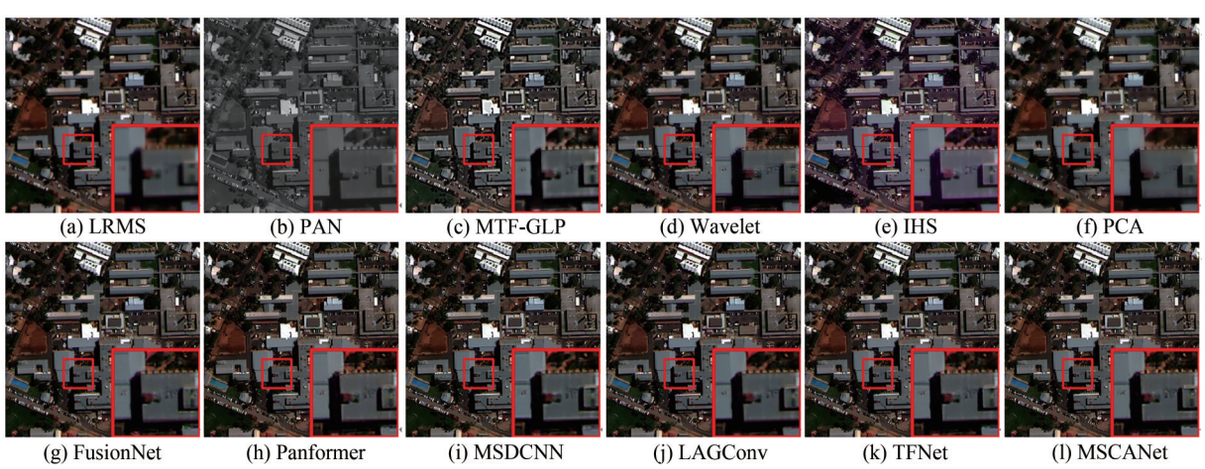

图 11 WorldView-4 真实数据集融合结果

Fig.11 Fusion result of WorldView-4 real dataset

表3　真实数据的客观评价指标

Table 3　Objective evaluation index of real dataset

Method	WV4			QB			WV2		
	D_λ ↓	D_S ↓	QNR ↑	D_λ ↓	D_S ↓	QNR ↑	D_λ ↓	D_S ↓	QNR ↑
MTF-GLP	0.065 5	0.050 9	0.887 1	0.095 7	0.150 9	0.768 9	0.094 2	0.065 3	0.847 0
Wavelet	0.014 1	0.039 8	0.946 7	0.133 5	0.151 4	0.738 2	0.046 9	0.073 1	0.883 6
PCA	0.034 8	0.064 7	0.902 8	0.016 4	0.083 9	0.901 1	0.069 5	0.056 8	0.877 6
IHS	**0.013 3**	0.067 0	0.920 6	0.018 0	0.091 9	0.891 8	0.025 9	0.047 2	0.928 2
MSDCNN	0.024 0	0.016 4	0.960 0	0.013 2	0.034 0	0.953 3	0.018 0	0.046 2	0.936 6
FusionNet	0.027 8	0.026 4	0.946 5	0.014 1	**0.029 1**	0.957 3	0.017 2	0.031 6	0.951 8
Panformer	0.040 0	0.018 8	0.942 0	0.015 1	0.037 3	0.948 2	0.020 3	0.031 4	0.948 9
LAGConv	0.030 6	0.018 9	0.951 1	0.014 9	0.054 1	0.931 8	0.017 7	0.029 5	0.953 4
TFNet	0.019 9	0.026 1	0.954 5	0.015 4	0.041 6	0.943 6	**0.014 9**	0.048 7	0.937 2
MSCANet	0.018 1	**0.008 8**	**0.973 2**	**0.011 0**	0.031 1	**0.958 2**	0.015 5	**0.023 4**	**0.961 5**

3.6　消融实验结果分析

本文融合方法使用的三个策略分别是：1）使用细节注入模型注入高频细节；2）使用多尺度窗口自注意力MSCA模块作为融合网络的主干结构；3）在MAE损失函数的基础上加入光谱损失和空间损失，在WorldView-4数据集上针对每个策略分别进行了消融实验并验证其有效性。

