文章编号　1004-924X(2024)23-3504-09

3D-CNN与Transformer混合结构的高光谱图像空谱联合分类

景海钊，陶丽杰，张号逵*
（西北工业大学，陕西 西安 710129）

摘要：针对高光谱图像（Hyperspectral Images，HSI）地面覆盖类的像素级分类问题，提出一种3D-ConvFormer的混合结构模型，该模型通过在浅层使用三维卷积（3D-CNN）操作提取高光谱图像的局部空间光谱特征，在深层利用自注意力（Slef-attention）机制在卷积窗口内提取空间光谱特征，实现了卷积网络的平移不变性与self-attention对特征的灵活提取能力的有效融合。在Indian Pines，PaviaU和WHU Hi Longkou 3组公开的高光谱图像数据集上进行实验，采用总体分类精度（OA）、平均分类精度（AA）和Kappa系数3个指标，对地物类别的像素级分类结果进行量化评估。实验结果表明，模型在Indian Pines数据集上OA为98.41%，AA为97.56%，Kappa为98.16%；在PaviaU数据集上OA为99.39%，AA为99.30%，Kappa为99.18%；在WHU-Hi-Longkou数据集上OA为98.53%，AA为98.97%，Kappa为98.06%。模型在3组高光谱图像分类任务中展示出的性能均优于对比的模型方法，取得了良好的分类性能，有效提升高光谱图像的分类精度。

关　键　词：计算机视觉；高光谱图像；卷积神经网络；视频转换；自注意力机制
中图分类号：TP394.1；TH691.9　　**文献标识码**：A　　**doi**：10.37188/OPE.20243223.3504

Specral-spatial classification of hyperspectral imagery with hybrid architecture of 3D-CNN and Transformer

JING Haizhao，TAO Lijie，ZHANG Haokui*

（*Northwestern Polytechnical University*，*Xi'an* 710129，*China*）
* *Corresponding author*，*E-mail*：*hkzhang@nwpu.edu.cn*

Abstract：To address pixel-level land cover classification in hyperspectral images（HSI），a hybrid model 3D-ConvFormer is proposed. The model integrates 3D convolutional neural networks（3D-CNN）and self-attention mechanisms to effectively extract spatial-spectral features. In the shallow layers，3D-CNN operations capture local spatial-spectral features，while in the deeper layers，the self-attention mechanism operates within convolutional windows to enhance feature extraction flexibility. This design achieves a synergistic fusion of the translation invariance of convolutional networks and the adaptive feature extraction capabilities of self-attention. The model's performance was evaluated on three publicly available hyperspectral image datasets—Indian Pines，PaviaU，and WHU-Hi-Longkou—using three metrics：Overall Accuracy（OA），Average Accuracy（AA），and the Kappa coefficient. Experimental results demonstrate that the

收稿日期：2024-09-30；**修订日期**：2024-10-28.
基金项目：国家自然科学基金资助项目（No. 62401471）；2024年姑苏创新创业领军人才计划（青年创新领军人才）资助项目（No. ZXL2024333）

proposed model achieved an OA of 98.41%, AA of 97.56%, and Kappa of 98.16% on the Indian Pines dataset; an OA of 99.39%, AA of 99.30%, and Kappa of 99.18% on the PaviaU dataset; and an OA of 98.53%, AA of 98.97%, and Kappa of 98.06% on the WHU-Hi-Longkou dataset. Compared to baseline models, 3D-ConvFormer consistently outperformed in classification tasks across all three datasets, significantly improving the accuracy of hyperspectral image classification.

Key words: computer vision; hyperspectral image; 3D Convolutional Neural Network (CNN); vision transformer; self-attention

1 引　言

高光谱图像(Hyperspectral Image, HSI)能够捕获连续波段的丰富光谱信息，在遥感[1-2]、环境监测[3]和农业评估[4]等领域具有重要意义。然而，HSI数据的高维性和复杂性给准确高效的分类带来了挑战[5]。随着深度学习的发展，卷积神经网络(Convolutional Neural Network, CNN)在HSI分类中取得了显著进展。早期方法主要利用一维卷积神经网络(1D-CNN)和二维卷积神经网络(2D-CNN)分别提取光谱和空间特征[6-7]。随着HSI三维数据结构研究的深入，三维卷积神经网络(3D-CNN)开始受到关注。3D-CNN的卷积核在空间和光谱维度上同时滑动，全面考虑像素的空间位置和光谱信息，提升了分类的准确性。Liu等设计了ConvNext模型，重新制定了激活函数和正则化策略，展示了三维卷积神经网络在各类任务中的竞争力[8]。Zhang等在对3D-CNN轻量化的基础上，提出一种名为LWNet模型，使用更深但参数更少的网络实现了更优的分类性能[9]。与1D-CNN和2D-CNN相比，3D-CNN的结构更适合处理HSI的三维数据，同时提取光谱和空间特征，避免了由降维等方法带来的信息损失，广泛应用于高光谱图像分类。

