ADC-CPANet:一种局部—全局特征融合的遥感图像分类方法

王威，李希杰，王新

长沙理工大学 计算机与通信工程学院, 长沙 410114

摘 要：遥感图像具有丰富的纹理信息和复杂的整体结构，因此在场景分类任务中进行多尺度的特征提取至关重要。基于此，设计了局部特征提取模块ADC模块ADC （Aggregation Depthwise Convolution Block）和全局—局部特征提取模块CPA模块CPA（Convolution Parallel Attention Block），并在ADC模块中提出一种非对称深度卷积组，以增强模型对图像翻转和旋转的鲁棒性；在CPA模块中提出一种能够扩大感受野并增强特征提取能力的多分组卷积头分解注意力。以ADC模块和CPA模块为基础构建全新的遥感图像场景分类模型ADC-CPANet，在各阶段采用堆叠ADC模块和CPA模块的策略，从而使模型具有更好的全局特征和局部特征提取能力。为验证ADC-CPANet的有效性，本文使用开源数据集RSSCN7数据集和SIRI-WHU数据集测试ADC-CPANet与其他深度学习网络的复杂度和识别能力。实验结果表明，ADC-CPANet的分类准确率分别高达96.43%和96.04%，优于其他先进的模型。

关键词：遥感图像，场景分类，卷积神经网络，Transformer，多分组卷积头分解注意力，ADC-CPANet模型

中图分类号：TP391.4/P2

引用格式：王威,李希杰,王新.2024.ADC-CPANet:一种局部—全局特征融合的遥感图像分类方法.遥感学报,28(10):2661-2672

Wang W, Li X J and Wang X. 2024. ADC-CPANet: A remote sensing image classification method based on local-global feature fusion. National Remote Sensing Bulletin, 28(10):2661-2672[DOI:10.11834/jrs.20232658]

1 引 言

随着卫星与无人机等遥感观测技术的飞速发展，高分辨率遥感图像的数据总量和数据类型得到极大丰富（徐从安 等，2021）。由于海量观测数据日益增长，因此，设计智能化信息提取和知识挖掘的方法已然成为遥感大数据应用的必然需求（徐科杰 等，2021）。高分辨率遥感图像识别技术对于城市规划、环境监测、土地利用等领域具有重要意义（刘康 等，2020；欧阳淑冰 等，2022；Zhu等，2022b），是目前计算机视觉领域的研究热点之一。

遥感图像场景分类是遥感图像解译的重要组成部分，旨在将每一幅遥感图像映射到预定的分类标签中（Li等，2020）。根据特征表示方法的不同，现有的遥感图像场景分类方法主要可分为两类：基于手工设计特征的方法和基于深度学习的方法（邓培芳 等，2021；Li等，2022）。基于手工设计特征的方法如尺度不变特征变换（Lowe，1999）、梯度直方图（Dalal和Triggs，2005）等，虽然这些方法在一些简单的场景分类任务中可以取得较好的分类效果，但是这类方法提取的特征可能存在不全面或者冗余等情况，因此在复杂场景中分类准确率仍然较低。

深度学习能够从文字、图像等海量数据中提取并分析特征，它可以解决许多复杂的模式识别难题。自2012年AlexNet（Krizhevsky等，2012）提出以来，卷积神经网络CNN（Convolutional Neural Networks）在图像分类领域取得了突破性的进展。随后，许多基于CNN的工作也在遥感图像场景分

收稿日期：2022-12-07；**预印本：**2023-03-17

基金项目：国防科技创新特区项目(编号:2019XXX00701)；湖南省重点研究开发计划(编号:2020SK2134)；湖南省自然科学基金(编号:2022JJ30625)

第一作者简介：王威,研究方向为计算机视觉、模式识别。E-mail：wangwei@csust.edu.cn

通信作者简介：王新,研究方向为计算机视觉、模式识别。E-mail：wangxin@csust.edu.cn

类任务中表现优异。余东行等（2020）通过综合迁移学习和集成学习，提出了联合卷积神经网络与集成学习的分类方法，该方法可有效提高当训练数据较少时或深层卷积神经网络难以训练时遥感图像场景分类的精度。Li等（2020）将挤压和激励SE（Squeeze-and-Excitation）注意力机制（Hu等，2018）加入到高分辨率网络模型HRNet（High-Resolution Network）（Sun等，2019）中，所提出的SE-HRNet可以更好地区分具有丰富特征的场景类别。Zhu等（2022a）提出全局联合注意机制模块来提取注意力区域，该模块能够学习更具辨别性的特征表达。徐从安等（2021）提出一种基于双重注意力机制的具有强鉴别性特征表示方法，该方法能够实现对重点区域和显著特征的关注，进而提高特征表示的鉴别性能力。虽然上述方法已经取得较好的分类结果，但仍有一定的局限性。CNN在提取局部特征方面表现很好，但难以捕获全局信息，而全局信息有助于理解遥感图像中的复杂内容。