3.6.1　注入模型的有效性分析

本文方法基于注入模型（Injection model），融合网络直接预测多光谱图像缺失的细节，再将网络输出的细节与原始多光谱图像相加，如式（2），而另一种常见的多光谱-全色融合模型是非注入模型（Non-injection model），网络直接预测目标高分辨率多光谱图像，其计算表达式为

$$M_{\mathrm{HS}}^{k}=F\left(M_{\mathrm{up}},P;\theta\right)\qquad(k=1,\cdots,c)\tag{10}$$

分别给出这两种模型在WorldView-4数据集上的融合结果，如表4，可以发现注入模型在各个客观评价指标上的结果均优于非注入模型，验证了本文采取的注入模型策略的有效性。

表4　注入模型在WV4数据集的消融实验结果

Table 4　Ablation result of injection model in WV4 dataset

Model	ERGAS↓	SAM↓	PSNR ↑	SCC ↑
Non-injection model	2.412	2.926	31.944	0.980
Injection model	**2.268**	**2.816**	**32.488**	**0.983**

3.6.2　多尺度窗口注意力单元MSCA的有效性分析

本文提出了多尺度窗口注意力单元MSCA，该单元将原始的Swin transformer中第二个STB的MLP结构替换成了多尺度卷积结构，并在多尺度卷积单元之后加入通道注意力（Channel Attention，CA）结构。为了验证MSCA的有效性，将三种结构进行对比，分别是原始的STB结构、MSCA消去注意力机制的结构以及本文提出的MSCA，分别对应图12中的（a）、（b）、（c）。

实验结果如表5，将MLP模块替换成多尺度单元之后，各项客观评价指标结果都比原始的STB结构有明显提升，证明了对遥感图像不同尺度的信息进行融合的策略相比于原始STB的MLP全连接结构更加有效。在多尺度单元后加入通道注意力机制，融合结果进一步提升，其中ERGAS和SAM分别减小了4.59%和3.43%，证明了在融合模块中加入通道注意力权重对空间和光谱信息的保留都是有益的。另外，将Swin transformer模块改进之后，模型的总参数量由2.20×10^{6}减少为1.99×10^{6}，证明改进后的模型不仅融合效果更好，运算效率也有所提升。

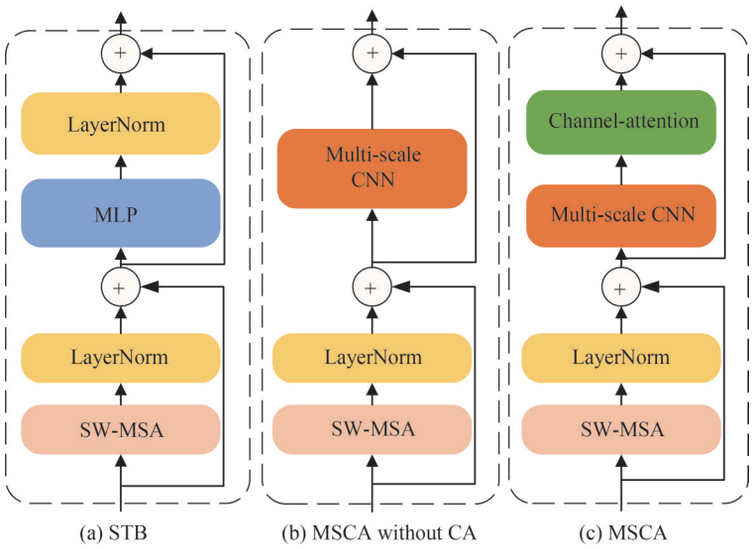

图 12 三种不同的窗口注意力单元结构

Fig.12 Three different window attention unit structures

表 5 MSCA 模块在 WV4 数据集的消融实验结果

Table 5 Ablation result of MSCA in WV4 dataset

Structure	MLP	Multi-scale CNN	Channel-attention	ERGAS↓	SAM↓	PSNR↑	SCC↑
Fig.12(a)	√			2.767	3.156	30.763	0.974
Fig.12(b)		√		2.377	2.916	32.100	0.981
Fig.12(c)		√	√	**2.268**	**2.816**	**32.488**	**0.983**

