激光与光电子学进展 先进成像

结合 Transformer 与多尺度残差机制的高光谱遥感分类

陈禹汗，王波*，严清赟，黄冰洁，贾桐，薛彬
南京信息工程大学遥感与测绘工程学院，江苏 南京 210044

摘要 卷积神经网络(CNNs)在高光谱图像分类中已经取得了令人瞩目的成果。但由于卷积运算的局限性，CNNs并不能很好地进行上下文信息交互。为了解决远距离捕获高光谱序列关系的问题，本文将Transformer用于高光谱分类。提出了一种基于Swin Transformer的多尺度混合光谱注意力模型(SMSaNet)。在提出的SMSaNet中使用多尺度光谱增强残差融合模块和光谱注意力模块对光谱特征进行建模，使用改进的Swin Transformer模块来提取空间特征，最后使用全连接层实现对高光谱图像的分类。在两个公开数据集Indian Pines和University of Pavia上将SMSaNet与其他5种分类方法进行对比实验，结果表明SMSaNet获得了最优的分类效果，总体分类精度分别达到了99.51%和99.56%。

关键词 图像处理；高光谱遥感；残差网络；注意力机制；Transformer；感受野
中图分类号 P237 **文献标志码** A **DOI**：10.3788/LOP220921

Hyperspectral Remote-Sensing Classification Combining Transformer and Multiscale Residual Mechanisms

Chen Yuhan, Wang Bo*, Yan Qingyun, Huang Bingjie, Jia Tong, Xue Bin
School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, *Jiangsu, China*

Abstract Convolutional neural networks (CNNs) have achieved impressive results in hyperspectral image classification. However, because of the limitations of convolution operations, CNNs cannot satisfactorily perform contextual information interaction. In this study, we use the Transformer for hyperspectral classification to address the problem of capturing hyperspectral sequence relationships at extended distances. We propose a multiscale mixed spectral attention model based on Swin Transformer (SMSaNet). The spectral features are modeled using the multiscale spectral enhancement residual fusion module and the spectral attention module in SMSaNet. The spatial features are then extracted using the improved Swin Transformer module, and hyperspectral image classification is realized using a fully connected layer. SMSaNet is compared with five other classification models on two public datasets, that is, the Indian Pines and University of Pavia. The results show that SMSaNet achieves the best classification effect compared to the other models. The overall classification accuracies reach 99.51% and 99.56%, respectively.

Key words image processing; hyperspectral remote sensing; resnet; attention mechanism; Transformer; receptive field

1 引 言

近年来，由于地球观测与导航项目的大力发展，高光谱传感器的运用受到了广泛的关注，运用星载或机载传感器能够捕获大量的高光谱图像(HSI)。这些图像具有丰富的光谱与空间信息，在资源管理、沿海湿地测绘和监测[1]、土地覆被变化监测[2]、环境污染监测[3]和农作物精细分类等多个领域得到了应用。

高光谱遥感影像被广泛运用于地物精细分类领域，在早期的高光谱分类研究中，运用了大量的传统机器学习方法，Moughal等[4]将支持向量机成功运用于高光谱图像的分类并同最大似然法和光谱角映射器进行了比较。Petropoulos等[5]将支持向量机和人工神经网络运用于高光谱图像的分类，发现使用低空间分辨率

收稿日期：2022-03-08；**修回日期**：2022-04-20；**录用日期**：2022-06-13；**网络首发日期**：2022-06-23
基金项目：南京信息工程大学大学生创新创业训练计划项目(XJDC202110300481)
通信作者：*wangbo@nuist.edu.cn

高光谱图像时可能出现混合问题，这会显著降低分类器的精度。虽然上述传统方法能够在一定程度上取得较好的性能，但这些方法都是基于人工特征构造的浅层分类器，且大多只关注于光谱信息，难以对复杂地物进行精准分类。况且这些方法要想取得较好性能需要依赖于人工进行分类特征的选择，特征的选择对研究人员来说是异常繁琐的。

近年来，随着深度学习的迅速发展，使得研究人员可不再手动进行分类特征选择。典型的深度学习方法有神经网络[6]、深度信念网络[7]、卷积神经网络(CNNs)和图卷积神经网络[8]等。在这些模型中，CNNs在高光谱的特征提取和分类中得到了广泛的应用，如冯凡等[9]将多特征融合和混合卷积网络运用于高光谱分类，解决了三维卷积网络在训练样本较少时对高光谱图像的分类精度不理想的问题。

虽然CNNs的性能比传统机器学习方法提升了不少，但其仍旧存在着源自于卷积运算本身的局部性，图像和卷积核之间的交互是相互独立的，在进行局部处理的原则下，卷积对于长距离依赖的建模是无效的。基于此，有研究者提出了注意力机制[10]，并且在计算机视觉领域取得了重大突破。这是因为CNNs与注意力机制的结合能聚焦于给定信息的局部位置，将其分配相应的权值，强调特征图中关键性征，削弱相对无用的特征，因而能够提高模型的性能。如王欣等[11]提出一种基于三维卷积神经网络(3D-CNN)并结合双分支双注意力机制的快速密集连接网络，提高了高光谱分类的精度并减少训练时间。

注意力机制的使用不只在视觉领域，在自然语言处理(NLP)领域，基于Transformer[12]的方法在各种任务中已经达到了先进水平。Transformer最早被应用于机器翻译任务，但在最近两年被大量应用于视觉领域。Qing等[13]提出了一个名为SATNet的端到端Transformer模型，该模型适用于高光谱分类并依赖于自注意机制，结果表明Transformer能够很好地应用于高光谱分类。

Dosovitskiy等[14]证实了在视觉领域中CNNs的使用并不是必要的，将Transformer直接应用在视觉领域可以很好地执行图像分类任务。实验表明，使用迁移学习的Vision Transformer(ViT)与最先进的卷积网络相比能够取得优异的结果，且训练所需的计算资源大量减少，但同基于CNNs的方法相比，ViT需要使用大型数据集进行预训练。

最近，在ViT的基础上已经有了一些优秀的方法，其中最具影响力的是Swin Transformer[15]，一种层级式的ViT，在多个任务上达到了先进水平。Swin Transformer可以被用来作为计算机视觉领域一个通用的骨干网络，它拥有像卷积神经网络一样分层的结构，因为这种多尺度的特征，它可以使用到计算机视觉各个领域。

Swin Transformer在视觉领域的分类效果要显著优于传统的CNNs[15]，但相关的研究并没有将Swin Transformer应用于高光谱图像的分类当中。为填补这一空缺，本文设计了基于Swin Transformer的多尺度混合光谱注意力模型(SMSaNet)，提高了在训练样本有限条件下高光谱图像的分类精度。

本文的主要贡献有：1)对Swin Transformer进行了改进，利用多尺度光谱增强残差融合模块和光谱注意力机制对光谱特征进行建模，利用改进的Swin Transformer特征提取模块对空间特征进行建模，让Swin Transformer能够在高光谱数据集上得以应用。在公开数据集上经过实验验证得出本文方法具有优异的性能和泛化能力；2)进行了消融实验，结果表明，Swin Transformer所提出的移位窗口与窗口自注意力的结合使用，能够显著地提高模型的建模能力；3)本文方法可与主流的CNNs方法相结合，可利用CNNs中的大量先验知识对模型进行优化。

2 基本原理

2.1 整体结构

因为高光谱图像数据中光谱维数是远大于空间邻域的，所以在模型分类过程中会出现Hughes现象[16]。为减少光谱冗余，本文运用主成分分析(PCA)方法对图像进行数据降维，提取具备有效光谱信息特征的波段。在光谱波段进行维度压缩由N_1维压缩至N_2维，得到压缩后的$w\times h\times N_2$数据立方体，其中w和h分别代表图像的宽和高，N_1和N_2代表光谱特征维度。本文选取的N_2值为18。