近些年，Dosovitskiy等将Transformer引入图像处理领域，提出了视频转换(Vision Transformer, ViT)模型。ViT在计算量的限制下，将输入图像切割为图块，转换为序列，输入Transformer进行特征提取，并引入位置嵌入和分类标记进行最终分类[10]。Transformer在长距离依赖关系上表现出良好的建模能力，广泛应用于HSI分类任务中。Hong等提出使用SpectralFormer模型，通过多层自注意力机制建模高光谱图像的光谱和空间特征，并结合位置编码有效提升分类性能[11]。Ayas等提出了SpectralSWIN网络，通过新设计的Swin-Spectral模块同时处理高光谱图像的空间和光谱特征[12]。在这些工作中，ViT都展示出对于HSI数据的光谱、空间特征的良好提取性能。

CNN由于其强大的归纳偏置，尤其是其平移不变性，能够有效学习早期的浅层语义信息，但由于卷积操作的灵活性一般，在处理复杂的三维数据时，会在一定程度上限制卷积网络的性能。而自注意力机制虽然能够灵活地处理远程依赖关系，但其计算复杂度随token数量呈二次方增长，直接应用于高分辨率图像时会造成计算复杂度升高等问题。因此，在视觉任务中采用CNN和Transformer的混合结构得到了研究人员的广泛关注。Parmar等指出，在模型的浅层阶段使用注意力机制而在深层阶段使用卷积，即使参数数量显著增加，模型性能也可能下降[13]，这说明卷积在捕获浅层特征方面表现更好，而注意力层则擅长整合全局信息。在MobileViT中，MobileNetv2的深层结构被MobileViT块替代，进一步提升了性能[14]。ParCNet中的位置感知循环卷积(ParC)则结合传统卷积设计了一种混合结构，有效捕获了全局特征[15]。近年来，混合结构也被应用于HSI分类任务中。Sun等提出的SSFTT采用3D-CNN提取浅层光谱特征，并通过2D-CNN提取浅层空间特征，之后将特征输入Transformer模块进行特征学习[16]。Hong等在SpectralFormer模型中通过集成Transformer与卷积来有效学习光谱空间关系[11]。综上，本文提出了一种3D-CNN与Transformer结合(3D-ConvFormer)的高光谱图像分类模型，该模型在3组公开的高光谱图像数据集的分类任务上展现出良好的稳定性。为了提高卷积操作在处理复杂3D数据时的灵活性，同时更加精准地提取HSI数据的空间和光谱特征，提出在卷积操作的窗口

内进行自注意力操作。具体来说，模型在浅层采用3D-CNN来提取局部空间光谱特征，随后在特征提取的窗口内实施注意力机制。该方法通过类似3D-CNN的操作处理输入数据，使得模型既保留了卷积网络的平移不变性，又在窗口中融合了自注意力对特征提取的灵活性，实现对空谱特征的有效提取，提升HSI地面覆盖类像素级分类的精度。

2　原　理

2.1　3D-ConvFormer模型架构

图1为3D-ConvFormer框架的流程。如图1所示，HSI分类框架主要分为3个部分：样本提取阶段、3D-CNN与自注意力结合的混合结构特征提取阶段，以及基于全局平均池化（Global Average Pooling，GAP）的分类模块。