3.6.3 损失函数的有效性分析

本文提出的损失函数在 MAE 损失的基础上增加了光谱损失(Spectral loss)和空间结构损失(Spatial loss),为了验证新加入的损失函数的有效性,设计了四种不同的损失函数组合,分别是 MAE、MAE+Spectral loss、MAE+Spatial loss 和 MAE+Spectral loss+Spatial loss。在 WorldView-4 数据集进行消融实验,实验结果如表 6。

表 6 损失函数在 WV4 数据集的消融实验结果

Table 6 Ablation result of loss function in WV4 dataset

MAE	Spectral loss	Spatial loss	ERGAS↓	SAM↓	PSNR↑	SCC↑
√			2.362	2.873	32.109	0.981
√	√		2.343	2.845	32.211	0.981
√		√	2.329	2.900	32.261	0.982
√	√	√	**2.268**	**2.816**	**32.488**	**0.983**

分析结果可知,相比单独的 MAE 损失,单独加入光谱损失可以提升其 SAM 结果,即减少融合图像的光谱扭曲程度,单独加入空间结构损失主要提升了 ERGAS 结果,增加了融合图像与参考图像的空间相似程度。同时加入光谱损失和空间损失之后各项评估指标比单独加入其中一项的结果更好,有利于空间信息和光谱信息的同时保持,证明了本文提出的组合损失函数的有效性。

3.7 网络性能分析

所有方法的训练和测试实验在 NVIDIA GeForce RTX 3090 Ti 显卡下实现,表 7 给出了各个网络训练所需的参数量和三个数据集平均每张图片所需的测试时间。其中传统方法无需训练,所以只给出平均测试时间,MTF-GLP 在各个方法中耗时最长,运行效率最低。

在深度学习方法中,自注意力运算相比卷积神经网络的方法时间开销更大,所以 MSCANet 和 Panformer 方法的测试时间略高于其他四种基于 CNN 的深度学习方法,但是本文方法的测试时间短于

表7 所有方法的平均测试时间和参数量

Table 7 Average test time and number of parameters for all methods

Method	Runtime/s	Parameters
MTF-GLP	0.919	—
Wavelet	0.095	—
PCA	0.122	—
IHS	0.105	—
MSDCNN	0.046	0.19×10^6
FusionNet	0.053	0.15×10^6
Panformer	0.197	1.85×10^6
LAGConv	0.079	0.05×10^6
TFNet	0.125	2.36×10^6
MSCANet	0.147	1.99×10^6

Panformer。MSCANet的模型参数量高于MSDCNN、FusionNet和LAGConv三种轻量级模型,但比TFNet的训练参数更少,虽然MSCANet网络运算复杂度适中,但是考虑到融合效果相比其他方法具有更好的光谱保真和空间结构相似性,本文方法依然更具优势。

4 结论

为了更有效地提升卫星捕获到的多光谱图像的空间分辨率,提出了一种基于改进Swin transformer的融合网络MSCANet。模型使用双流分支提取多光谱图像和全色图像的特征,降采样之后的特征图像在通道维级联并送入融合网络。为了提高在各种复杂地面场景中特征提取的鲁棒性,在融合部分集成了一个多尺度窗口注意力单元MSCA,该单元是对Swin transformer的改进,将第二个STB中的MLP模块替换成了多尺度卷积和通道注意力机制。最后,将高层特征与低层特征进行跳跃连接,采用注入模型恢复出高分辨率的多光谱图像。为了实现空间结构信息保真和光谱保真,用MAE损失、光谱损失和空间结构损失的组合损失函数优化模型。分别在三种商用卫星的仿真数据和真实数据集进行对比试验,MSCANet相比其他方法的视觉表现效果和客观评估结果都显著提升。针对本文提出的三个融合策略进行了消融实验,实验结果表明,本文采用的注入模型、MSCA模块的搭建以及损失函数的组合对于光谱保真和空间分辨率提升均是有效的。在未来的工作中,可以将本文提出的MSCANet迁移到多光谱图像和高光谱图像的融合、可见光图像红外图像融合等类似的任务中,提高本文模型的泛化性。

参考文献

[1] DADRASS J F, SAMADZADEGAN F, MEHRAVAR S, et al. A review of image fusion techniques for pan-sharpening of high-resolution satellite imagery[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2021, 171: 101-117.

