doi:10. 12068/j. issn. 1005－3026. 2024. 08. 014

多尺度特征融合的Transformer遥感影像超分辨率重建

王 植, 王 坤, 王梦晴
(东北大学 资源与土木工程学院, 辽宁 沈阳 110819)

摘 要: 针对现有遥感影像超分辨率重建算法,在处理复杂场景时,存在无法充分提取和利用特征,且计算复杂度高的问题,提出一种多尺度特征融合的Transformer遥感影像超分辨率重建网络模型. 该模型引入了多尺度残差Swin Transformer模块,在充分提取特征的同时,减少用于提取浅层特征的模块冗余;建立了一个特征细化融合模块,可以充分提取图像特征来提高网络性能. 基于UC Merced Land Use公开数据集进行实验,结果表明:提出的模型所需参数数量仅为目前主流超分辨率重建方法EDSR模型的61. 6%,重建结果在不同尺度下的峰值信噪比和结构相似度相对EDSR分别平均提高了0. 82 dB和0. 024. 通过对比分析,证明本文提出的模型在提高图像质量的同时,有效地减少了网络参数冗余,可明显提高重建图像质量,满足高分辨率遥感影像处理需要.

关 键 词: 遥感影像;超分辨率重建;Transformer;特征提取;特征细化融合
中图分类号: P 237 **文献标志码:** A **文章编号:** 1005-3026(2024)08-1178-07

Super-resolution Reconstruction of Remote Sensing Image Based on Transformer of Multi-scale Feature Fusion

WANG Zhi, *WANG Kun*, *WANG Meng-qing*
(School of Resources & Civil Engineering, Northeastern University, Shenyang 110819, China. Corresponding author: WANG Zhi, E-mail: wangzhi@ mail.neu.edu.cn)

Abstract: To address the limitation of the existing super-resolution reconstruction of remote sensing image algorithms in fully extracting and utilizing features and coping with high computational complexity in complex scenes, a Transformer network model for super-resolution reconstruction of remote sensing image based on multi-scale feature fusion was proposed. The multi-scale residual Swin Transformer module was introduced to fully extract features and reduce the module redundancy used for flat feature extraction. A feature fusion refinement module was established that can fully extract image features to improve network performance. Based on the public UC Merced Land Use dataset, the experimental results show that the number of parameters required by the proposed model is only 61. 6% of the parameters compared with the current mainstream super-resolution reconstruction method EDSR model. The peak signal-to-noise ratio and structural similarity of the reconstruction results at different scales are increased by 0. 82 dB and 0. 024 on average compared with the EDSR model. Through comparative analysis, it is proved that the model proposed can effectively reduce the redundancy of network parameters while improving the quality of the image. It can significantly improve the quality of the reconstructed image to meet the requirements of high-resolution remote sensing image processing.

Key words: remote sensing image; super-resolution reconstruction; Transformer; feature extraction; feature refinement fusion(FRF)

近年来,随着遥感技术的发展,高分辨率遥感影像已广泛应用于国民经济各领域,例如,城

收稿日期: 2023-04-11
基金项目: 中央高校基本科研业务费专项资金资助项目(N170113027).
作者简介: 王 植(1979-),男,辽宁沈阳人,东北大学副教授.

市规划、土地利用、农业、林业和环境保护等方面.目前提高遥感影像的分辨率的方法主要从改善硬件和改进算法两方面入手.通过提升传感器性能来获取高分辨率遥感影像具有成本高、开发周期长等问题,这在一定程度上限制了遥感技术的应用[1].因此,研发实现遥感影像超分辨率(super-resolution,SR)重建的算法具有重要的应用价值.