本文模型输入数据为通过PCA降维的数据集$\hat{I}\in\mathbb{R}^{w\times h\times 18}$，从$\hat{I}$立方体中创建近邻块子集$P\in\mathbb{R}^{Q\times Q\times 18}$，其中$Q$代表近邻块的宽和高。为了能够充分利用地物标签信息，取重叠的近邻块作为本文模型的输入，以其中间像素为该近邻块的类别标签。

传统CNNs模型通过利用不同大小的卷积核由浅到深地提取特征，其理论感受野可以达到全图，但大量的研究[17]证实了CNNs的实际感受野是远小于其理论感受野的，这不利于充分利用上下文信息进行特征提取，而利用Transformer中的自注意机制能够进行远程信息建模。因此本文先使用CNNs对光谱维度进行特征提取，之后使用改进的Swin Transformer特征提取模块，逐渐地扩大感受野，利用注意力机制建立起不同地物的空间相关性。

SMSaNet的网络结构如图1所示，其中共有6个Transfomer模块，2个CNNs模块，包含1个多尺度光谱残差融合模块和1个光谱注意力模块。首先运用PCA法去除高光谱图像光谱维度上的冗余，减少图像处理的复杂度，接着连接一个多尺度光谱增强残差融合模块，使用不同尺寸的卷积核能够充分利用光谱信息。

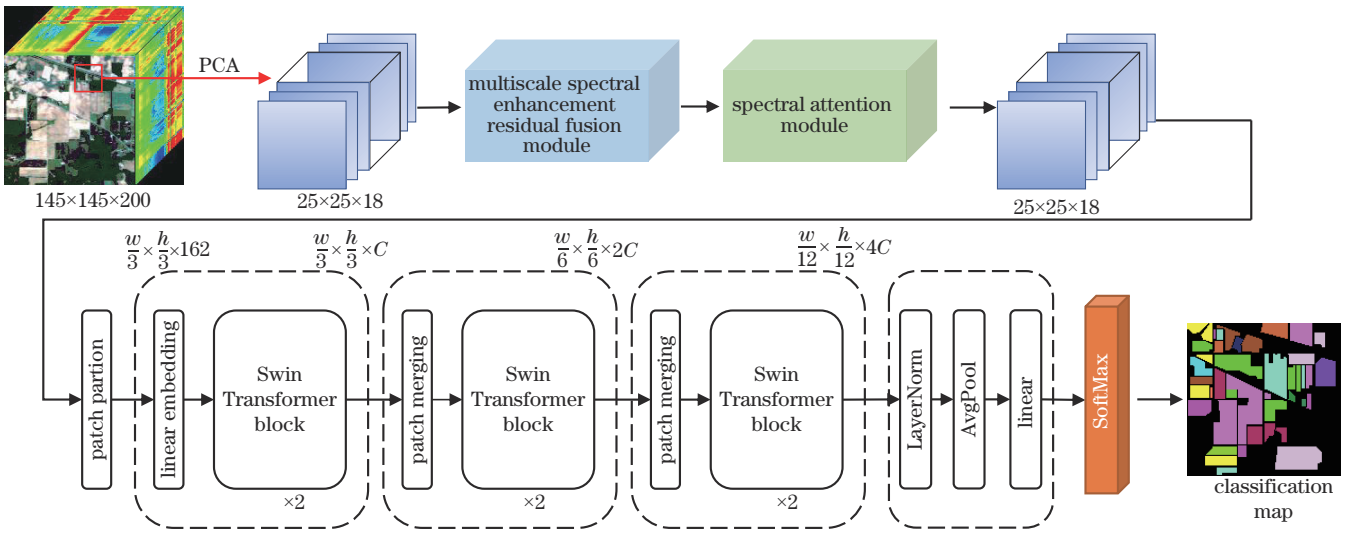

图 1　SMSaNet 网络结构

Fig. 1　Network structure of SMSaNet

接着连接 1 个光谱注意力模块，利用不同光谱通道之间的重要性进行建模。之后连接改进的 Swin Transformer 特征提取模块，对空间信息进行建模。最后连接两层全连接层，加入 dropout 策略[18]防止过拟合。

2.2　多尺度光谱增强残差融合模块

Sun 等[19]的研究表明，多尺度特征对于高光谱地物分类是必要的，本文使用通道分组结构在每个分支上增加光谱特征增强残差模块，对每个分组特征块运用不同大小的卷积核，用以获取不同尺度所表示的输出特征，取得增强的重要通道的特征图。本文设计的多尺度光谱残差融合模块如图 2 所示，为获得更细粒度的多个感受野，因而在光谱维度上进行维度通道拆

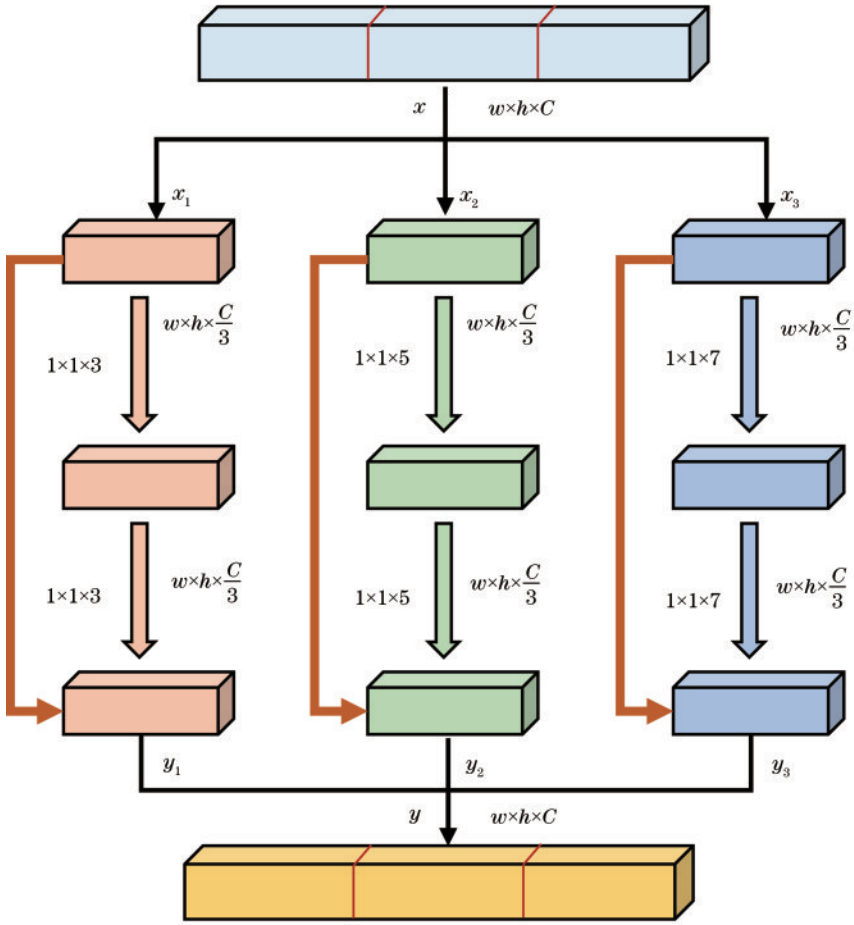

图 2　多尺度光谱增强残差融合模块

Fig. 2　Multiscale spectral enhancement residual fusion module

分，假设输入特征数据 $x\in\mathbb{R}^{w\times h\times C\times b}$，$w\times h$ 表示特征图的空间维度大小；C 表示特征图中光谱维度大小；b 表示特征图中特征维度大小。为了实现分层操作，将光谱通道数划分为 3 个部分，表示为 $x_i\in\mathbb{R}^{w\times h\times \hat{C}\times b}(i=1,2,3)$，其中每个特征图子集 x_i 的空间维度大小相同；$\hat{C}=C/3$。x_1 中光谱特征增强残差模块取 $1\times1\times3$ 的卷积核进行卷积，x_2 中光谱特征增强残差模块取 $1\times1\times5$ 的卷积核进行卷积，x_3 中光谱特征增强残差模块取 $1\times1\times7$ 的卷积核进行卷积。每个部分的输出为

$$y_i=\begin{cases}\mathrm{Conv}_{1\times1\times k}\left(\sigma\left(\mathrm{Conv}_{1\times1\times k}(x_i)\right)\right)\oplus x_i, i=1, k=3\\ \mathrm{Conv}_{1\times1\times k}\left(\sigma\left(\mathrm{Conv}_{1\times1\times k}(x_i)\right)\right)\oplus x_i, i=2, k=5,\\ \mathrm{Conv}_{1\times1\times k}\left(\sigma\left(\mathrm{Conv}_{1\times1\times k}(x_i)\right)\right)\oplus x_i, i=3, k=7\end{cases}\tag{1}$$