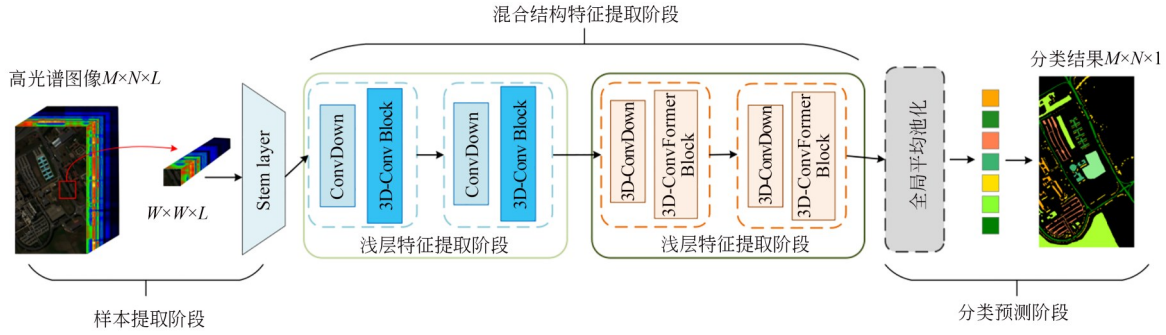


图1　3D-ConvFormer框架流程

Fig. 1　Workflow of 3D-ConvFormer framework

样本提取阶段，从HSI数据中抽取空间大小为W、光谱带数为L的立方块作为样本，样本尺寸为$W \times W \times L$。每个立方体块是以HSI数据中的某一点为中心像素，在该像素的邻域窗口内进行提取。训练中，为了确保立方体块标签的准确性，W设置为奇数。混合特征提取阶段，使用基于3D-CNN的模块对浅层特征进行初步提取；在深层特征提取阶段，设计了一种3D-ConvFormer Block（3D-CF Block）的混合模块，该模块结合3D-CNN和ViT对深层语义特征进行进一步提取。在分类阶段，设计采用基于全局平均池化分类模块预测分类结果。

首先，将样本提取阶段得到的立方体块输入Stem层，对光谱带数量进行压缩，并初步提取光谱信息。在浅层特征提取阶段，通过卷积下采样操作降低分辨率并增加通道数，减少信息丢失，同时在3D-Conv block中提取初级特征。在深层特征提取阶段，将原有的3D-Conv block替换为3D-ConvFormer block，通过自注意力机制灵活捕获窗口中的依赖关系。每个阶段的卷积下采样操作同时作用于光谱和空间分辨率，具体的下采样率分别为2，4，8，16。实验结果表明，3D-ConvFormer block在深层特征提取过程中能够更好地学习高级语义信息，提高模型对复杂特征的表达能力。在最终的分类阶段，利用全局平均池化将特征图压缩成固定长度的向量，使模型能够获取全局特征，从而提升分类效率，并增强模型对位置变化和未见过数据的泛化能力。

2.2　3D-ConvFormer Block

图2为本文提出的3D-ConvFormer Block中所包含的特征提取操作的具体流程。该模块结合卷积操作和自注意力机制的优势，能够更好地提取高光谱图像3D数据中的特征信息。具体而言，它采用与卷积相同的滑动重叠窗口机制，窗口大小设置为3。在每个提取出的窗口内，将数据展开后进行自注意力操作，从而赋予网络与self-attention相同的灵活性。此外，在重叠的小窗口中顺序执行attention操作，确保网络继承卷积网络的平移不变性，从而有效提取HSI数据中的深层特征。

首先，3D卷积通过聚合输入特征中每个元素的局部区域信息来实现特征提取。它通过滑动窗口操作在输入X上执行逐步提取，能够有效捕捉局部空间和光谱特征。同时，卷积层的参数在整个输入上共享，这显著减少了模型的参数数量。更重要的是，卷积的平移不变性使得模型能

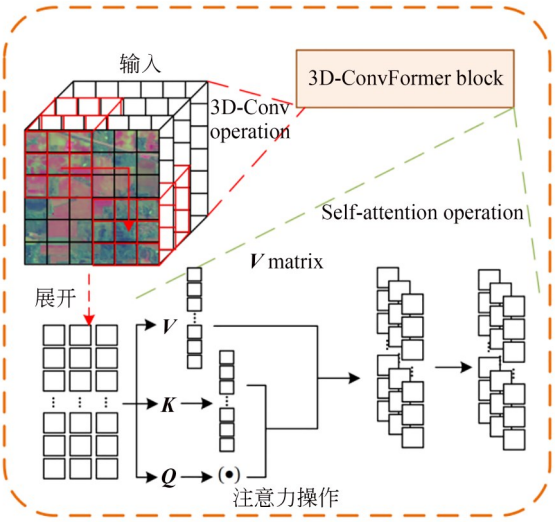


图 2　3D-ConvFormer Block 操作

Fig. 2　Operation of 3D-ConvFormer block

够处理输入中不同位置的特征，从而增强模型的泛化能力，使其能够更好地适应新数据。3D 卷积的操作可以表示为：

$$Y(i,j,k)=\sum_{h}\sum_{w}\sum_{d}X(i+h,j+w,k+d)\cdot K(h,w,d),\quad(1)$$

式中：Y 是输出特征图，X 是输入图像，K 是卷积核。

在展开数据后进行自注意力操作。自注意力机制通过引入 3 个可学习的权重矩阵 $\boldsymbol{W}^{Q}$，$\boldsymbol{W}^{K}$ 和 $\boldsymbol{W}^{V}$ 来分别对应 Query，Key 和 Value。每个输入标记都会通过这些矩阵线性映射，生成查询矩阵 $\boldsymbol{Q}$、键矩阵 $\boldsymbol{K}$ 和值矩阵 $\boldsymbol{V}$。自注意力的核心在于通过查询矩阵和键矩阵之间的点积来计算注意力分数，并通过 softmax 函数归一化这些分数，从而建立输入数据内的依赖关系。这一过程可以表示为：