目前图像超分辨率重建方法主要分为基于插值的方法、基于重建的方法和基于深度学习的方法[2].基于插值的方法主要应用在图像放大、缩小和增强等过程中,通过使用插值函数估计待插入像素点的取值从而生成更加平滑和连续的图像,提高图像的分辨率,该方法具有较高的重建效率,但是在像素突变区域的处理效果差,导致重建结果不理想[3].基于重建的方法是通过利用低分辨率(low-resolution,LR)影像和先验知识建立优化求解模型来取得更好的重建结果[4],然而对于一些场景复杂的影像,该方法仍然存在重建后边缘模糊、纹理不清晰等问题.因此,近年来更多的研究关注基于深度学习的方法.Dong等[5]首次提出超分辨率卷积神经网络(super-resolution of convolutional neural networks,SRCNN),但其训练时间长且不适用于高倍率的超分辨率重建.为了解决这些问题,Shi等[6]设计了高效的亚像素卷积网络(efficient sub-pixel convolutional network,ESPCN),但是随着网络层数的加深会引起梯度消失和梯度爆炸的问题.此后Kim等[7]将残差引入到影像的超分辨率重建中,极大地消除了由于深度过大引起的梯度消失和梯度爆炸问题.Lim等[8]提出了增强型深度超分辨率(enhanced deep super-resolution,EDSR)网络,即通过去除多余的残差块来进一步提高模型性能.Zhang等[9]提出了残差密集网络(residual dense networks,RDN),该网络中设计的残差密集块通过密集连接的卷积层来充分提取图像的局部特征.上述方法均采用堆叠多个卷积层提取特征进行非线性映射,以达到提高影像分辨率的目的.然而,它们都难以充分挖掘全局信息和特征之间的关系,从而导致影像细节丢失、边缘模糊等问题.Vaswani等[10]设计了更为简单的Transformer网络架构,该算法仅基于注意力机制,放弃了循环和卷积,此后在计算机视觉方面开始逐步得到广泛的应用.Liu等[11]使用了Swin Transformer算法通过移位窗口方案将自注意力计算限制在非重叠的局部窗口中,捕捉全局信息和特征之间的关系更为准确,因此被成功应用到图像分类、目标检测、语义分割等视觉任务中.

虽然以上方法在图像超分辨率重建方面已经取得了很好的效果,但是对于需要更多高频细节和纹理信息的复杂遥感影像,这些算法忽略了前一层提取的特征对SR重构性能的影响,造成了特征利用率不足、冗余特征信息等问题.

针对上述算法存在的问题,本文提出了一种多尺度特征融合的Transformer遥感影像超分辨率重建方法,该方法在传统卷积神经网络模型的基础上引入了Transformer的全局自注意力机制和特征融合策略,可以更好地捕捉全局信息和局部特征之间的关系.经过实验证明,本文提出的模型在处理遥感影像超分辨率重建方面效果更佳.

1　研究方法

1.1　网络结构

本文提出的网络模型主要包括特征提取模块和特征重建模块,网络总体框架如图1所示.将输入的低分辨率影像进过一个3×3的卷积层用于提取输入影像的低级特征,并将提取的特征直接传递给重建模块,以保留低频信息;深层特征提取模块由两个区域级非局部(region-level non-local,RL-NL)模块和多个带有残差的Swin Transformer模块(residual Swin Transformer block,RSTB)组成.其中,RL-NL模块用于捕获远距离上下文信息,它能够在图像的特定区域内捕捉到全局的关系,采用多个RSTB进行深层特征的提取.此外,为了加速本文模型的收敛,将第一个RL-NL模块的输出依次传播到RSTB.最后,利用特征细化融合(feature-refinement fusion,FRF)模块将提取的所有特征进行拼接,并使用1×1卷积降维,最后与原始特征一起输入重建模块进行重建从而提高重构性.特征重建模块由2个3×3卷积层和1个亚像素卷积层组成,主要用于最终的遥感影像超分辨率重建.

1.2　Swin Transformer残差模块

遥感影像与自然图像不同,遥感影像尺度跨度大、纹理细节复杂,传统的卷积神经网络用相同的卷积核在学习特征的过程中忽略了局部特征和全局特征之间的关系,且无法区分遥感影像在不同通道中的关键信息,使重建结果在边缘区

域和纹理细节方面表现不佳.因此可以引入Transformer,它可以被视为空间变化卷积的一个具体实例,同时也包含了注意力机制,使网络能够在学习局部特征和全局特征之间关系的同时也能够自适应地关注图像中的高频信息,从而更好地利用特征来提高网络的重构性能.其次,RSTB由多个带有跳跃连接的Swin Transformer层(Swin Transformer layers,STL)组成.RSTB模块结构如图2所示,这种结构允许聚合不同级别的特征,从而提高特征的利用率.