式中：$\mathrm{Conv}_{1\times1\times k}(\ )$ 表示具有 $1\times1\times k$ 的卷积核的三维 CNNs 层；σ 表示 Relu 激活函数；i 表示特征图子集的序号。然后对 y_i 进行逐一级联（Concat），得到模块的输出 y。

2.3　光谱注意力模块

为防止上述光谱维度上的通道分组操作使得通道间的相关性丢失，因而引入光谱注意力机制，光谱注意力模块如图 3 所示。将多尺度光谱增强残差融合模块的输出 y 作为光谱注意力模块的输入，$y\in\mathbb{R}^{w\times h\times C}$，其中 $w\times h$ 代表空间维度大小；C 代表特征图光谱特征维度大小。运用平均池化（AvgPool）和最大池化（MaxPool）来压缩空间维度信息，经过权重共享的多层感知器（MLP）模块第一层分别得到 $1\times1\times\frac{C}{r}$ 的特征图，其中 r 表示压缩率，MLP 模块第二层得到 $1\times1\times C$ 的特征图。将二者相加后运用 Sigmiod 激活函数进行激活得到特征矩阵 $\boldsymbol{M}_C\in\mathbb{R}^{1\times1\times C}$，最后与 y 相乘，得到光谱维加权的特征图 $\hat{y}\in\mathbb{R}^{w\times h\times C}$。特征图计算式为

$$\hat{y}=\boldsymbol{M}_C\otimes y,\tag{2}$$

式中：$\hat{y}\in\mathbb{R}^{w\times h\times C}$ 表示光谱维度加权的特征图；$\boldsymbol{M}_C\in\mathbb{R}^{1\times1\times C}$ 表示特征矩阵；$\boldsymbol{y}$ 为多尺度光谱增强残差融合模块的输出。

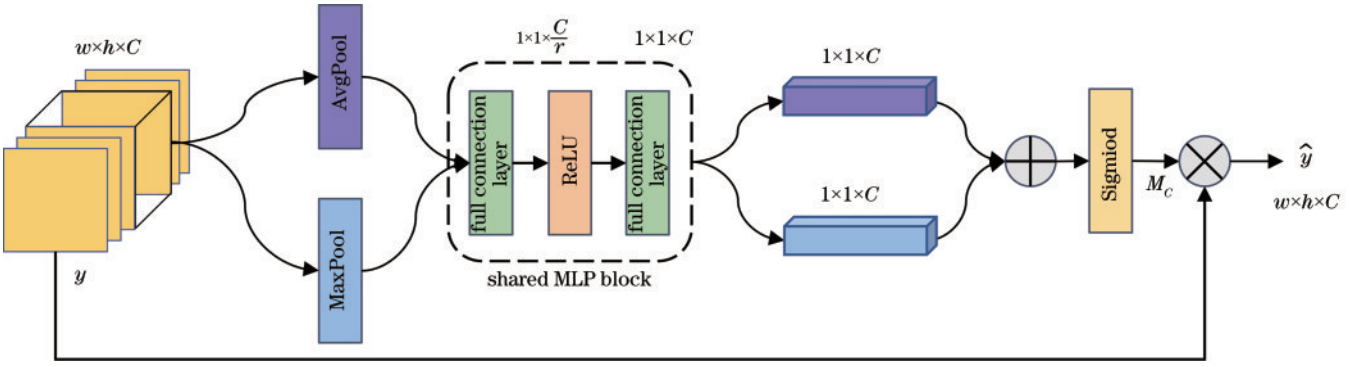

图 3　光谱注意力模块

Fig. 3　Spectral attention module

2.4　Swin Transformer 特征提取模块

本文设计的改进的 Swin Transformer 特征提取模块与层级式的 CNNs 非常相似，模块由 3 个 stage 叠加而成。对于输入特征，应用 patch partition 方法将特征图的最小单位由像素转换为 patches，将输入特征图 $\hat{y}\in\mathbb{R}^{w\times h\times C}$ 划分为不同的 patches 集合。本文设置 C 的值为 96，每个 patch 的尺寸大小为 3×3；通过 patch partition 每个 patch 的特征维度为 3×3×18；patches 的总数量为 $\frac{w}{3}\times\frac{h}{3}$。输入的 patches 进入 stage 1 部分，在 stage 1 部分使用 linear embedding 方法将划分后的 patch 特征维度转换为 C，然后经过一组 Swin Transformer block；在 stage 2 部分先使用 patch merging 方法将输入的相邻 patches 进行合并，使得 patches 的总数量为 $\frac{w}{6}\times\frac{h}{6}$，特征维度变为 $2C$，再通过一组 Swin Transformer block；最后进入 stage 3，使用 patch merging 方法将输入的相邻 patches 合并，使得 patches 的总数量为 $\frac{w}{12}\times\frac{h}{12}$，特征维度变为 $4C$，再经过一组 Swin Transformer block。经过上述层级式的叠加，模块在最后可达到全局感受野。

图 4 中，patch partition 使用一个 2D 卷积层，其参数 stride、kernel_size 都设置为 3，输出通道值设置为 3×3×18。对经过 2D 卷积的特征图的最后两个维度进行拼接操作，将通道维度与拼接的新维度进行交换。linear embedding 为一个 1D 卷积层，卷积核大小为 1，输入通道值为 3×3×18，输出通道值为 C。patch merging 能够使不同 patches 合并，使得通道数为原来的 4 倍，再通过线性层进行降维，使得通道数降低为原来的 1/2。