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=soft\max\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V},\quad(2)$$

式中 $\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}$ 是查询和键的点积，表示它们之间的相关性，接着除以 $\sqrt{d_k}$ 进行缩放，然后通过 softmax 函数得到注意力权重。最终，3D-ConvFormer Block 根据这些依赖关系对值矩阵进行加权求和，以聚合特征信息，从而增强特征的表达能力。这一过程不仅帮助模型捕捉全局特征，还能有效建模各特征之间的关系。

3　实验结果与分析

3.1　实验数据集

在对比实验中，选取 Indian Pines[17]，Pavia University[18] 和 WHU-Hi-LongKou[19] 数据集，将上述 3 个数据集划分为训练集和测试集，分别如表 1～表 3 所示。对于 Pavia University 和 WHU-Hi-LongKou 数据集，从每个类别中随机抽取 150 个样本作为训练集，其他作为测试集。而在 Indi-

表 1　Indian Pines 数据集土地覆盖类别

Tab. 1　Land cover class in Indian Pines dataset

Class	Train num	Test num
Alfalfa	46	10
Corn-notill	1 428	150
Corn-mintill	830	150
Corn	237	10
Grass-pasture	483	150
Grass-trees	730	150
Grass-pasture-mowed	28	10
Hay-windrowed	478	150
Oats	20	10
Soybean-notill	972	150
Soybean-mintill	2 455	150
Soybean-clean	593	150
Wheat	205	10
Woods	1 265	150
Buildings-Grass-Trees-Drives	386	150
Stone-Steel-Towers	93	10
Total	10 249	1 560

表 2　Pavia University 数据集土地覆盖类别详情

Tab. 2　Land cover class in Pavia University dataset

Class	Train num	Test num
Asphalt	6 631	150
Meadows	18 649	150
Gravel	2 099	150
Trees	3 064	150
Painted metal sheets	1 345	150
Bare Soil	5 029	150
Bitumen	1 330	150
Self-Blocking Bricks	3 682	150
Shadows	947	150
Total	42 776	1 350

表3　WHU-Hi-LongKou数据集土地覆盖类别详情

Tab. 3　Land cover class in WHU-Hi-LongKou dataset

Class	Train num	Test num
Corn	34 511	150
Cotton	8 374	150
Sesame	3 031	150
Broad-leaf soybean	63 212	150
Narrow-leaf soybean	4 151	150
Rice	11 854	150
Water	67 056	150
Roads and houses	7 124	150
Mixed weed	5 229	150
Total	204 542	1 350

an Pines数据集上，有6个类别的总样本数均小于300个，分别是Alfalfa，Corn，Grass-pasture-mowed，Oats，Wheat和Stone-Steel-Towers。对于这6个类别，随机抽取10个样本作为训练集，其他类别中随机抽取150个样本作为训练集，其余样本用作测试集。

3.1.1　Indian Pines数据集

Indian Pines数据集是1992年由机载可见/红外成像光谱仪（AVIRIS）传感器在美国印第安纳州的一片农业区获取的高光谱遥感图像数据。该数据集包含145×145像素的高光谱图像，每个像素拥有224个光谱波段，波长为0.4～2.5 μm。由于部分波段位于水汽吸收区，通常有24个波段会被移除，剩下的200个波段用于16种地面覆盖物的分类研究。

3.1.2　Pavia University数据集

Pavia University数据集是在意大利Pavia市的Pavia University区域获取的。该数据集由机载可见/红外成像光谱仪（ROSIS）传感器采集，图像空间分辨率为1.3 m，覆盖610×340像素的区域。该数据集包含103个有效的光谱波段，波长为0.43～0.86 μm，涵盖从可见光到近红外区域的电磁波谱。由于水汽吸收等原因，部分波段被移除，最终保留的波段具有较高的光谱信息密度。数据集标注了9类不同的地物类型，包括草地、树木、沥青、砖瓦、金属屋顶和阴影等。