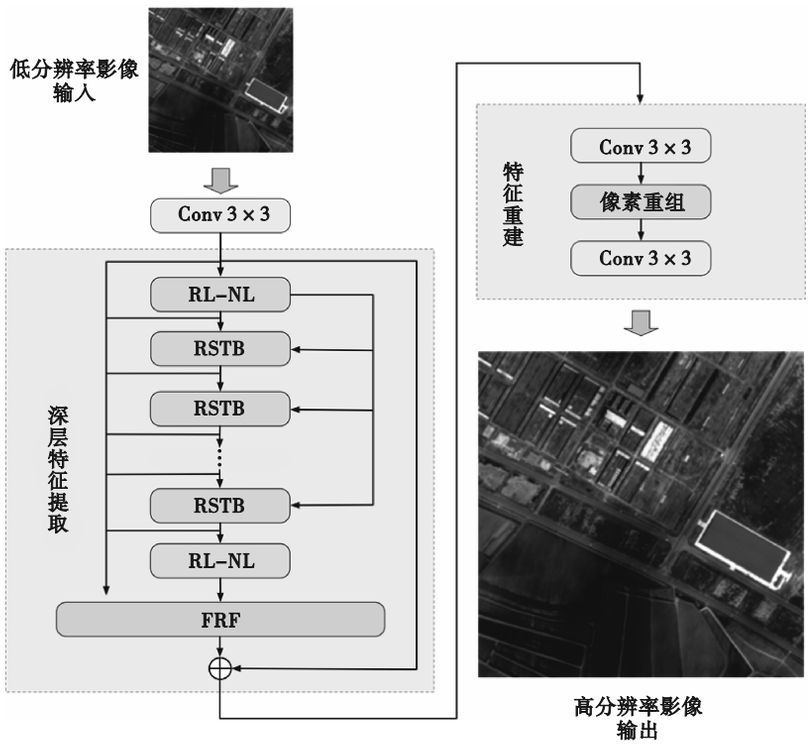


图1 本文算法网络总体框架

Fig. 1 Overall framework of the proposed algorithm network

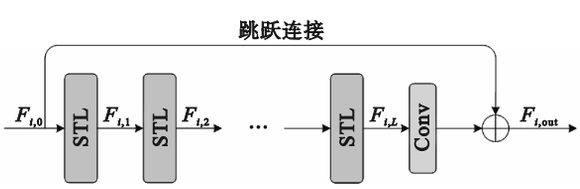


图2 RSTB结构示意图

Fig. 2 Schematic diagram of the RSTB structure

在RSTB模块中,给定第i个RSTB的输入特征$F_{i,0}$通过L个STL模块提取中间特征$F_{i,1}$,$F_{i,2},\cdots,F_{i,L}$,特征提取可用式(1)表示:

$$F_{i,j}=H_{\text{STL}}\left(F_{i,j-1}\right),j=1,2,\cdots,L. \tag{1}$$

式中:$F_{i,j}$为第i个RSTB中的第j个STL模块的输出;$F_{i,j-1}$为第i个RSTB中的第$(j-1)$个STL模块的输出.RSTB的输出可用式(2)表示:

$$F_{i,\text{out}}=H_{\text{Conv}}\left(F_{i,L}\right)+F_{i,0}. \tag{2}$$

式中:$F_{i,\text{out}}$为第i个RSTB的输出;$F_{i,L}$为第i个RSTB中最后一个STL的输出;$F_{i,0}$为第i个RSTB的输入;H_{Conv}为卷积操作.

如图3所示,每个STL首先将上一级的特征输入通过LayerNorm(LN)层进行归一化操作,然后将结果在多头自注意力(multi-head self-attention,MSA)中进行拼接,在MSA机制中,每个头计算一个注意力矩阵,表示输入序列中不同位置的权重.通过连接多个头的注意力矩阵,综合各头的信息以更好捕捉输入序列特征.最终,拼接后的特征矩阵作为下一层网络输入,进行后续特征处理.接下来,使用一个带有两个全连接层和GELU非线性激活函数的多层感知机(multi-layer perceptron,MLP)来进一步对特征进行变换.这样设计的目的是在具有大尺度跨度的遥感影像数据中,由于图像中存在长距离像素之间的关联,传统的局部感知方法可能无法有效捕捉这些长距离关系.因此,通过引入全局感知机制和多通道信息交互,将空间位置的特征进行全局感知,可以更好地捕捉图像中不同位置之间的长距离关系,从而更好地处理遥感影像尺度跨度大的问题.