2.4.1　Swin Transformer block

标准的 Tranformer 编码器由多头自注意力（MSA）和 MLP 组成，并且在各个模块之前使用 LayerNorm（LN），在每个块之后运用残差连接。标准

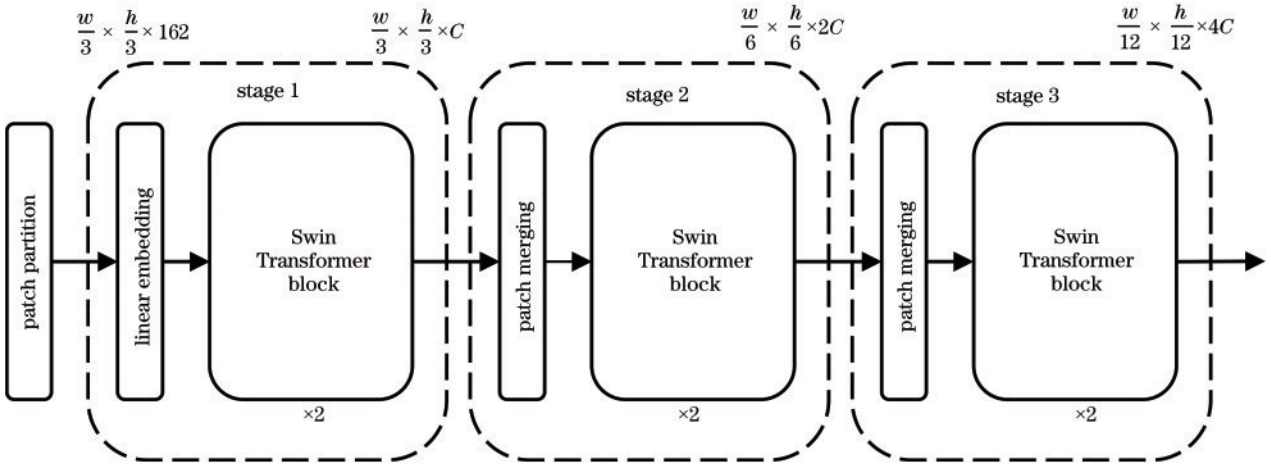

图 4　Swin Transformer特征提取模块

Fig. 4　Swin Transformer feature extraction module

Transformer中的MLP是具有GeLu非线性的两层全连接层。基于Transformer编码器结构构建的Swin Transformer block如图5所示，在Swin Transformer block中MSA块被替换为窗口多头自注意力(W-MSA)和移位窗口多头自注意力(SW-MSA)，Swin Transfomer block以两层block连在一起作为一个基本单元。第一层的Swin Transformer block由MLP和W-MSA组成，之后的Swin Transformer block由MLP和SW-MSA组成。在W-MSA、SW-MSA和MLP之前使用LN，并且在W-MSA、SW-MSA和MLP之后使用残差连接。

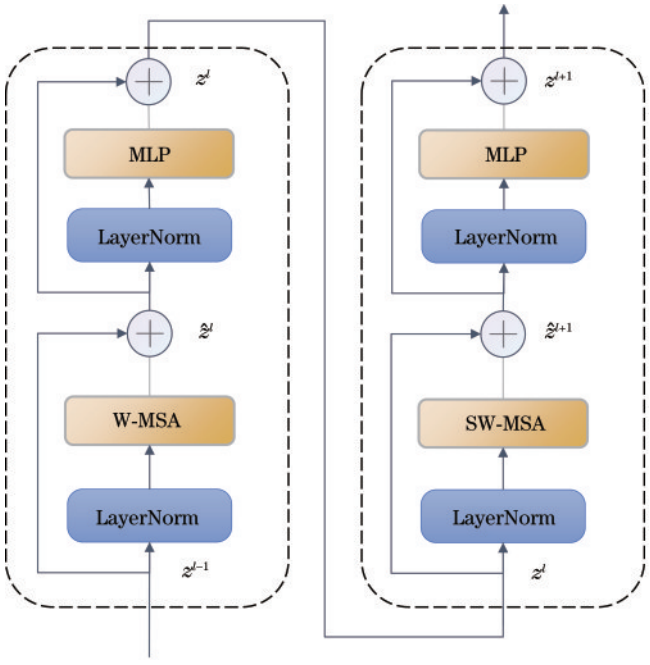

图 5　Swin Transformer块

Fig. 5　Swin Transformer block

对于一组连续的Swin Transformer blocks计算过程如下式所示：

$$\hat{z}^{l}=\text{W-MSA}(\text{LN}(z^{l-1}))+z^{l-1},\tag{3}$$

$$z^{l}=\text{MLP}(\text{LN}(\hat{z}^{l}))+\hat{z}^{l},\tag{4}$$

$$\hat{z}^{l+1}=\text{SW-MSA}(\text{LN}(z^{l}))+z^{l},\tag{5}$$

$$z^{l+1}=\text{MLP}(\text{LN}(\hat{z}^{l+1}))+\hat{z}^{l+1},\tag{6}$$

式中，$\hat{z}^{l}$和z^{l}分别表示Swin Transformer块l中的(S)W-MSA模块和MLP模块的输出特征。

2.4.2　W-MSA和SW-MSA

在计算机视觉领域，MSA的使用取得了巨大的成功，MSA的使用不但可以提高精度，还能够使得损失的landscape变得更加平整，从而提高模型的泛化能力。近期研究表明，MSA和卷积(Convs)通常表现出相反的行为，最后阶段的MSA在预测中起着关键作用[20]。MSA虽然大量应用在计算机视觉领域，但其计算的复杂度较高，应用W-MSA能够降低计算复杂度。标准的MSA计算的具体过程如式(7)～(10)所示。假设存在i个head，patches的数目为$W\times H$。

$$\boldsymbol{Q}_i=\boldsymbol{X}\boldsymbol{W}_i^{Q},\boldsymbol{K}_i=\boldsymbol{X}\boldsymbol{W}_i^{K},\boldsymbol{V}_i=\boldsymbol{X}\boldsymbol{W}_i^{V},\tag{7}$$

$$\text{head}_i=\text{Attention}(\boldsymbol{Q}_i,\boldsymbol{K}_i,\boldsymbol{V}_i),\tag{8}$$

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{SoftMax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V},\tag{9}$$

$$\text{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Concat}(\text{head}_1,\cdots,\text{head}_i)\boldsymbol{W}^{O},\tag{10}$$

式中：$\boldsymbol{X}\in\mathbb{R}^{WH\times C}$表示输入矩阵；$\boldsymbol{Q}_i,\boldsymbol{K}_i,\boldsymbol{V}_i\in\mathbb{R}^{WH\times\frac{C}{i}}$分别表示第$i$个head的query、key、value矩阵；$\boldsymbol{W}_i^{Q},\boldsymbol{W}_i^{K},\boldsymbol{W}_i^{V}\in\mathbb{R}^{C\times\frac{C}{i}}$分别表示第$i$个head的query、key、value的参数矩阵；$\boldsymbol{W}^{O}\in\mathbb{R}^{C\times C}$表示转换矩阵；$d$为query和key的维度；$\text{Concat}(\text{head}_1,\cdots,\text{head}_i)\in\mathbb{R}^{WH\times C}$表示将矩阵进行拼接。

W-MSA是在大小为$M\times M$的窗口中进行计算，总共通过$\left[\frac{W}{M}\times\frac{H}{M}\right]$个窗口进行MSA计算。如图6所示，图中深色区域代表窗口，灰色区域代表patch。W-MSA通过将输入图片划分成不重合的窗口，然后

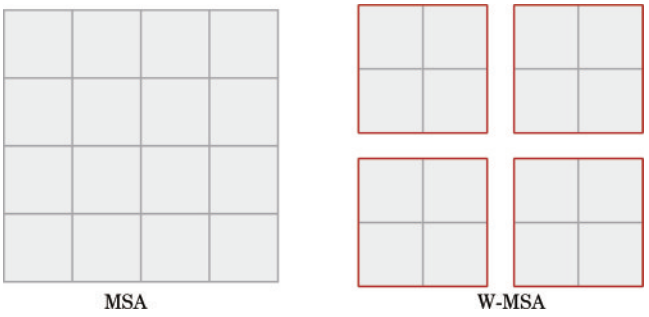

图 6　MSA 和 W-MSA
Fig. 6　MSA and W-MSA

在不同的窗口内进行 MSA 计算。对于一张图片来说含有 $W\times H$ 的 patches，每个窗口包含 $M\times M$ 个 patches，对于每个窗口的自注意力计算如下式所示：

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{SoftMax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}+\boldsymbol{B}\right)\boldsymbol{V},\tag{11}$$