3.1.3　WHU-Hi-LongKou数据集

WHU-Hi-LongKou数据集是2018年7月17日在中国湖北省龙口镇采集的高光谱遥感数据。数据采集使用安装在DJI Matrice 600 Pro（DJI M600 Pro）无人机平台上的Headwall Nano-Hyperspec成像传感器，焦距为8 mm。数据采集区域为简单的农业场景，包含6种作物：玉米、棉花、芝麻、宽叶大豆、窄叶大豆和水稻。所获取的图像尺寸为550×400像素，包含400～1 000 nm的270个光谱波段。

3.2　实验设置

实验在运行Linux操作系统的服务器上进行，该服务器配备了Intel(R) Xeon(R) E5-2640 v4 @ 2.40 GHz处理器和Nvidia GeForce RTX 2080 Ti显卡，并使用PyTorch 2.3.1框架进行实现。在训练过程中，3个数据集均被裁剪为空间分辨率为27×27的三维立方体作为输入。所有数据集的批次大小均为64，训练总共进行300个epoch。使用AdamW优化器，学习率初始值为5×10^{-4}，并结合余弦退火策略，权重衰减率设置为1×10^{-5}。前10%的epoch作为预热阶段，学习率从初始值的10%逐渐增加至初始值，在接下来的90% epoch中学习率逐渐降低。

3.3　对比实验结果及分析

在对比实验中，选取4种性能表现良好的HSI分类方法与3D-ConvFormer进行了比较，具体为3D-CNN[6]，LWNet[9]，SSFTT[16]和Spectral-Former[11]。使用Indian Pines，Pavia University和WHU Hi LongKou 3个主流数据集评估这些模型的分类性能。实验采用HSI分类领域常用的3项关键指标：总体精度（OA）、平均精度（AA）和Kappa系数。这些指标能够更加系统地评估各模型在不同数据集及各类上的分类表现，从而确保实验结果的准确性与全面性。

如表4所示，在Indian Pine数据集的分类预测中，3D-ConvFormer的OA，AA和Kappa系数分别为98.41%，97.56%和98.16%。在该数据集上，与其他组实验结果相比，本文方法在这3个评估指标上均表现最佳。图3(f)展示了3D-ConvFormer在Indian Pine数据集上的可视化分类结果。与其他4种方法相比，3D-ConvFormer在类别边界的清晰度和细节表现上更为出色，其分类结果也更接近Ground Truth，进一步证明了该方法在处理细节方面的优越性。

表 4　在 Indian Pines 数据集上对比方法的分类结果

Tab. 4　Comparison of classification results of different methods on Indian Pines dataset

Class	3D-CNN	LWNet	SpectralFormer	SSFTT	3D-ConvFormer
C1	97.22	100.00	97.83	100.00	91.67
C2	97.81	98.28	91.39	95.07	95.93
C3	98.82	99.85	96.02	100.00	99.56
C4	93.39	82.82	47.68	92.51	91.63
C5	100.00	100.00	98.34	99.70	100.00
C6	100.00	97.59	98.90	99.48	100.00
C7	100.00	100.00	100.00	100.00	100.00
C8	100.00	100.00	100.00	100.00	100.00
C9	100.00	100.00	100.00	100.00	100.00
C10	97.93	99.39	94.34	99.15	99.03
C11	98.57	99.05	84.97	99.22	99.13
C12	99.77	97.52	98.31	95.26	99.77
C13	85.64	82.56	71.22	85.64	89.74
C14	100.00	99.91	98.81	100.00	99.28
C15	100.00	100.00	98.70	100.00	100.00
C16	86.75	92.77	87.10	78.31	95.18
OA(%)	98.37	98.22	91.98	96.52	**98.41**
AA(%)	97.24	96.86	91.48	97.97	**97.56**
KAPPA(%)	98.11	97.94	90.89	97.65	**98.16**