[2] WANG Z, MA Y, ZHANG Y. Review of pixel-level remote sensing image fusion based on deep learning[J]. Information Fusion, 2023, 90: 36-58.

[3] VIVONE G, GARZELLI A. A benchmarking protocol for pansharpening: dataset, preprocessing, and quality assessment[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 17.

[4] VIVONE G, ALPARONE L, CHANUSSOT J, et al. A Critical comparison among pansharpening algorithms[J]. IEEE Transactions on Geoscience and Remote Sensing, 2015, 53(5): 2565-2586.

[5] AIAZZI B, ALPARONE L, BARONTI S, et al. MTF-tailored multiscale fusion of high-resolution MS and pan imagery[J]. Photogrammetric Engineering & Remote Sensing, 2006, 72(5): 591-596.

[6] SIRGUEY P, MATHIEU R, ARNAUD Y, et al. Improving MODIS spatial resolution for snow mapping using wavelet fusion and ARSIS Concept[J]. IEEE Geoscience and Remote Sensing Letters, 2008, 5(1): 78-82.

[7] SHAH V P, YOUNAN N H, KING R L. An efficient pan-sharpening method via a combined adaptive PCA approach and contourlets[J]. IEEE Transactions on Geoscience and Remote Sensing, 2008, 46(5): 1323-1335.

[8] CHOI M. A new intensity-hue-saturation fusion approach to image fusion with a tradeoff parameter[J]. IEEE Transactions on Geoscience and Remote Sensing, 2006, 44(6): 1672-1682.

[9] BALLESTER C, CASELLES V, IGUAL L, et al. A variational model for P+XS image fusion[J]. International Journal of Computer Vision, 2006, 69(1): 43-58.

[10] MASI G, COZZOLINO D, VERDOLIVA L, et al. Pansharpening by Convolutional Neural Networks[J]. Remote Sensing, 2016, 8(7): 594.
[11] YANG J, FU X, HU Y, et al. PanNet: a deep network architecture for pan-sharpening[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 10012-10022.
[12] YUAN Q, WEI Y, MENG X, et al. A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2018, 11(3): 978-989.
[13] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[J]. arXiv Preprint arXiv: 2010.11929, 2020.
[14] 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, 2021: 10012-10022.
[15] ZHOU H, LIU Q, WANG Y. PanFormer: a transformer based model for pan-sharpening[C]. IEEE International Conference on Multimedia and Expo (ICME), 2022: 1-6
[16] FAN Wensheng, LIU Fan, LI Ming. Remote sensing image fusion based on double-branch u-shaped transformer [J]. Acta Photonica Sinica, 2023, 52(4): 0428002.
范文盛, 刘帆, 李明. 基于双分支U形Transformer的遥感图像融合[J]. 光子学报, 2023, 52(4): 0428002.
[17] LI Ming, LIU Fan, LI Jingzhi. Remote sensing image fusion with detail injection combining convolutional attention module and convolutional autoencoder [J]. Acta Photonica Sinica, 2022, 51(6): 0610005.
李明, 刘帆, 李婧芝. 结合卷积注意模块与卷积自编码器的细节注入遥感图像融合[J]. 光子学报, 2022, 51(6): 0610005.
[18] LI B, CUI W, WANG W, et al. Two-stream convolution augmented transformer for human activity recognition[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(1): 286-293.
[19] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[J]. Medical Image Computing and Computer-Assisted Intervention, 2015: 234-241.
[20] MENG X, XIONG Y, SHAO F, et al. A large-scale benchmark data set for evaluating pansharpening performance: overview and implementation[J]. IEEE Geoscience and Remote Sensing Magazine, 2021, 9(1): 18-52.
[21] XIANG Z, XIAO L, YANG J, et al. Detail-injection-model-inspired deep fusion network for pansharpening[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-15.
[22] IN Z R, ZHANG T J, JIANG T X, et al. LAGConv: local-context adaptive convolution kernels with global harmonic bias for pansharpening[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(1): 1113-1121.
[23] LIU X, LIU Q, WANG Y. Remote sensing image fusion based on two-stream fusion network[J]. Information Fusion, 2020, 55: 1-15.