式中：$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\in\mathbb{R}^{M^2\times d}$ 表示 query、key、value 矩阵；M^2 和 d 分别表示窗口中 patches 的数量和 query 与 key 的维数；$\boldsymbol{B}$ 由偏置矩阵 $\hat{\boldsymbol{B}}\in\mathbb{R}^{(2M-1)\times(2M+1)}$ 得到。对于 MSA、W-MSA 每个窗口的计算量以及 W-MSA 的计算量的总体计算量如下式所示：

$$\Omega\,\mathrm{MSA}=4WHC^2+2(WH)^2C,\tag{12}$$

$$\Omega\,\mathrm{Window}=4M^2C^2+2M^4C,\tag{13}$$

$$\Omega\,\mathrm{W\text{-}MSA}=4WHC^2+2M^2WHC,\tag{14}$$

式中：$W\times H$ 为 patches 的总数；W 和 H 分别表示特征图的宽和高；C 代表特征维度；M 代表每个窗口的大小。

基于窗口计算自注意力的方法降低了计算的复杂度，但是窗口与窗口却不能够进行交互，为了在保持非重叠窗口高效计算的同时引入跨窗口连接，在连续 Swin Transformer block 中使用 SW-MSA 来改变 W-MSA 的范围。如图 7 所示，左侧为 layer 1 使用了 W-MSA，右侧为 layer1+1 使用了 SW-MSA。窗口会从特征图的 $\left[\frac{M}{2}\times\frac{M}{2}\right]$ 位置处开始，进行 Window Partition 操作，通过左右图像的对比能够发现窗口发生了明显偏移。

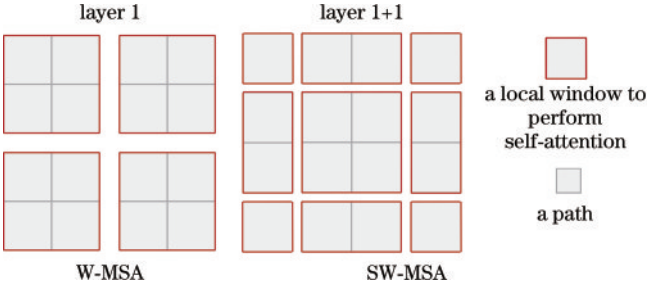

图 7　W-MSA 和 SW-MSA
Fig. 7　W-MSA and SW-MSA

3　分析与讨论

3.1　实验平台参数设置

所有实验均是在配备 Intel(R) Core(TM) i5 10400CPU@2.90 GHz 处理器、Nvidia GeForce RTX 3060 显卡的 Windows10 系统上运行的。为了减小实验误差，模型从训练集中随机抽取有限样本进行训练，epoch 设置为 100，批次大小设置为 128，所有实验结果取 10 次实验的平均值。模型使用通用的 Adam 优化器，设置为默认参数，设置初始学习率大小为 0.001。

在了验证本文方法的有效性，将 SMSaNet(本文方法)与神经网络(Baseline)、1D-CNN[21]、3D-CNN[22]、3D+2D-CNN[23]，以及 Swin Transformer[15](Swin-T)的分类效果进行对比。在 Indian Pines 和 University of Pavia 两个公开数据集上进行实验，数据集参数信息如表 1 和表 2 所示。为了评估不同模型对高光谱图像分类的性能，使用总体精度(OA)、平均精度(AA)和 Kappa 系数(Kappa)作为评价标准。

表 1　India 数据集的地物类别和样本数
Table 1　Figure categories and sample counts of India dataset

Class No.	Land cover/use type	Training	Test
1	Alfalfa	23	23
2	Corn-notill	300	1128
3	Corn-min	300	530
4	Corn	118	119
5	Grass-pasture	241	242
6	Grass-trees	300	430
7	Grass-pasture-moved	14	14
8	Hay-windrowed	239	239
9	Oats	10	10
10	Soybean-notill	300	672
11	Soybean-mintill	300	2155
12	Soybean-clean	296	297
13	Wheat	102	103
14	Woods	300	965
15	Buildings-grass-trees-crives	193	193
16	Stone-steel-towers	46	47
Total		3082	7167

3.2　高光谱数据集

Indian Pines(India)数据集拍摄于美国印第安纳西北部的农场测试地，运用机载传感器 AVIRIS 采集。数据集的波段范围为 400～2500 nm，空间分辨率为 20 m，图像原始尺寸为 145×145，本文使用剔除水吸收和低信噪比波段后为 200 个波段的数据进行分类实验，实验过程中采用的训练样本和测试样本的划分如表 1 所示。对于 India 数据集中样本数少于 600 的类，本文随机选取该类的一半作为训练样本，当样本数大

表 2 PU 数据集的地物类别和样本数

Table 2 Figure categories and sample counts of PU dataset

Class No.	Land cover/use type	Training	Test
1	Asphalt	663	5968
2	Meadows	1864	16785
3	Gravel	209	1890
4	Trees	306	2758
5	Painted metal sheets	134	1211
6	Bare soil	502	4527
7	Bitumen	133	1197
8	Self-blocking bricks	368	3314
9	Shadows	94	853
Total		4273	38503

于 600 时，随机取该类 300 个作为训练样本，剩余样本定义为测试样本[24]。

University of Pavia（PU）数据集拍摄于意大利北部帕维亚大学区域，运用机载传感器 ROSIS 采集，空间分辨率为 1.3 m，图像原始尺寸为 610×340，本文使用剔除噪声影响波段后为 103 个波段的数据进行分类实验。同 India 数据集相比，该数据集样本数量较多，类别较少。实验过程中采用的训练样本和测试样本的划分如表 2 所示。对于 PU 数据集，其样本数量较多，且各类地物较易区分，所以本文随机选取每类的 10% 为训练样本，其余为测试样本。

3.3 实验结果与分析

3.3.1 空间尺寸以及 dropout 率的选择

在实验过程中不同大小的 dropout 率会对实验结果造成影响。神经网络在训练过程中，由于神经元之间的依赖性，容易导致训练数据的过拟合。dropout 对于神经网络单元，按照指定概率将其临时从网络中丢弃，有助于减少网络相互依赖的学习和防止过拟合。dropout 作为超参数，对于不同的网络值是不同的，根据表 3 中的实验结果，本文对 India 数据集选择 dropout 率为 0.3、对 PU 数据集选择 dropout 率为 0.1 来进行后续实验。

表 3 不同 dropout 率在 India 和 PU 数据集上的实验结果

Table 3 Experimental results of different dropout rates on India and PU datasets

Dropout rate	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
OA /%（India）	99.47	95.60	**99.51**	97.41	98.23	99.01	98.44	96.91
OA /%（PU）	**99.56**	99.04	99.31	99.19	99.07	99.02	98.99	98.45

输入图像的尺寸对模型的分类性能至关重要[25]，为了得到模型的最佳输入图像尺寸，采用不同尺寸的输入图像在 India 数据集上对本文方法进行实验，将输入尺寸设置为 9×9～35×35，本文方法的 OA、AA、Kappa 随输入图像尺寸的变化情况如表 4 所示。可发现在输入尺寸取 25×25 时能够取得最好的 OA、AA、Kappa 值，因而本文选择空间尺寸为 25×25 来进行后续实验。

表 4 不同空间尺寸在 India 数据集上的实验结果

Table 4 Experimental results of different spatial sizes on India dataset

Spatial dimension	9×9	11×11	13×13	15×15	17×17	19×19	21×21
OA /%	97.92	98.51	98.93	99.02	99.22	99.21	99.33
AA /%	98.10	98.68	98.95	99.05	99.11	99.16	99.36
Kappa /%	98.21	98.51	98.73	99.17	99.25	99.31	99.34
Spatial dimension	23×23	25×25	27×27	29×29	31×31	33×33	35×35
OA /%	99.47	**99.51**	99.39	99.35	99.31	99.26	99.21
AA /%	99.52	**99.66**	99.45	99.26	99.18	99.15	99.10
Kappa /%	99.42	**99.44**	99.32	99.21	99.05	98.72	98.69