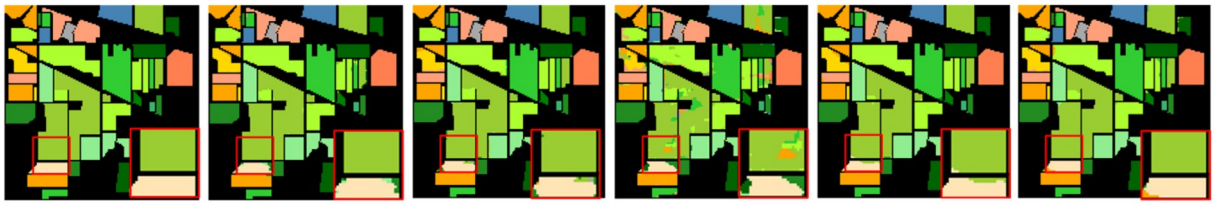


图 3　在 Indian Pine 数据集上不同方法的可视化预测结果

Fig. 3　Visualization results of different methods on Indian Pine dataset

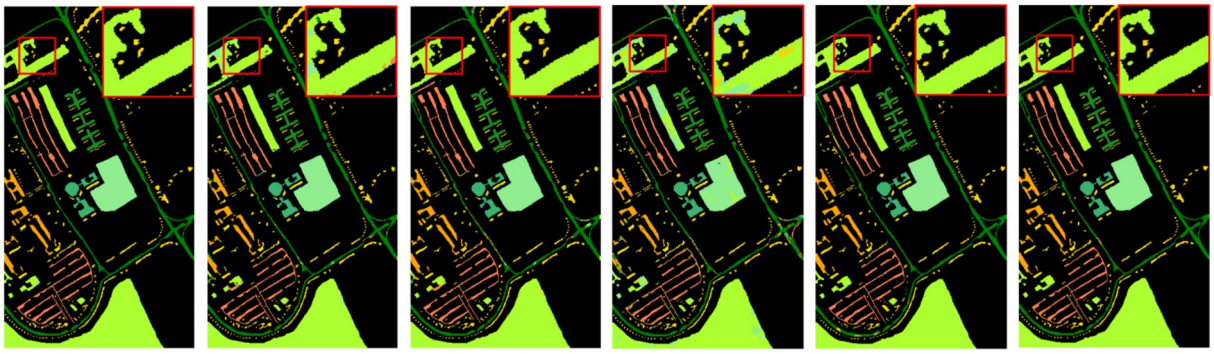
如表 5 所示，在 Pavia University 数据集上 3D-ConvFormer 的 OA，AA 和 Kappa 系数分别为 99.39%，99.30% 和 99.18%。卷积方法中表现最好的是 LWNet，OA 为 97.20%，ViT 方法中表现最好的是 SSFTT，OA 值为 96.11%。相比之下，3D-ConvFormer 比这两种方法分别高出 2.19% 和 3.28%。图 4(f) 展示了 3D-ConvFormer 在该数据集上的分类可视化结果，它在地物分类任务中表现出色。与其他 4 个对照组相比，3D-ConvFormer 的分类可视化结果呈现出更少的误分类现象，尤其在特定区域，不同类别的像素颜色区分更加显著。这表明 3D-ConvFormer 模型不仅在大范围分类任务中表现优异，同时在捕捉细微地物特征和处理边缘细节上也具有极高的精确性。

同样地，在 WHU-Hi-LongKou 数据集上，3D-ConvFormer 继续展现出卓越的分类性能。如表 6 所示，该模型的 OA，AA 和 Kappa 系数分别为 98.53%，98.97% 和 98.07%。相对于 LWNet 和 SSFTT 在预测 OA 值上分别提升了 2.11% 和 0.41%。在所有评价指标上，3D-ConvFormer 优于所有对照组模型。图 5 显示了 3D-ConvFormer 在该数据集上的分类可视化结果。可以看到，边界分割更加清晰有效，尤其在特定的农业场景中，框选区域的像素分类更加一致，

表 5　在 Pavia University 数据集上对比方法的分类结果

Tab. 5　Comparison of classification results of different methods on Pavia University dataset

Class	3D-CNN	LWNet	SpectralFormer	SSFTT	3D-ConvFormer
C1	96.62	93.18	88.80	96.76	99.15
C2	94.37	97.71	84.10	99.34	99.70
C3	91.74	97.69	79.66	98.92	99.69
C4	96.12	93.45	94.88	81.67	97.25
C5	100.00	99.83	99.78	98.24	99.92
C6	99.80	100.00	91.41	99.92	100.00
C7	99.92	99.92	95.34	99.49	100.00
C8	96.72	99.04	89.35	99.52	98.58
C9	97.62	97.24	99.79	91.09	99.37
OA(%)	95.94	97.20	87.88	96.11	**99.39**
AA(%)	96.99	97.56	91.45	97.57	**99.30**
KAPPA(%)	94.62	96.26	84.34	96.74	**99.18**



(a) Ground truth　(b) 3D-CNN　(c) LWNet　(d) SpectralFormer　(e) SSFTT　(f) 3D-ConvFormer

图 4　在 Pavia University 数据集上不同方法的可视化预测结果

Fig. 4　Visualization results of different methods on Pavia University dataset

表 6　在 WHU-Hi-LongKou 数据集上对比方法的分类结果

Tab. 6　Comparison of classification results of different methods on WHU-Hi-LongKou dataset