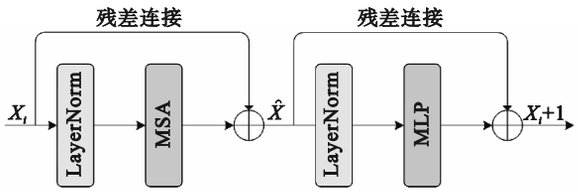


图3　STL结构示意图

Fig. 3　Schematic diagram of the STL structure

在每个MSA模块和每个MLP之前应用LN层，并且在每个模块之后使用残差连接，整个计算过程可表示为

$$\hat{X}_i = D_{\mathrm{MSA}}\left(L_{\mathrm{LayerNorm}}\left(X_i\right)\right) + X_i\ , \tag{3}$$

$$X_{i+1} = P_{\mathrm{MLP}}\left(L_{\mathrm{LayerNorm}}\left(\hat{X}_i\right)\right) + \hat{X}_i\ . \tag{4}$$

式中：D_{MSA}表示多头注意力计算；P_{MLP}表示多层感知机；$L_{\mathrm{LayerNorm}}$表示归一化操作；X_i表示STL模块输入特征；$\hat{X}_i$表示MSA模块输出特征；X_{i+1}表示MLP模块输出特征.

注意力矩阵是通过局部窗口中的自注意机制计算得出的，在式(5)中，通过将$\boldsymbol{Q}$和$\boldsymbol{K}^{\mathrm{T}}$进行点积操作得到相似度得分，然后对得分进行归一化，得到每个位置的注意力权重.最后，将每个位置的注意力权重与对应位置的值向量$\boldsymbol{V}$相乘并求和，得到对输入序列的加权和.

$$G_{\mathrm{attention}}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right) = T_{\mathrm{softmax}}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}} + B\right)\boldsymbol{V}. \tag{5}$$

式中：$G_{\mathrm{attention}}$表示注意力计算；T_{softmax}为softmax函数；$\boldsymbol{Q}$表示查询向量；$\boldsymbol{K}$表示键向量；$\boldsymbol{V}$表示值向量；B表示可学习的相对位置偏差；d表示键向量的维度.

1.3　特征细化融合模块

特征细化融合(FRF)模块是一种在超分辨率重建任务中广泛应用的模块，通过利用上一级模块的输出特征来优化当前特征，以提高重建性能.该模块将模型多级提取的特征作为输入，利用卷积和池化操作来降低特征的维数，并将降维后的特征进行拼接操作得到细化后的特征.最后，将细化后的特征作为重建模块的输入，在重建模块中生成重建图像.具体过程如图4所示.

假设第$(x+1)$个模块的输出为O_{x+1}，则融合过程可以表示为

$$\dot{O}_x = f\left(O_x + O_{x+1}\right). \tag{6}$$

式中：$f(\cdot)$表示卷积运算；O_x表示第x个模块输出；O_{x+1}为第$x+1$个模块输出；$\dot{O}_x$为融合后特征输出.

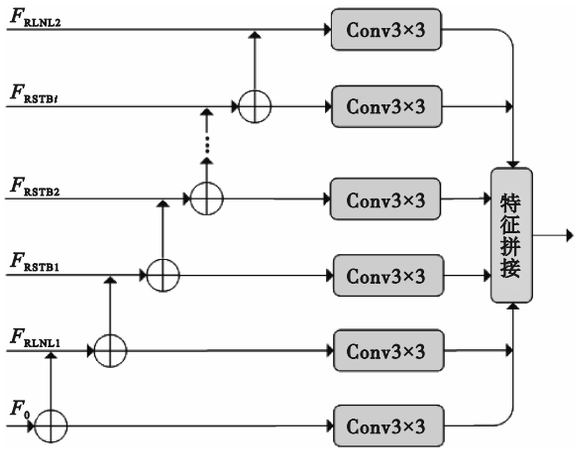


图4　FRF 结构示意图

Fig. 4　Schematic diagram of the FRF structure

2　实验与分析

2.1　数据集和模型参数设置

本文使用了UC Merced Land Use公开数据集[12]，其中包含了21类不同的土地使用类型，每个类别有100张尺寸为256像素×256像素的遥感图像.实验随机选择800张图像，其中80张用于测试集，720张用于训练集.每张遥感图像被切割为64像素×64像素的块，并将其进行模糊和降采样处理，作为输入低分辨率图像.