3.3.2 结果与分析

实验结果表明，在训练样本数量有限的情况下，通过对比可以发现使用 3D-CNN 分类策略优于 1D-CNN 分类策略，这是因为三维卷积核更符合高光谱的空谱特性[26]，而使用 3D-CNN 与 2D-CNN 相结合的卷积神经网络结构，在分类过程中充分利用了局部空谱信息，因此分类效果较好，Swin Transformer 比 3D-CNN 的分类策略较好是因为其充分利用了空间特征。但仅使用 Swin Transformer 未考虑光谱波段之间的相关性，本文方法充分利用光谱间的相关性和空间特征使得分类结果在 OA、AA、Kappa 等指标上均取得最好的成绩。

SMSaNet 与 Baseline、1D-CNN、3D-CNN、3D+2D-CNN 以及 Swin-T 这 5 种方法的比较结果如表 5 和表 6 所示。从表 5 和表 6 中可以看出，本文方法的分类精度均优于其他方法。从 India 数据集上看，本文方法比 Baseline、1D-CNN、3D-CNN、3D+2D-CNN、Swin-T 在 OA 上分别提高了 14.3 百分点、5.3 百分点、0.5 百分点、0.2 百分点、0.9 百分点；在 AA 上分别提高了 10.67 百分点、10.53 百分点、1.21 百分点、0.22 百分

表 5 India 数据集分类结果

Table 5 Classification results on India dataset

No.	Baseline	1D-CNN	3D-CNN	3D+2D-CNN	Swin-T	SMSaNet
1	96.77	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
2	77.91	78.15	98.60	98.70	96.44	**98.95**
3	70.16	74.39	99.32	98.97	98.80	**99.48**
4	66.83	68.45	**100.00**	**100.00**	98.22	99.09
5	93.43	91.57	**100.00**	**100.00**	99.41	99.85
6	95.38	95.95	99.41	99.61	99.22	**99.80**
7	95.24	86.36	95.24	90.91	**100.00**	**100.00**
8	99.11	98.82	99.41	**100.00**	**100.00**	99.85
9	68.75	70.59	**100.00**	**100.00**	**100.00**	**100.00**
10	79.55	77.82	99.27	99.27	97.84	**99.85**
11	90.81	88.87	99.53	99.47	**99.76**	99.33
12	74.07	65.96	99.52	98.54	97.42	**99.76**
13	99.28	97.92	**100.00**	**100.00**	**100.00**	**100.00**
14	97.22	97.92	97.77	99.77	99.77	**99.89**
15	76.55	72.99	96.09	98.54	95.07	**98.90**
16	93.65	91.18	98.46	**100.00**	**100.00**	99.22
OA /%	85.21	84.21	99.01	99.31	98.63	**99.51**
AA /%	88.99	89.13	98.45	99.44	98.74	**99.66**
Kappa /%	83.29	82.17	98.87	99.22	98.44	**99.44**

表 6 PU 数据集分类结果

Table 6 Classification results on PU dataset

No.	Baseline	1D-CNN	3D-CNN	3D+2D-CNN	Swin-T	SMSaNet
1	93.00	92.97	98.90	99.17	**99.28**	99.20
2	96.18	96.52	99.77	**99.84**	99.77	99.80
3	77.03	81.45	97.78	98.99	**99.07**	97.44
4	94.63	94.88	**99.70**	98.27	99.44	99.56
5	99.92	**100.00**	**100.00**	99.67	99.92	**100.00**
6	90.91	92.99	99.98	**100.00**	99.96	**100.00**
7	82.69	86.91	99.09	99.50	99.50	**99.58**
8	82.69	82.73	98.05	**99.60**	97.30	99.08
9	**99.65**	98.95	96.51	98.91	97.41	98.81
OA /%	92.66	93.40	99.32	99.54	99.38	**99.56**
AA /%	90.15	90.81	98.39	98.95	98.72	**99.18**
Kappa /%	90.26	91.23	99.10	99.39	99.18	**99.41**

点、0.92 百分点；在 Kappa 上分别提高了 16.15 百分点、17.27 百分点、0.57 百分点、0.22 百分点、1.00 百分点。由于 PU 数据集各类样本的数据量充足，因此本文方法的分类精度在不同地物之间的提升没有特别明显。从表 6 可以看出，本文方法在 PU 数据集上在 OA、AA、Kappa 这 3 个方面都达到了最优，分别为 99.56%、99.18%、99.41%。本文方法比 Baseline、1D-CNN、3D-CNN、3D+2D-CNN、Swin-T 在 OA 上分别提高了 6.90 百分点、6.16 百分点、0.24 百分点、0.02 百分点、0.18 百分点；在 AA 上分别提高了 9.03 百分点、8.37 百分点、0.79 百分点、0.23 百分点、0.46 百分点；在 Kappa 上分别提高了 9.15 百分点、8.18 百分点、0.31 百分点、0.02 百分点、0.23 百分点。

分类结果图如图 8 和图 9 所示，可以看出神经网络和 1D-CNN 存在明显的噪声，对于 3D-CNN、3D+2D-CNN、Swin-T 来说分类效果图对地物分类相对清晰，但 3D-CNN 等方法仍然存在相似地物错分现象，本文方法的分类结果与真实地物标签最为接近，几乎不存在噪声现象。从图中可以看出，本文方法确实提高了高光谱图像的分类精度。这与表 5～6 中的分类精度相符合。另外为验证多尺度光谱增强残差融合模块和光谱注意力模块的有效性，本文给出了 India 数据集不同模块的类别激活映射图(CAM)[27]，从图 10 中可以看出，不同模块所学习到的特征对最终分类结果影响是不一致的，证明了模块叠加使用的必要性。

3.3.3 消融实验

为验证移位窗口算法、多尺度光谱增强残差融合模块以及光谱注意力模块在高光谱地物分类任务上的有效性，本文进行了消融实验，消融实验结果如表 7 所示，表中 shift 表示移位窗口，w/o 表示去除模块，M 表示多尺度光谱增强残差融合模块，S 表示光谱注意力模块。在两个数据集上运用移位窗口划分的网络的性能都优于基于单个窗口划分的模型，在 India 数据集上的 OA、AA、Kappa 分别为 0.21%、0.73%、0.56%，在 PU 数据集上的 OA、AA、Kappa 分别为 0.19%、0.41%、0.85%。从表 7 中可以看出多尺度光谱增强残差融合模块的使用对模型的性能有重要影响，在 India 数据集上的 OA、AA、Kappa 分别为 0.50%、0.69%、0.73%，在 PU 数据集上的 OA、AA、Kappa 分别为 0.45%、0.13%、0.59%。光谱注意力模块对模型的性能也有一定的提升，在 India 数据集上的 OA、AA、Kappa 分别为 0.30%、0.50%、0.31%，在 PU 数据集上的 OA、AA、Kappa 分别为 0.28%、0.04%、0.54%。

结果表明，使用移位窗口在窗口之间建立连接是很有效的，同时利用多尺度光谱增强残差融合模块以及光谱注意力模块，能够使模型对高光谱地物分类的效果有显著提升。

3.3.4 模型性能分析

每秒所执行的浮点运算次数(FLOPs)是衡量深度学习模型计算复杂度的指标，其值越大，说明模型计算复杂度越高。对多个模型的参数量(param)和计算复杂度进行统计如表 8 所示(其中 MFLOPs 表示每秒百万个浮点操作)。相较于 1D-CNN 和神经网络，本