Class	3D-CNN	LWNet	SpectralFormer	SSFTT	3D-ConvFormer
C1	98.00	97.05	98.91	99.56	99.66
C2	98.20	98.54	60.45	99.29	99.77
C3	99.79	99.93	98.81	100.00	99.90
C4	91.47	93.18	84.90	98.02	96.29
C5	99.58	99.93	86.27	99.48	99.63
C6	98.79	98.97	98.76	98.10	98.89
C7	98.64	98.38	98.86	96.81	99.73
C8	92.63	92.47	93.95	95.23	98.48
C9	98.41	97.99	98.41	96.63	98.39
OA(%)	96.12	96.42	92.53	98.12	**98.53**
AA(%)	97.28	97.38	91.03	97.68	**98.97**
KAPPA(%)	94.94	95.32	90.37	97.20	**98.07**

且不同类别之间的像素分布区分明显，具有更好的可分离性。整体上，该方法的分类结果显示出极少的误分类现象，进一步证明了它在复杂场景下的可靠性与精确度。

在3组复杂度不同的数据集上进行实验，结果如图6所示。由图可知，3D-ConvFormer的性能最佳，能够有效提取特征，同时具有良好的稳定性和泛化性。

通过在不同场景复杂度和类型的3个公开数据集上进行对比分析，3D-ConvFormer的捕捉细微差异能力更强，尤其在类别边界的处理上体现了更高的精度。

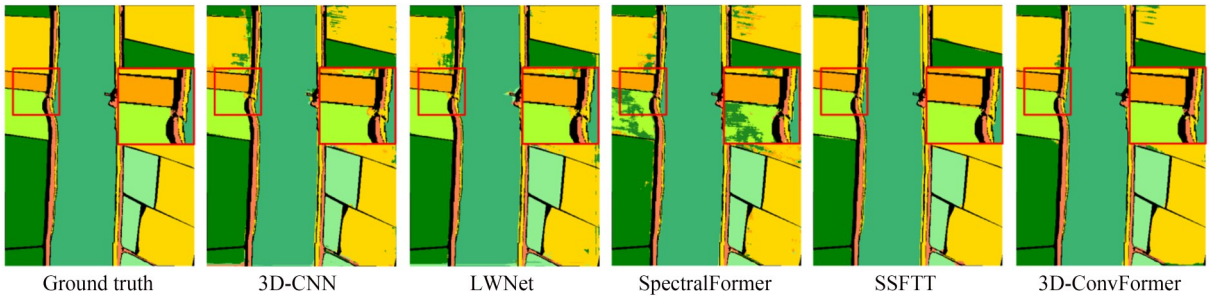


图5　在WHU-Hi-LongKou数据集上不同方法的可视化预测结果

Fig. 5　Visualization results of different methods on WHU-Hi-LongKou dataset

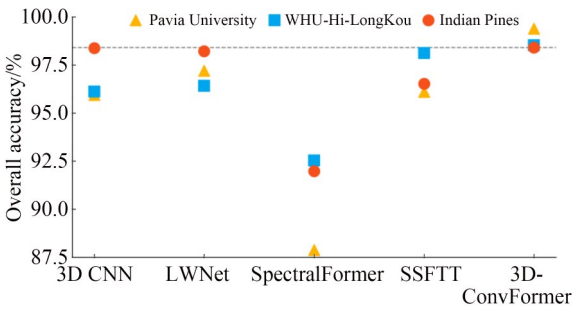


图6　在3组数据集上5种方法预测的总体精度的散点图

Fig. 6　Scatter plot of overall accuracy predicted by five methods on three datasets

4　结　论

本文提出了3D-ConvFormer模型架构，融合卷积操作和自注意力机制，对高光谱图像空间和光谱特征进行有效提取，实现了高光谱图像的地面覆盖像素级分类。通过在浅层利用3D-CNN提取局部空间和光谱特征，同时在深层结合自注意力机制的灵活特征提取能力，有效融合两种操作机制，优化了模型在处理HSI复杂3D数据时的性能。模型在Indian Pines数据集上的OA为98.41%，AA为97.56%，Kappa为98.16%；在PaviaU数据集上的OA为99.39%，AA为99.30%，Kappa为99.18%；在WHU Hi Longkou数据集上的OA为98.53%，AA为98.97%，Kappa为98.06%，各项指标均优于对比方法。实验结果表明，模型在多组数据集上的性能始终最优，证明了其强大的稳定性。该模型有效融合了卷积网络的平移不变性和自注意力机制的灵活性，为高光谱图像分类提供了一种性能优越、稳定性强的解决方案。

参考文献：

[1] 闫敬文，陈宏达，刘蕾. 高光谱图像分类的研究进展[J]. 光学 精密工程，2019，27(3)：680-693.
YAN J W，CHEN H D，LIU L. Overview of hyperspectral image classification [J]. *Opt. Precision Eng.*，2019，27(3)：680-693.(in Chinese)

[2] LI Y，ZHANG H K，XUE X Z，*et al*. Deep learning for remote sensing image classification：a survey [J]. *WIREs Data Mining and Knowledge Discovery*，2018，8(6)：e1264.