网络结构中，RSTB模块数量、STL的数量和窗口移动大小，分别设置为4，6和8.在模型中，除了用于融合特征的1×1卷积核外，其余卷积核大小均为3×3.训练过程中，采用ADAM优化器[13]更新权重参数，设置指数衰减率为$\beta_1 = 0.9$，$\beta_2 = 0.99$，网络训练的初始学习率设为0.000 1，每迭代1 000次后学习率减半，并采用$L1$损失函数作为ADAM的优化目标.

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left\| X_{\mathrm{HR}}^{i} - D\left(X_{\mathrm{LR}}^{i}\right)\right\|_1. \tag{7}$$

式中：$L(\theta)$为平均绝对误差；N为训练对的数量；X_{HR}^{i}为输入的高分辨率影像；$D\left(X_{\mathrm{LR}}^{i}\right)$为重建低分辨率影像.

2.2　实验环境与评价指标

本文进行网络训练的硬件配置采用Windows 11系统，CPU为Intel i7−12700 H，配备32 GB的NVIDIA RTX 3060显卡进行开发，实验主要利用的开发编程语言为Python，使用的深度学习框架是PyTorch.

本文评价指标采用峰值信噪比(peak signal-to-noise ratio, PSNR)和结构相似度

(structural similarity index measure, SSIM). 其中PSNR计算方法是通过计算两个图像之间的均方误差(mean square error, MSE)得到的. MSE越小,PSNR值越高,表示重建图像的质量越好.

SSIM是一种考虑了图像亮度、对比度和结构等方面特征的基于人眼视觉特性的图像质量评价指标,它能够更好地反映人眼对图像质量的主观感受,因此通常被用于评估图像的清晰度、细节等方面. SSIM的取值范围为[-1, 1],在实际应用中,SSIM的取值通常在0到1之间,当SSIM接近1时,表明超分辨率重建后的图像与真实图像在结构、纹理和细节等方面高度相似,因此可作为超分辨率重建算法的评价指标.

2.3 实验对比分析

2.3.1 RSTB模块数量对模型性能的影响

在UC Merced Land Use数据集上,采用不同数量的RSTB模块进行超分辨率重建实验,以验证本文提出的网络模型中RSTB模块数量对模型性能的影响.实验结果见表1.

从表1的实验结果可以看出,当增加RSTB模块的数量时,PSNR和SSIM也随之提高,当模块数量为4之后开始趋于稳定,当模块数量增加到6时,PSNR仅仅提高了0.02 dB,为了权衡模型性能和模型大小,因此选取RSTB的数量为4.

表1 RSTB模块数量对模型性能的影响

Table 1 Effect of the number of RSTB modules on model performance

RSTB模块数量/个	PSNR/dB	SSIM
2	34.78	0.951 2
4	34.89	0.953 8
6	34.91	0.952 1
8	34.87	0.952 9

2.3.2 FRF模块对模型性能的影响

为了进一步验证所提模型中使用FRF模块的有效性,实验比较了有无FRF模块在UC Merced Land Use数据集上的训练结果.图5分别展示了在2倍和4倍放大因子下,迭代100次带有和不带有FRF模块的PSNR曲线.实验结果表明了FRF的引入使得重建效果更好.