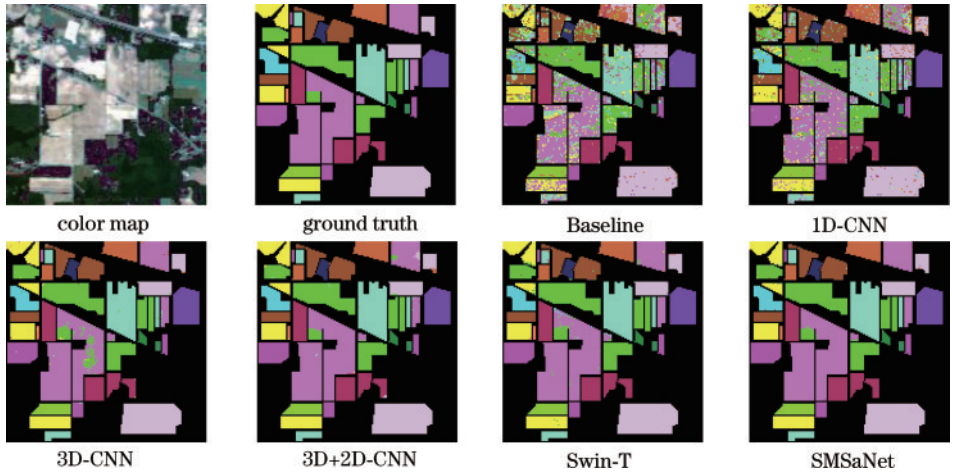

图 8　India数据集分类结果图

Fig. 8　Classification result chart on India dataset

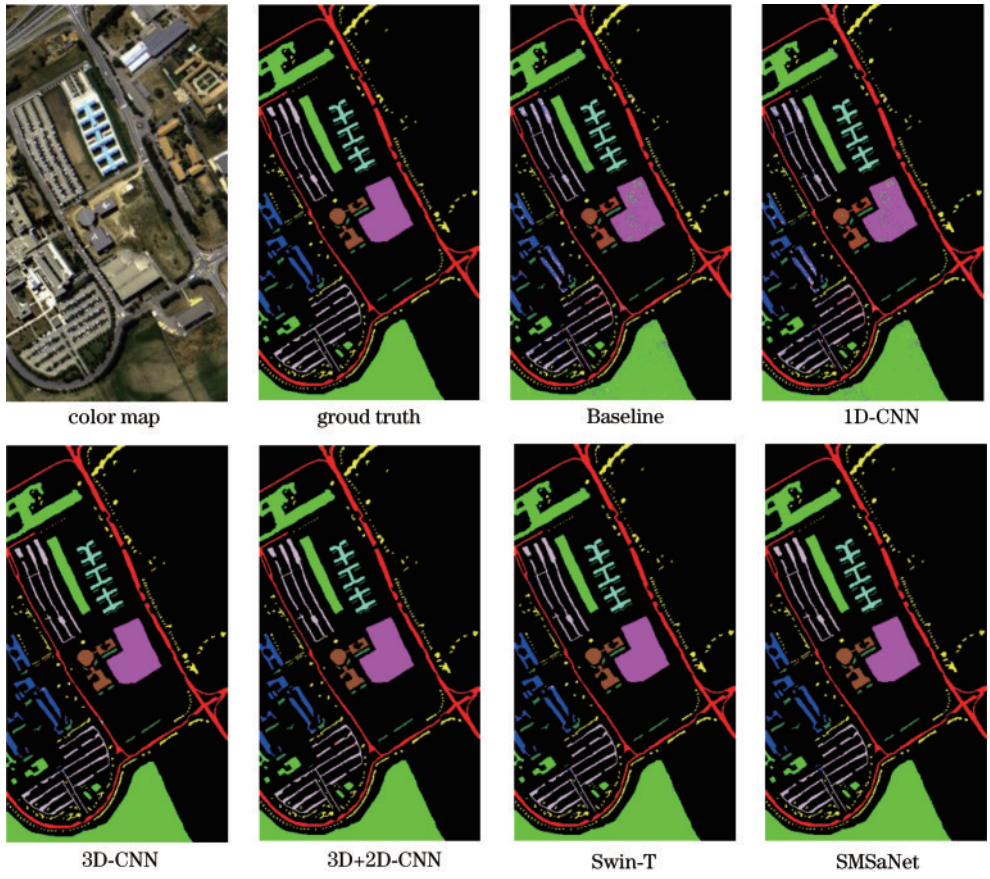

图 9　PU数据集分类结果图

Fig. 9　Classification result chart on PU dataset

文方法复杂度较高，参数量与神经网络相当；相较于3D-CNN，本文参数量较多，但模型的计算量较低；相较于3D+2D-CNN，本文方法在未增加参数量的前提下便能够取得较好结果，提高了分类精度，降低了计算复杂度；相较于Swin-T，本文方法大量减少了参数数量，降低了计算复杂度，且在多个指标上取得最好结果，因此，相比其他模型，本文方法在参数量适中的前提下可提高分类精度，降低计算复杂度。

3.3.5　不同训练样本测试

为了进一步证明SMSaNet在不同比率的训练样本下的有效性，本文对表1～2中所选的训练样本数量乘以固定比率作为实验的训练样本标准，其余作为测试样本。不同算法在小样本条件下的OA如图11所示，从图中可看出，即使在训练样本有限的情况下，本文所提出的SMSaNet仍然能够获得最高的分类精度，证明了SMSaNet在样本有限条件下的有效性与鲁棒性。

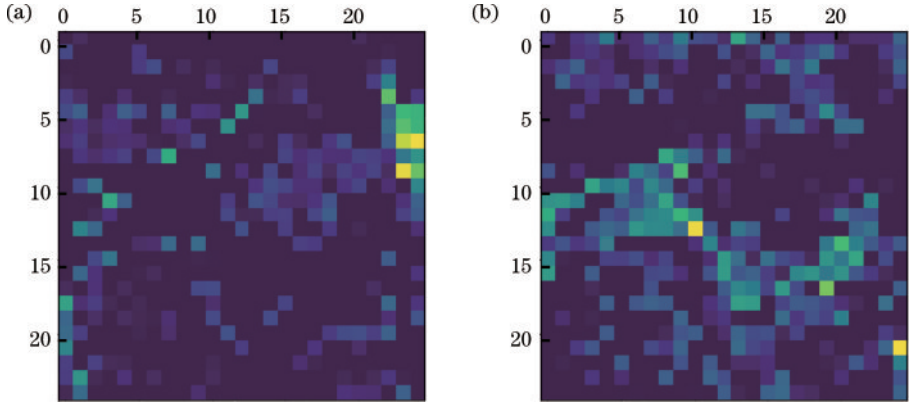

图 10　类别激活映射图(CAM)。(a)多尺度光谱增强残差融合模块的CAM;(b)光谱注意力模块的CAM

Fig. 10　Class activation mapping (CAM). (a) CAM of multiscale spectral enhanced residual fusion module; (b) CAM of spectral attention module

表 7　消融实验

Table 7　Ablation experiments

Method and module				India			PU		
Method	Shift	M	S	OA /%	AA /%	Kappa /%	OA /%	AA /%	Kappa /%
w/o shift		√	√	99.30	98.93	98.88	99.37	98.77	98.56
SMSaNet	√	√	√	99.51	99.66	99.44	99.56	99.18	99.41
w/o M	√		√	98.86	98.63	98.71	99.11	99.05	98.82
w/o S	√	√		99.21	99.16	99.13	99.28	99.09	98.87
w/o M and S	√			98.56	98.17	98.06	98.89	98.65	98.48

表 8　不同模型的参数量和FLOPs

Table 8　Params and FLOPs for different models

Method	Image size	Param /M	FLOPs
Baseline	1×1	4.230	4.23 MFLOPs
1D-CNN	1×1	0.036	0.11 MFLOPs
3D-CNN	25×25	0.772	93.82 MFLOPs
3D+2D-CNN	25×25	5.009	152.68 MFLOPs
Swin-T	25×25	27.495	89.78 MFLOPs
SMSaNet	25×25	5.056	46.65 MFLOPs