近年来，随着Transformer（Vaswani等，2017）的深度学习方法被提出，其在视觉领域引起广泛关注。与CNN局部建模的特点不同的是，Transformer通过自注意力机制能够关联起图像中的每个像素，因此具有较强的全局建模能力。许多研究者尝试将Transformer架构的网络应用于遥感图像场景分类任务中，并获得不错的效果。Bashmal等（2021）提出一种基于视觉转换器（Dosovitskiy等，2021）的场景分类方法，并应用不同的数据增强策略训练模型。近些年的研究表明，CNN-Transformer的混合架构有利于结合两种架构的优势。Zhang等（2021）提出一种名为遥感转换器的遥感图像场景分类方法，该方法将自注意力融合到ResNet（He等，2016）中，增强了模型的特征提取能力。Li等（2022）提出一种用于遥感图像场景分类的双分支网络，该网络可以融合基于CNN的局部特征和基于Transformer的全局特征，从而提高分类精度。目前大部分混合架构采取的策略是在网络浅层阶段采用卷积块，在最后几个阶段采用Transformer块。然而这些混合架构在浅层阶段无法捕获全局信息，进而限制模型在遥感图像场景分类任务中的性能。

针对上述问题，本文设计了一种在网络各阶段均能提取全局特征和局部特征的遥感图像场景分类模型ADC-CPANet。本文的主要贡献：（1）提出局部特征提取模块ADC模块，该模块可以有效提取局部特征并加强模型对图像翻转和旋转的鲁棒性。（2）提出全局—局部特征提取模块CPA模块，该模块可以有效提取全局—局部特征并实现特征有效融合，同时在CPA模块中提出一种能够扩大感受野并加强特征提取能力的多分组卷积头分解注意力。（3）基于ADC模块和CPA模块提出了遥感图像场景分类模型ADC-CPANet。

2 模型结构设计

遥感图像场景分类具有类内差异性大、类间相似性高的特点和难点，因此遥感图像场景分类任务非常考验分类模型对多尺度特征的提取能力。针对上述问题，本文提出局部特征提取模块ADC模块和全局—局部特征提取模块CPA模块，并基于两种模块设计出了一种在模型各阶段均能提取全局特征和局部特征的遥感图像场景分类网络ADC-CPANet。

2.1 ADC模块

遥感图像具有丰富的纹理信息，因此在遥感图像场景分类任务中对遥感图像进行局部特征提取十分重要。基于此，本文提出局部特征提取模块ADC模块，该模块遵循MetaFormer（Yu等，2022）的一般架构。其中，MetaFormer提出将Transformer模块中令牌混合器（Token Mixer）替换为池化操作后仍可以取得较好的效果，但池化操作容易导致局部特征提取不够充分，为实现更有效地提取局部特征，本文设计了一种非对称深度卷积组作为高效的Token Mixer。非对称深度卷积组由3×3深度卷积、3×1深度卷积和1×3深度卷积并联组成，每个深度卷积之后均有GELU激活函数（Hendrycks和Gimpel，2023）和批标准化（Ioffe和Szegedy，2015）。非对称卷积组能够有效提取局部特征并提升模型对图像翻转和旋转的鲁棒性。研究（Ding等，2019）表明，将3×3卷积替换成非对称卷积组的并联结构可提高模型的表征能力。最后在MetaFormer的范式中使用非对称深度卷积组和多层感知机—双维注意力层构建ADC模块，具体如图1所示。ADC模块中的多层感知机—双维注意力层由多层感知机（扩展率为4）和双维注意力机制组成，多层感知机可以提取更基本和明显的特征，双维注意力机制可以有效加强前馈网络在空间和通道上的建模，从而提高性能。