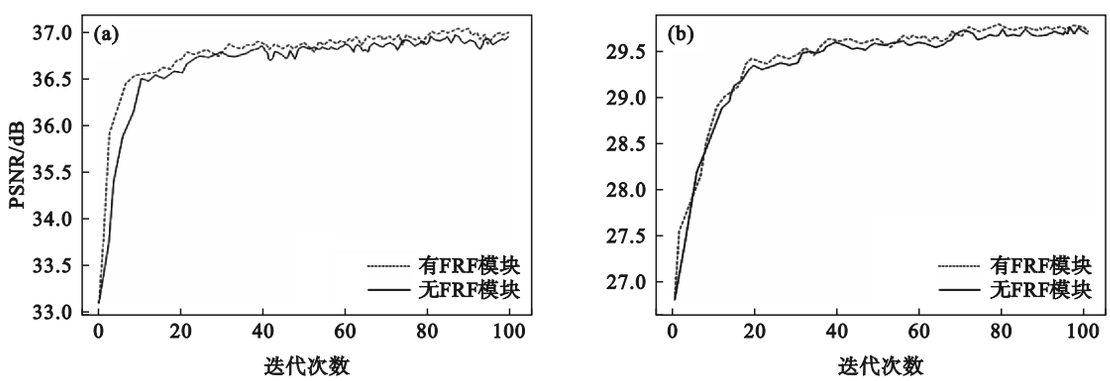


图5 迭代100次时有无FRF模块对模型性能的影响

Fig. 5 Influence of FRF module with or without FRF module on model performance with 100 iterations

(a)—2倍放大因子;(b)—4倍放大因子.

2.3.3 模型尺寸分析

为了验证本文所提出的模型性能,在UC Merced Land Use数据集上分别对SRCNN,RDN和EDSR标准模型进行训练,对不同模型的参数量与本文模型对比分析,得出不同模型的参数数量和PSNR的关系,实验结果如表2所示,从表2可以看出,与超分辨率重建中性能较好的EDSR模型相比,本文模型所需参数数量为2.68×10^7,仅为EDSR所需数量的61.6%,并且重建结果的PSNR比使用EDSR提高了0.82 dB.实验表明,本文提出的模型可以在不大幅度增加参数数量的情况下实现更好的重建结果.

表2 不同模型参数量的对比

Table 2 Comparison of the number of parameters of different models

模型名称	参数数量/个	PSNR/dB
SRCNN	1.56×10^7	25.82
RDN	4.12×10^7	26.95
EDSR	4.35×10^7	27.47
本文模型	2.68×10^7	28.29

2.3.4 与其他模型SR结果对比

为了评估本文所提出的模型在遥感影像超分辨率重建方面的效果,设计了在相同的实验

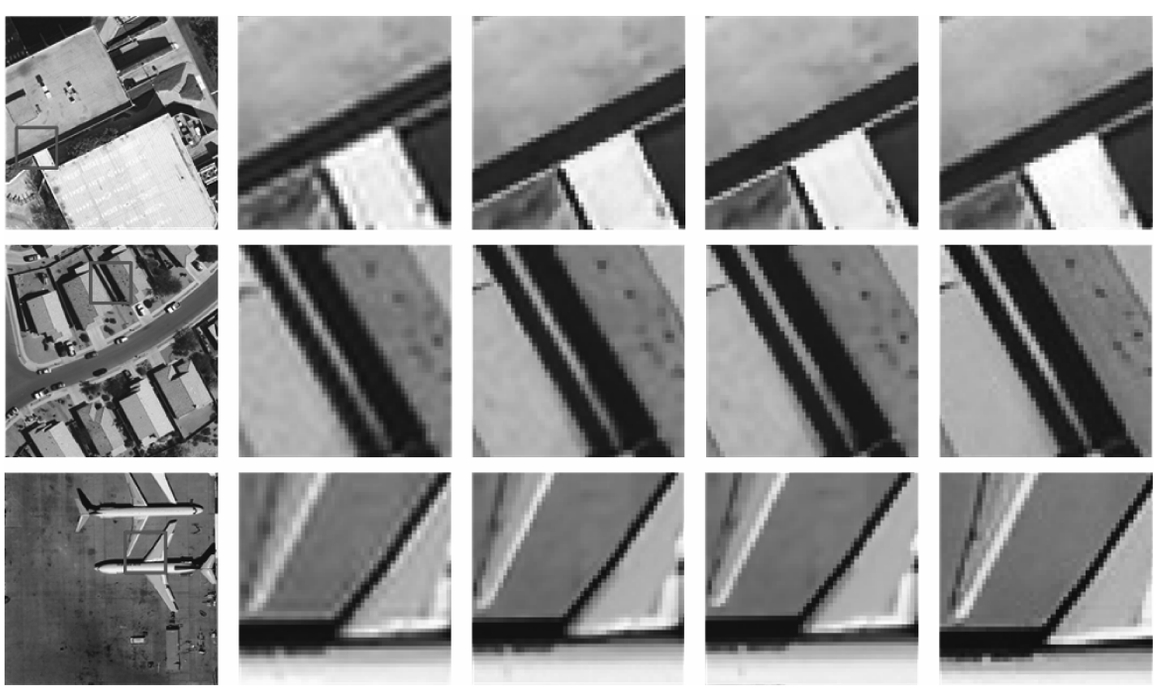
条件下的不同算法在不同测试集上的对比实验,将本文模型与SRCNN,RDN和EDSR模型进行了定量分析比较,并总结了不同尺度下的不同影像类型的重建结果.实验主要选择遥感影像中的建筑物、密集住宅区和机场这3种不同影像类型进行测试,影像的定量评估结果见表3.从表3可以看出,在不同的尺度上,本文提出的模型在测试集中的PSNR和SSIM指标均优于其他模型.