Remote Sensing Image Fusion Method Based on Improved Swin Transformer

LI Zitong, ZHAO Jiankang, XU Jingran, LONG Haihui, LIU Chuanqi
(*School of Electronic Information and Electrical Engineering*, *School of Perceptual Science and Engineering*, *Shanghai Jiao Tong University*, *Shanghai* 200240, *China*)

Abstract: Remote sensing images are widely used in land monitoring, environmental perception, disaster prediction and urban analysis. Most commercial satellites such as WorldView-4, QuickBird and WorldView-2 are equipped with sensors that can obtain panchromatic images and multispectral images at the same time. Panchromatic images have high spatial resolution but have only one band. The spatial resolution of multispectral images is low due to the bandwidth limitation of the equipment. In order to obtain more accurate details of the measured object, panchromatic image and multispectral image can be fused to generate images with both high spatial resolution and high spectral resolution. Fusion methods of multispectral and panchromatic images can be divided into four categories: multi-resolution analysis method, component substitution method, variational optimization method and deep learning method. Compared with traditional methods, deep learning has stronger feature extraction ability, so it is widely used. Currently, transformer structure is introduced into advanced remote sensing image fusion method. Aiming at the problem that existing methods based on transformer fail to fully integrate multi-scale features

of remote sensing images, this paper proposes a multispectral-panchromatic fusion network MSCANet, based on improved Swin transformer. The model extracts features of multispectral images and panchromatic images respectively by using two-flow branches. The downsampled feature images are cascaded and fed into the fusion network. In order to improve the robustness of feature extraction in various complex ground scenes, a Multiscale Swin-transformer with Channel Attention (MSCA) unit is integrated in the fusion part. The unit replaces the MLP part of Swin transformer into a cascade module of multi-scale convolution and channel attention, which can better fuse the feature information of ground objects of different sizes in remote sensing images and use the long-range dependence between regions. The fusion network focus on predicting the high-frequency details lost in multispectral images. Then high frequency details are added to the original image to restore a high resolution multispectral image. Simulation experiment and real experiment of three commercial satellites are conducted. In the experiment of simulation data, the fusion results were evaluated by calculating the difference between the reference image and the simulation dataset. Compared with other methods, MSCANet has the best performance in visual performance and quantitative metrics. Compared with the method with the second performance, the ERGAS index of MSCANet in the three datasets decreased by 11.99%, 0.4% and 3.43%, respectively. In the experiment of three real datasets, combining visual effect and quantitative metrics analysis, the result of MSCANet is the best. Ablation experiments were conducted for the three fusion strategies proposed in this paper. The experimental result shows that the injected model used in this paper outperforms the non-injected model. It also proves that the replacement of MLP module in MSCA module and the addition of attention mechanism are conducive to the improvement of fusion performance. Also, the addition of spectral loss and spatial structure loss on the basis of MAE loss is effective for the improvement of spectral fidelity and spatial resolution. In conclusion, the effectiveness of the proposed method was verified by comparison and ablation experiments. In future work, MSCANet is expected to be migrated to the fusion of multispectral image and hyperspectral image, visible image and infrared image, and other similar tasks to improve the generalization of the model proposed in this paper.

Key words: Remote sensing; Image fusion; Multispectral image; Panchromatic image; Swin transformer
OCIS Codes: 100.2000; 350.2660; 100.3010; 110.4234

Foundation item: National Natural Science Foundation of China (No. 62171283), Shanghai Commercial Aircraft Systems Engineering Joint Research Foundation (No. CASEF-2022-MQ01)