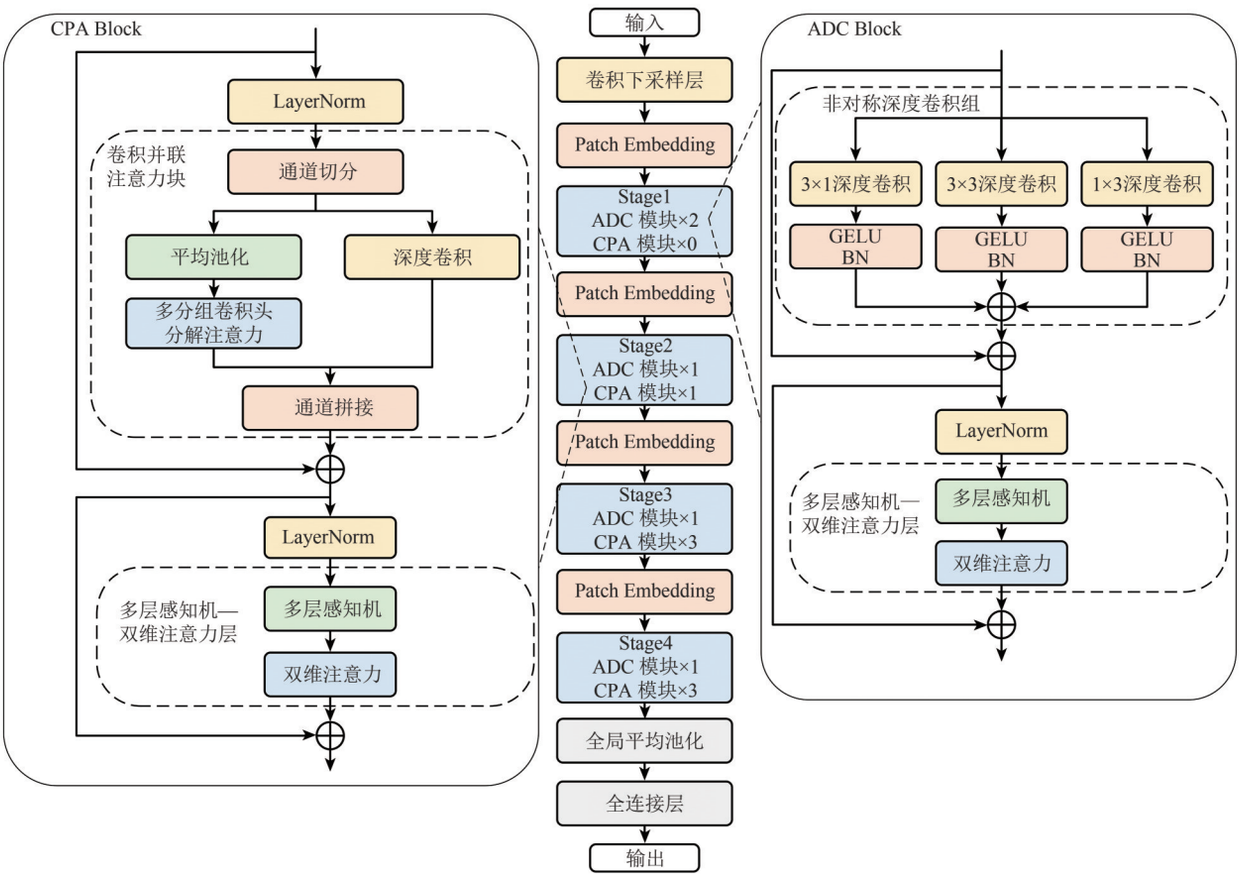


图1 ADC-CPANet结构

Fig. 1 ADC-CPANet structure

2.2 CPA模块

虽然ADC模块可以有效地提取局部特征，但是在遥感图像场景分类任务中实现全局信息的捕获仍然是迫切需要解决的问题。研究（Park和Kim，2022）表明，以自注意力为核心的Transformer模块会在一定程度上局限模型提取局部特征的能力，从而削弱模型的整体建模能力，因此将全局信息和局部信息以某种特定形式融合有助于模型提取到更本质、全面的特征。基于此，本文提出全局—局部特征提取模块CPA模块，该模块遵循MetaFormer的一般架构。为实现更有效地提取全局特征，同时保留局部特征，本文设计了一种卷积并联注意力块作为高效的Token Mixer。

卷积并联注意力块是一种高效的并行结构模块，该模块可以有效地提取全局—局部特征并实现特征融合。卷积并联注意力块中两条支路分别采用深度卷积和多分组卷积头分解注意力层。假设当前卷积并联注意力块的输入特征图大小为$H\times W\times C$，其中H、W分别代表特征图的高度和宽度，C代表特征图的通道数。卷积并联注意力块把输入的特征图在通道域上按1∶1比例进行切分，得到两组通道数相同、空间维度相同的特征图。通往卷积支路的特征图大小为$H\times W\times \frac{C}{2}$，该特征图会经过3×3深度卷积完成局部特征提取，得到卷积特征图。通往注意力支路的特征图大小为$H\times W\times \frac{C}{2}$，该特征图会依次经过平均池化和多分组卷积头分解注意力完成全局特征提取，得到注意力特征图。其中，卷积并联注意力模块的核心思想是融合局部特征和全局特征。因此，将上述两组经特征提取后的特征图在通道维度上进行拼接，得到大小为$H\times W\times C$的特征图作为当前模块的输出。最后在MetaFormer的范式中使用卷积并联注意力块和多层感知机—双维注意力层构建CPA模块。具体如图1所示。CPA模块同样引入多层感知机—双维注意力层来提高模型性能。