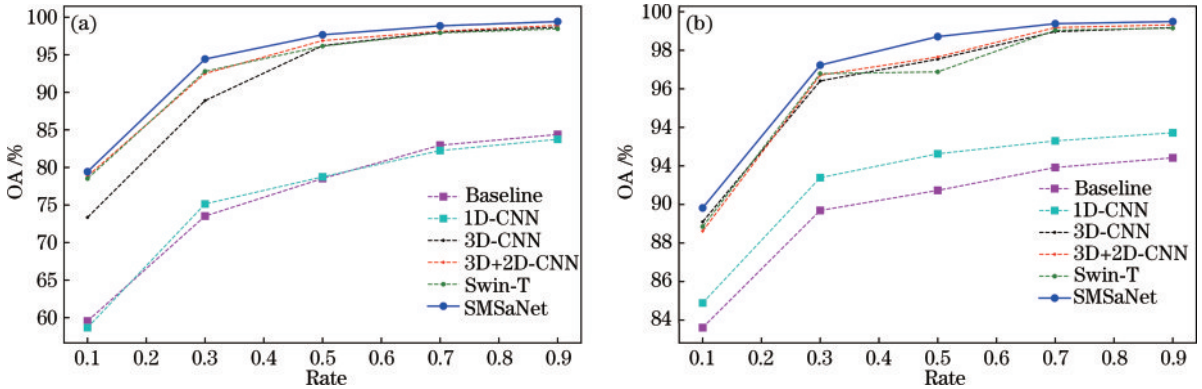

图 11　不同比率训练样本对应的OA值。(a)Inida数据集;(b)PU数据集

Fig. 11　OA values corresponding to different ratios of training samples. (a) Inida dataset; (b) PU dataset

4　结　　论

本文尝试将CNNs和视觉领域的Swin Transformer方法相结合,应用在小型数据集上,并取得了比较好的结果。实验结果表明,Swin Transformer模块能够利用注意力的方式充分提取空间特征,从而得到其空间上不同地物间的相关性。将Transformer当做特征提取器,再结合CNNs中的先验

知识能够有效解决高光谱分类问题，在公开数据集上使用较小的数据量无需过多地对数据进行处理就可以达到先进水平。结合多尺度光谱残差融合模块和Swin Transformer block将高光谱图像“图谱合一”的特点物尽其用，进一步优化空间特征与光谱特征，通过多种实验对Transformer与CNNs的混合使用进行了分析，实验结果表明，本文方法具有良好的分类性能和泛化能力。

参 考 文 献

[1] Sun W W, Liu K, Ren G B, et al. A simple and effective spectral-spatial method for mapping large-scale coastal wetlands using China ZY1-02D satellite hyperspectral images[J]. International Journal of Applied Earth Observation and Geoinformation, 2021, 104: 102572.

[2] Stuart M B, McGonigle A J S, Willmott J R. Hyperspectral imaging in environmental monitoring: a review of recent developments and technological advances in compact field deployable systems[J]. Sensors, 2019, 19(14): 3071.

[3] Su H J, Yao W J, Wu Z Y, et al. Kernel low-rank representation with elastic net for China coastal wetland land cover classification using GF-5 hyperspectral imagery [J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2021, 171: 238-252.

[4] Moughal T A. Hyperspectral image classification using Support Vector Machine[J]. Journal of Physics: Conference Series, 2013, 439(1): 012042.

[5] Petropoulos G P, Arvanitis K, Sigrimis N. Hyperion hyperspectral imagery analysis combined with machine learning classifiers for land use/cover mapping[J]. Expert Systems With Applications, 2012, 39(3): 3800-3809.

[6] Golhani K, Balasundram S K, Vadamalai G, et al. A review of neural networks in plant disease detection using hyperspectral data[J]. Information Processing in Agriculture, 2018, 5(3): 354-371.

[7] Chen Y S, Zhao X, Jia X P. Spectral-spatial classification of hyperspectral data based on deep belief network[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2015, 8(6): 2381-2392.

[8] Hong D F, Gao L R, Yao J, et al. Graph convolutional networks for hyperspectral image classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 59(7): 5966-5978.

[9] 冯凡，王双亭，张津，等. 基于多特征融合和混合卷积网络的高光谱图像分类[J]. 激光与光电子学进展，2021，58(8)：0810010.
Feng F, Wang S T, Zhang J, et al. Hyperspectral images classification based on multi-feature fusion and hybrid convolutional neural networks[J]. Laser & Optoelectronics Progress, 2021, 58(8): 0810010.

[10] Hommel B, Chapman C S, Cisek P, et al. No one knows what attention is[J]. Attention, Perception, & Psychophysics, 2019, 81(7): 2288-2303.

[11] 王欣，樊彦国. 基于改进DenseNet和空谱注意力机制的高光谱图像分类[J]. 激光与光电子学进展，2022，59(2)：0210014.
Wang X, Fan Y G. Hyperspectral image classification based on modified DenseNet and spatial spectrum attention mechanism[J]. Laser & Optoelectronics Progress, 2022, 59(2): 0210014.

[12] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[EB/OL]. (2017-10-06) [2022-04-15]. https://arxiv.org/abs/1706.03762.

[13] Qing Y H, Liu W Y, Feng L Y, et al. Improved transformer net for hyperspectral image classification[J]. Remote Sensing, 2021, 13(11): 2216.

[14] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale[EB/OL]. (2021-06-03) [2022-04-15]. https://arxiv.org/abs/2010.11929.

[15] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]// 2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2021: 9992-10002.

[16] Cao J Y, Chen Z, Wang B. Deep Convolutional networks with superpixel segmentation for hyperspectral image classification[C]//2016 IEEE International Geoscience and Remote Sensing Symposium, July 10-15, 2016, Beijing, China. New York: IEEE Press, 2016: 3310-3313.

[17] Liu Y G, Yu J Z, Han Y H. Understanding the effective receptive field in semantic image segmentation[J]. Multimedia Tools and Applications, 2018, 77(17): 22159-22171.

[18] Garbin C, Zhu X Q, Marques O. Dropout *vs.* batch normalization: an empirical study of their impact to deep learning[J]. Multimedia Tools and Applications, 2020, 79(19): 12777-12815.

[19] Sun W W, Shao W J, Peng J T, et al. Multiscale low-rank spatial features for hyperspectral image classification [J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 5501605.

[20] Raghu M, Unterthiner T, Kornblith S, et al. Do vision transformers see like convolutional neural networks? [EB/OL]. (2021-08-19) [2022-02-04]. https://arxiv.org/abs/2108.08810.

[21] Hu W, Huang Y Y, Wei L, et al. Deep convolutional neural networks for hyperspectral image classification[J]. Journal of Sensors, 2015, 2015: 258619.

[22] Hamida A B, Benoit A, Lambert P, et al. 3-D deep learning approach for remote sensing image classification [J]. IEEE Transactions on Geoscience and Remote Sensing, 2018, 56(8): 4420-4434.

[23] 廖金雷，张磊，周湘山，等. 融合植被指数的3D-2D-CNN高光谱图像植被分类方法[J]. 科学技术与工程，2021，21(27)：11656-11662.
Liao J L, Zhang L, Zhou X S, et al. A hyperspectral image vegetation classification method using 2D-3D CNNs and vegetation index[J]. Science Technology and

Engineering, 2021, 21(27): 11656-11662.

[24] 徐沁, 梁玉莲, 王冬越, 等. 基于SE-Res2Net与多尺度空谱融合注意力机制的高光谱图像分类[J]. 计算机辅助设计与图形学学报, 2021, 33(11): 1726-1734.
Xu Q, Liang Y L, Wang D Y, et al. Hyperspectral image classification based on SE-Res2Net and multi-scale spatial spectral fusion attention mechanism[J]. Journal of Computer-Aided Design & Computer Graphics, 2021, 33(11): 1726-1734.

[25] Touvron H, Vedaldi A, Douze M, et al. Fixing the train-test resolution discrepancy: FixEfficientNet[EB/OL]. (2020-11-18) [2022-04-15]. https://arxiv. org/abs/2003.08237.

[26] Ji S W, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221-231.

[27] Zhou B L, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 2921-2929.