表3　不同模型在不同测试集上的PSNR和SSIM的对比结果

Table 3　Comparison results of PSNR and SSIM of different models on different test sets

影像类型	放大因子倍数	SRCNN		RDN		EDSR		本文模型	
		PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM
建筑物	2	28.99	0.910 5	31.68	0.943 7	32.45	0.947 9	33.76	0.949 9
	3	25.52	0.815 3	27.42	0.870 2	27.74	0.879 2	29.36	0.910 2
	4	23.68	0.726 5	24.90	0.785 2	25.31	0.802 5	26.47	0.870 1
密集住宅区	2	30.67	0.900 7	32.82	0.931 4	33.93	0.941 6	34.52	0.942 5
	3	27.52	0.814 2	29.26	0.864 8	29.92	0.878 9	30.31	0.892 8
	4	25.66	0.732 6	26.73	0.783 7	27.31	0.806 3	28.26	0.832 6
机场	2	32.61	0.920 4	35.11	0.949 4	36.19	0.958 0	36.41	0.969 1
	3	29.89	0.858 3	31.66	0.895 3	32.23	0.904 2	33.15	0.924 7
	4	28.12	0.802 8	29.23	0.835 9	29.78	0.849 1	30.14	0.894 6

为了更直观地观察遥感影像超分辨率重建的效果,实验分别在2倍和4倍放大因子下,对测试集遥感影像中的建筑物、密集住宅区、机场这3种影像类型进行可视化重建,并与其他算法进行对比,各算法超分辨率重建后的影像可视化结果如图6所示.从放大后的影像可以清晰地看出,本文模型能够重建出更为精细的边缘纹理细节.与其他模型相比较,本文模型超分辨率重建后的遥感影像在纹理细节和整体效果上都更接近真实的遥感影像.



原始影像　SRCNN　RDN　EDSR　本文模型

(a)