2.2.1 多分组卷积头分解注意力

Transformer中的计算开销主要来自于自注意力

层，为了改善这一问题同时加强全局特征提取能力，本文提出了具有线性复杂度的多分组卷积头分解注意力MGDA（Multi-Gconv Head Decomposition Attention）。

特征图$\boldsymbol{X}\in\boldsymbol{R}^{H\times W\times C}$输入MGDA后，首先生成查询矩阵$\boldsymbol{Q}$、键矩阵$\boldsymbol{K}$以及值矩阵$\boldsymbol{V}$，其表达式分别为：$\boldsymbol{Q}=\boldsymbol{W}_G^Q\boldsymbol{W}_P^{(\cdot)}\boldsymbol{X}$，$\boldsymbol{K}=\boldsymbol{W}_G^K\boldsymbol{W}_P^{(\cdot)}\boldsymbol{X}$，$\boldsymbol{V}=\boldsymbol{W}_G^V\boldsymbol{W}_P^{(\cdot)}\boldsymbol{X}$。其中，$\boldsymbol{W}_P^{(\cdot)}$表示的是1×1的点卷积，$\boldsymbol{W}_G^{(\cdot)}$是3×3空洞率为2的分组卷积（组数为$C$），该分组卷积旨在获得更大的感受野，从而提取到更多的图像特征。接着分别对$\boldsymbol{Q}$和$\boldsymbol{K}^{\mathrm{T}}$进行Softmax操作，得到Softmax（$Q$）和Softmax（$\boldsymbol{K}^{\mathrm{T}}$）。不同于大小为$HW\times HW$的较大的常规注意力图，本文将Softmax（$\boldsymbol{K}^{\mathrm{T}}$）和$\boldsymbol{V}$点积得到的注意力图大小仅为$C\times C$。然后将大小为$C\times C$的注意力图和一个可学习矩阵$\boldsymbol{Y}$进行点积，得到具有更强全局依赖关系的注意力图。最后该注意力图与Softmax（$\boldsymbol{Q}$）点积得到MGDA的输出结果。MGDA采用两次Softmax操作计算两次注意力图，实现线性化的同时获得更多的权重系数。注意力如下式（1）所示。

$$\text{Attention}=\left(\text{Softmax}\left(\boldsymbol{Q}\right)/\partial\right)\left(\left(\left(\text{Softmax}\left(\boldsymbol{K}^{\mathrm{T}}\right)/\partial\right)\boldsymbol{V}\right)\boldsymbol{Y}\right) \tag{1}$$

式中，矩阵$\boldsymbol{Q}\in\boldsymbol{R}^{HW\times C}$、$\boldsymbol{K}^{\mathrm{T}}\in\boldsymbol{R}^{C\times HW}$和$\boldsymbol{V}\in\boldsymbol{R}^{HW\times C}$是从原始大小$H\times W\times C$的张量进行变形获得。$\boldsymbol{Y}$为可学习的参数矩阵，$\partial$为缩放参数。类似于多头注意力机制，本文将通道数分成“头数”来并行学习单独的注意力特征图。具体如图2所示。

2.2.2 双维注意力机制

对于常规视觉转换器来说，前馈网络部分通常缺少特征空间层次上的建模，同时由于本模型采用较少的通道维度，加入注意力机制可以有效地提升模型性能。本文受Huang等（2022）提出的注意力启发，引入双维注意力机制。

双维注意力机制由两个分支组成：通道注意力分支和空间注意力分支。如图3所示，左边为通道注意力分支，用于捕捉全局信息；右边为空间注意力分支，用于捕捉局部信息。通道注意力分支首先经过全局平均池化，再通过两个全连接层。其中r代表缩减率，值为4。空间注意力分支在第二个全连接层将通道数调整为1之前，将左侧分支的全局信息输出与右侧分支的局部信息输出拼接在一起，由此融合全局信息和局部信息。最后将空间注意力分支、通道注意力分支融合后得到输出特征图。