[3] 张达，郑玉权. 高光谱遥感的发展与应用[J]. 光学与光电技术，2013，11(3)：67-73.
ZHANG D，ZHENG Y Q. Hyperspectral remote sensing and its development and application review [J]. *Optics & Optoelectronic Technology*，2013，11

(3): 67-73. (in Chinese)

[4] 黄鸿, 郑新磊. 加权空-谱与最近邻分类器相结合的高光谱图像分类[J]. 光学 精密工程, 2016, 24(4): 873-881.
HUANG H, ZHENG X L. Hyperspectral image classification with combination of weighted spatial-spectral and KNN[J]. *Opt. Precision Eng.*, 2016, 24(4): 873-881. (in Chinese)

[5] FAUVEL M, TARABALKA Y, BENEDIKTSSON J A, *et al*. Advances in spectral-spatial classification of hyperspectral images[J]. *Proceedings of the IEEE*, 2013, 101(3): 652-675.

[6] CHEN Y S, JIANG H L, LI C Y, *et al*. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2016, 54(10): 6232-6251.

[7] LI Y, ZHANG H K, SHEN Q. Spectral-spatial classification of hyperspectral imagery with 3D convolutional neural network[J]. *Remote Sensing*, 2017, 9(1): 67.

[8] LIU Z, MAO H Z, WU C Y, *et al*. A ConvNet for the 2020s[C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June* 18-24, 2022. *New Orleans, LA, USA*. IEEE, 2022: 11976-11986.

[9] ZHANG H K, LI Y, JIANG Y N, *et al*. Hyperspectral classification based on lightweight 3-D-CNN with transfer learning[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2019, 57(8): 5813-5828.

[10] DOSOVITSKIY A. An image is worth 16x16 words: Transformers for image recognition at scale[J]. *arXiv preprint* arXiv:2010.11929, 2020.

[11] HONG D F, HAN Z, YAO J, *et al*. SpectralFormer: rethinking hyperspectral image classification with transformers[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 1-15.

[12] AYAS S, TUNC-GORMUS E. SpectralSWIN: a spectral-swin transformer network for hyperspectral image classification[J]. *International Journal of Remote Sensing*, 2022, 43(11): 4025-4044.

[13] PARMAR N, VASWANI A, USZKOREIT J, *et al*. *Image Transformer*[EB/OL]. 2018: 1802.05751. https://arxiv.org/abs/1802.05751v3.

[14] MEHTA S, RASTEGARI M. *MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer*[EB/OL]. 2021: 2110.02178. https://arxiv.org/abs/2110.02178v2.

[15] ZHANG H K, HU W Z, WANG X Y. *ParC-Net: Position Aware Circular Convolution with Merits from ConvNets and Transformer*[M]. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022: 613-630.

[16] SUN L, ZHAO G R, ZHENG Y H, *et al*. Spectral-spatial feature tokenization transformer for hyperspectral image classification[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 3144158.

[17] BAUMGARDNER M, BIEHL L, LANDGREBE D. 220 band AVIRIS hyperspectral image data set: June 12, 1992 Indian pine test site 3[J]. *Purdue University Research Repository*, 2015, 10(7): 991.

[18] LICCIARDI G, PACIFICI F, TUIA D, *et al*. Decision fusion for the classification of hyperspectral data: outcome of the 2008 GRS-S data fusion contest[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2009, 47(11): 3857-3865.

[19] ZHONG Y F, HU X, LUO C, *et al*. WHU-Hi: UAV-borne hyperspectral with high spatial resolution (H2) benchmark datasets and classifier for precise crop identification based on deep convolutional neural network with CRF[J]. *Remote Sensing of Environment*, 2020, 250: 112012.

作者简介：

景海钊(2000—),男,陕西西安人,硕士研究生,2023年于陕西科技大学获得学士学位,主要从事深度学习、机器视觉与图像处理方面的研究。E-mail: haizhao_jing@mail.nwpu.edu.cn

通讯作者：

张号逵(1991—),男,陕西西安人,博士,教授,博士生导师,2021年于西北工业大学获得博士学位,主要从事深度学习、计算机视觉等方面的研究。E-mail: hkzhang@nwpu.edu.cn