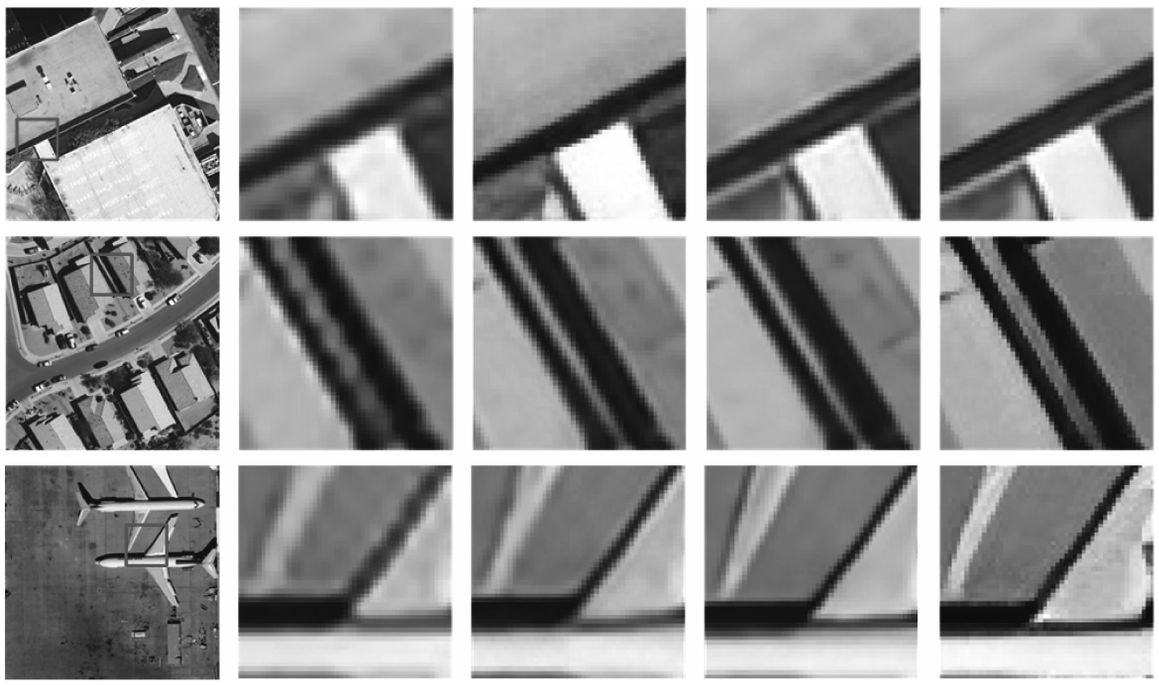


图6 不同模型在不同测试集上的可视化结果对比

Fig. 6 Comparison of visualization results of different models on different test sets

(a)—2倍放大因子；(b)—4倍放大因子.

3 结 语

本文提出了一种多尺度特征融合的Transformer遥感影像超分辨率重建方法.该方法在传统CNN模型的基础上引入了多尺度残差Swin Transformer模块,使得模型充分并有效地利用特征,同时加入了特征细化融合模块,有效地融合了多个尺度的特征信息.通过与目前主流的超分辨率重建网络在3种不同尺度上的实验结果表明,本文提出的模型在有效地减少网络参数冗余的基础上,不牺牲模型性能的同时,能够生成更加清晰的纹理细节,有效地解决了遥感影像重建过程中地物轮廓不清晰和产生伪影等问题,在定量评价和视觉效果方面均优于现有主流超分辨率重建算法,能够明显提高图像质量,满足高分辨率影像处理需要.

参考文献:

[1] Huang B, He B, Wu L N, et al. Deep residual dual-attention network for super-resolution reconstruction of remote sensing images[J]. *Remote Sensing*, 2021, 13(14): 2784-2802.

[2] 王植,李安翼,方锦雄.基于密集卷积神经网络的遥感影像超分辨率重建[J].测绘与空间地理信息,2020,43(8):4-8.(Wang Zhi, Li An-yi, Fang Jin-xiong. Super resolution reconstruction of remote sensing images based on dense convolution neural network [J]. *Geomatics & Spatial Information Technology*, 2020, 43(8): 4-8.)

[3] Chang K, Ding P L K, Li B X. Single image super-resolution using collaborative representation and non-local self-similarity[J]. *Signal Processing*, 2018, 149: 49-61.

[4] Ma Y C, Lyu P Y, Liu H, et al. Remote sensing image super-resolution based on dense channel attention network [J]. *Remote Sensing*, 2021, 13(15): 2966-2986.

[5] Dong C, Loy C C, He K M, et al. Learning a deep convolutional network for image super-resolution [C]// European Conference on Computer Vision (ECCV). Cham: Springer, 2014: 184-199.

[6] Shi W Z, Caballero J, Huszár F, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016: 1874-1883.

[7] Kim J, Lee J K, Lee K M. Accurate image super-resolution using very deep convolutional networks [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016: 1646-1654.

[8] Lim B, Son S, Kim H, et al. Enhanced deep residual networks for single image super-resolution [C]//IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Honolulu, 2017: 1132-1140.

[9] Zhang Y L, Tian Y P, Kong Y, et al. Residual dense network for image super-resolution [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 2472-2481.

[10] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [EB/OL]. (2017-06-12) [2023-04-11]. http://arxiv.org/abs/1706.03762.

[11] Liu Z, Lin Y T, Cao Y, et al. Swin Transformer: hierarchical vision transformer using shifted windows [C]//IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, 2021: 9992-10002.

[12] Avcibas I, Sankur B, Sayood K. Statistical evaluation of image quality measures [J]. *Journal of Electronic Imaging*, 2002, 11(2): 206-223.

[13] Singarimbun R N, Nababan E B, Sitompul O S. Adaptive moment estimation to minimize square error in backpropagation algorithm [C]//International Conference of Computer Science and Information Technology (ICoSNIKOM). Medan: IEEE, 2019: 1-7.