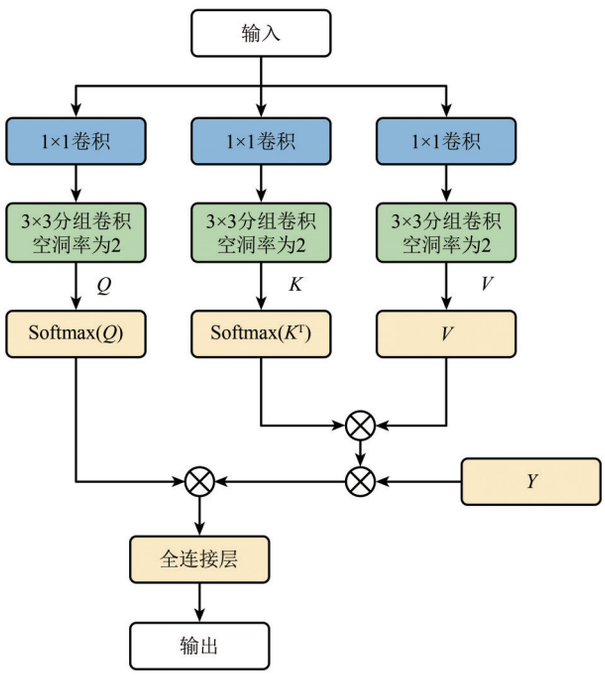


图2 多分组卷积头分解注意力

Fig. 2 Multi-Gconv head decomposition attention

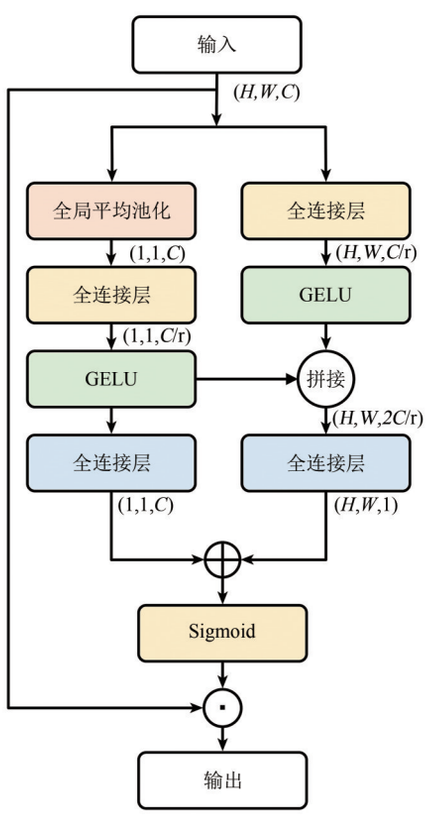


图3 双维注意力

Fig. 3 Two dimensional attention

2.3 ADC-CPANet

针对遥感图像纹理信息丰富和整体结构复杂的特点，基于ADC模块和CPA模块，本文设计了一种遥感图像场景分类模型ADC-CPANet。其结构如图1所示，参数如表1所示。

表1 ADC-CPANet详细参数设置

Table 1 Detailed parameter configuration of ADC-CPANet

	输出大小	层	
输入网络	$\frac{H}{2}\times\frac{W}{2}$	卷积下采样层	3×3卷积，C=64，S=2 3×3卷积，C=64，S=1 3×3卷积，C=64，S=1
第1阶段	$\frac{H}{4}\times\frac{W}{4}$	Patch Embedding	平均池化，S=2 1×1卷积，C=96
		模块	ADC模块×2
第2阶段	$\frac{H}{8}\times\frac{W}{8}$	Patch Embedding	平均池化，S=2 1×1卷积，C=128
		模块	［ADC模块×1］+ ［CPA模块×1］
第3阶段	$\frac{H}{16}\times\frac{W}{16}$	Patch Embedding	平均池化，S=2 1×1卷积，C=192
		模块	［ADC模块×1］+ ［CPA模块×3］
第4阶段	$\frac{H}{32}\times\frac{W}{32}$	Patch Embedding	平均池化，S=2 1×1卷积，C=192
		模块	［ADC模块×1］+ ［CPA模块×3］
分类器	1×1	模块	全局平均池化 全连接层，C=预测类别

预处理后的遥感图像首先经过卷积下采样层，该卷积下采样层由3个3×3的标准卷积组成，每个卷积之后均有批标准化（BN）和GELU激活函数；然后进入4个阶段，每个阶段先使用平局池化进行下采样和1×1卷积改变通道维度，再通过堆叠ADC模块和CPA模块进行多尺度的特征提取；最后利用分类器输出标签。

ADC-CPANet有4个阶段，考虑到在网络浅层通常需要更多的局部信息，网络的第1阶段只包含ADC模块。其余的每个阶段均包括1个ADC模块和若干个CPA模块。随着网络层数加深，特征图包含的信息更为抽象，此时网络的特征提取需要偏向于全局信息，因此CPA模块数量需呈递增趋势。保留ADC模块有助于模型提取更重要的图像级特征，同时帮助CPA模块捕获到更丰富的特征。综上所述，ADC-CPANet通过结合ADC模块的局部信息和CPA模块的全局信息可以学习到更全面的特征表示。

3 实验结果处理与分析

3.1 数据集

本实验数据采用来自武汉大学所发布的遥感图像数据集RSSCN7数据集（Zou等，2015）和SIRI-WHU数据集（Zhao等，2016）。两种数据集为遥感图像场景分类任务公开的常用数据集。

RSSCN7数据集包含2800幅遥感图像，这些图像来自于7个典型的场景类别：草地、森林、农田、停车场、住宅区、工业区和河湖，其中每个类别包含400张图像。该数据集中每张图像的像素大小为400×400，场景图像的多样性导致其具有较大的挑战性，这些图像成像于不同季节和天气，并以不同的比例进行采样。本实验将RSSCN7数据集的图像按照4∶1的比例划分为训练集和测试集，即训练集中每类有320张图像，共2240张；测试集中每类80张，共有560张。该数据集样本如图4所示。

图4 RSSCN7数据集

Fig. 4 RSSCN7 dataset

SIRI-WHU数据集包含2400幅遥感图像，这些图像来自于12个典型的场景类别：农场、商业区、港口、工业区、草地、立交桥、池塘、公园、闲置用地、居民区、河流、水体。其中每个类别包含200张图像，每一幅影像大小为200×200，该数据集资源来自谷歌地球，主要覆盖了中国的城市地区。本实验将数据集的图像按照4∶1的比例划分为训练集和测试集，即训练集中每类有160张图像，共1920张；测试集中每类40张，共有480张。该数据集样本如图5所示。

图5 SIRI-WHU数据集

Fig. 5 SIRI-WHU dataset

实验采用计算机图像分类任务中常用的评价标准，即采用准确率（Accuracy）、精确率（Precision）、召回率（Recall）、特异性（Specifity）、平衡分数（F1-score）作为模型的评价标准。设模型预测正确的正样本数量为T_P，预测错误的正样本数为F_N，预测正确的负样本数为T_N，预测错误的负样本数为F_P，则上述评价指标的计算公式如下：

$$\text{Accuracy} = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{2}$$

$$\text{Precision} = \frac{T_P}{T_P + F_P} \tag{3}$$

$$\text{Recall} = \frac{T_P}{T_P + F_N} \tag{4}$$

$$\text{Specifity} = \frac{T_N}{T_N + F_P} \tag{5}$$

$$\text{F1 - score} = \frac{2 \times T_P}{2 \times T_P + F_P + F_N} \tag{6}$$

3.2 实验设置

实验中，为了增强图像样本分布的多样性，对数据进行数据增强处理。首先通过随机裁剪将图片大小固定为224×224，然后进行随机水平旋转、归一化处理等操作。

在实验过程中，大部分模型的初始学习率设置为0.0005，为提高收敛质量，个别模型的初始学习率需要另作调整。例如，SwinTransformer（Liu等，2021）、CMT（Guo等，2022）、ViT（Dosovitskiy等，2021）的初始学习率另设为0.0002。为加快模型收敛速度，实验模型均采用AdamW优化器，权重衰减系数（weight decay）设置为0.05，损失函数采用交叉熵损失函数，每组实验迭代次数（epoch）设置为400，训练集和测试集的批尺寸（batch size）设置为16。所用操作系统版本号为Ubuntu 20.04.4 LTS，CPU型号为Intel(R)Xeon(R) Silver 4214 CPU@2.20 GHz，GPU型号为GeForce RTX 2080Ti，CUDA版本为11.1.74，网络搭建框架为Pytorch。所有模型的训练与测试在同一配置环境下进行。

3.3 消融实验

为了验证网络结构的有效性，本文分别从堆叠方式、模块内部结构、深层特征提取等角度设置多组消融实验，在RSSCN7数据集上的实验结果如下表2所示。其中，将ADC-CPANet中ADC模块替换为CPA模块得到CPANet；将ADC-CPANet中CPA模块替换为ADC模块得到ADCNet；将CPA模块和ADC模块堆叠顺序交换得到CPA-ADCNet。应注意，CPANet和CPA-ADCNet在网络第1阶段使用的模块均为CPA模块。对比其在RSSCN7数据集上的分类性能，单独堆叠形式的ADCNet、CPANet和交换堆叠顺序的CPA-ADCNet的准确率均低于ADC-CPANet。该实验证明了采用特定堆叠方式的ADC-CPANet既可以在模型各阶段提取到局部特征和全局特征，又可以提高模型的整体性能。ADC-CPANet-T将ADC-CPANet中第3和第4阶段的CPA模块数量设置为2，当弱化深层特征提取能力后，其准确率和各项评价指标均低于ADC-CPANet。NoDconvNet删除CPA模块中卷积并联注意力块中的深度卷积，此时CPA模块仅能提取全局特征，实验结果表明，网络性能有所下降。由此可见，CPA模块中提取的局部特征能够帮助模型提取到更本质、全面的特征。MHSANet则将CPA模块中的MGDA替换成多头自注意力，实验结果表明，注意力被替换后，网络分类效果不及ADC-CPANet。下面3种模型，准确率和各项评价指标均低于ADC-CPANet。NoTdAttenNet删除ADC模块和CPA模块中的双维注意力机制，导致前馈网络对空间和通道维度的建模减少；NoGconvNet删除CPA模块的MGDA中空洞率为2的分组卷积，导致感受野减小和特征提取能力被弱化；NoAcNet将ADC模块中的非对称深度卷积组替换成深度卷积，导致模型的局部特征提取能力下降，同时降低了模型对图像翻转和旋转的鲁棒性。

表2 消融方法评价指标对比

Table 2 Comparison of evaluation indicators of ablation methods

/%

模型	准确率	精确率	召回率	特异性	F1-score
ADCNet	96.07	96.13	96.09	99.37	96.09
CPANet	94.64	94.74	94.66	99.13	94.66
CPA-ADCNet	95.54	95.60	95.54	99.27	95.54
NoDconvNet	95.36	95.53	95.36	99.24	95.37
NoTdAttenNet	95.54	95.57	95.51	99.27	95.54
NoGconvNet	95.89	95.94	95.87	99.33	95.89
NoAcNet	95.54	95.54	95.53	99.29	95.53
MHSANet	95.18	95.23	95.17	99.21	95.19
ADC-CPANet-T	96.25	96.33	96.24	99.39	96.26
ADC-CPANet	**96.43**	**96.53**	**96.43**	**99.41**	**96.46**

注：粗体表示最优值。

3.4 对比实验

为了验证ADC-CPANet网络在解决遥感图像场景分类任务中的有效性，本文设置了3类对照组。第1类对照组包含CNN结构中经典的网络结构，如ResNet50（He等，2016），以及目前较为先进的EfficientNetV2（Tan和Le，2021）和ConvNeXt（Liu等，2022）。第2类对照组包含Transformer结构中分类表现较为突出的ViT（Dosovitskiy等，2021）、SwinTransformer（Liu等，2021）和PoolFormer（Yu等，2022）。第3类对照组包含近两年一些基于全局特征和局部特征相融合的混合模型BotNet（Srinivas等，2021）、CMT（Guo等，2022）、CoAtNet（Dai等，2021）和VAN（Guo等，2022）。

实验首先测试ADC-CPANet与一些深度学习模型在分类实验中的参数量与计算量，设输入图像尺寸为224×224，测试结果如表3所示。由表3可知，ADC-CPANet的复杂度远低于深层CNN模型，如ResNet50（He等，2016）、EfficientNetV2（Tan和Le，2021）和ConvNeXt（Liu等，2022）。与目前较先进的计算机视觉模型CoAtNet（Dai等，2021）、SwinTransformer（Liu等，2021）等相比也拥有较低的复杂度。与混合模型CMT（Guo等，2022）、VAN（Guo等，2022）相比，ADC-CPANet的参数量同样小于这些混合模型。由此可见，ADC-CPANet具有较低的复杂度，尤其在参数量方面占有较大的优势。

表3 各分类网络参数量与计算量对比

Table 3 Comparison of network parameters and computation for each classification

/M

模型	计算量	参数量
ResNet50(He等,2016)	4111.51	25.56
EfficientNetV2(Tan和Le,2021)	2874.31	21.46
ConvNeXt(Liu等,2022)	4457.48	27.81
ViT(Dosovitskiy等,2021)	16848.65	86.38
SwinTransformer(Liu等,2021)	4350.76	28.24
PoolFormer(Yu等,2022)	1818.83	11.89
BotNet(Srinivas等,2021)	3997.89	20.85
CMT(Guo等,2022)	1211.64	9.44
CoAtNet(Dai等,2021)	3320.71	17.75
VAN(Guo等,2022)	**879.57**	3.85
ADC-CPANet	2057.60	**3.73**

注：粗体表示最优值。

实验中，通过加权平均求和的方法求得各网络的评价结果如表4、表5所示。表4为RSSCN7数据集上对比实验结果，表5为SIRI-WHU数据集上对比实验结果。实验表明，ADC-CPANet在实验中各项评价指标取得了最好的效果，这也说明本网络分类器对数据集中的实例有着极强的识别能力。就表4而言，ADC-CPANet在RSSCN7数据集上准确率达到96.43%。与分类效果较好的ResNet50（He等，2016）和CoAtNet（Dai等，2021）相比，在参数量和计算量大幅减少的情况下，准确率分别提高了1.07%和0.89%，更加的轻量和高效。就表5而言，ADC-CPANet在SIRI-WHU数据集上准确率达到96.04%。与ResNet50（He等，2016）和CoAtNet（Dai等，2021）实验结果相比较，准确率仍提高了0.21%。ViT（Dosovitskiy等，2021）在RSSCN7数据集和SIRI-WHU数据集上分类准确率相对较低，因为ViT（Dosovitskiy等，2021）在大数据集上的分类效果较优，而在小数据集上的分类效果比较差。PoolFormer（Yu等，2022）的准确率和各项评价指标均低于ADC-CPANet，因为池化操作不能有效提取局部特征和全局特征，这也说明本文提出的两种模块对于提取多尺度特征更具有优势。同时在与基于全局特征和局部特征相融合的混合模型的比较中，ADC-CPANet也显著优于这些混合模型，这也说明本文所提出的ADC-CPANet模型拥有在各阶段提取全局特征和局部特征的能力。综上所述，本文所提出网络较其他图像分类网络具有更高的分类准确率和更优秀的性能。

3.5 可视化分析

ADC-CPANet在RSSCN7数据集中测试集上实验得到混淆矩阵如图6所示。由图6可见，分类准确率略低的只有“停车场”和“河湖”类别，因为“停车场”和“河湖”类别与其他类别有着众多相似性。在“河湖”类别测试样例中，部分样例中存在其他样例的信息，例如森林中存在湖水，工业区旁有河流等，导致该类别分类的准确的略低于其他类别。其次，“停车场”类别中可能包含“住宅区”和“工业区”类别的局部语义信息，容易造成预测错误。总体来说，本实验模型在整体上都有着很高的分类准确率，这说明ADC-CPANet网络模型能够有效捕获遥感图像的全局特征和局部特征，从而能够适应多尺度的RSSCN7数据集，并取得很好的分类效果。

表4 RSSCN7数据集上各分类网络实验结果对比

Table 4 Comparison of experimental results of classification networks on RSSCN7 dataset

/%

模型	准确率	精确率	召回率	特异性	F1-score
ResNet50(He等,2016)	95.36	95.40	95.34	99.24	95.34
EfficientNetV2(Tan和Le,2021)	94.29	94.31	94.27	99.07	94.29
ConvNeXt(Liu等,2022)	91.61	91.76	91.61	98.61	91.59
ViT(Dosovitskiy等,2021)	89.82	90.01	89.80	98.31	89.79
SwinTransformer(Liu等,2021)	93.75	93.80	93.76	98.96	93.74
PoolFormer(Yu等,2022)	92.86	92.93	92.87	98.81	92.86
BotNet(Srinivas等,2021)	94.11	94.41	94.10	99.04	94.14
CMT(Guo等,2022)	93.75	93.87	93.74	98.97	93.74
CoAtNet(Dai等,2021)	95.54	95.63	95.51	99.26	95.50
VAN(Guo等,2022)	94.11	94.23	94.11	99.03	94.10
ADC-CPANet	**96.43**	**96.53**	**96.43**	**99.41**	**96.46**

注:粗体表示最优值。

表5 SIRI-WHU数据集上各分类网络实验结果对比

Table 5 Comparison of experimental results of classification networks on SIRI-WHU dataset

/%

模型	准确率	精确率	召回率	特异性	F1-score
ResNet50(He等,2016)	95.83	95.97	95.83	**99.62**	95.83
EfficientNetV2(Tan和Le,2021)	95.21	95.29	95.21	99.55	95.19
ConvNeXt(Liu等,2022)	91.04	91.21	91.04	99.18	91.00
ViT(Dosovitskiy等,2021)	89.38	89.68	89.38	99.03	89.35
SwinTransformer(Liu等,2021)	94.79	95.00	94.79	99.52	94.79
PoolFormer(Yu等,2022)	93.75	93.85	93.75	99.43	93.75
BotNet(Srinivas等,2021)	90.00	90.19	90.00	99.09	90.03
CMT(Guo等,2022)	94.79	94.91	94.79	99.52	94.77
CoAtNet(Dai等,2021)	95.83	95.97	95.83	99.60	95.86
VAN(Guo等,2022)	95.00	95.07	95.00	99.54	94.99
ADC-CPANet	**96.04**	**96.02**	**96.04**	**99.62**	**96.02**

注:粗体表示最优值。

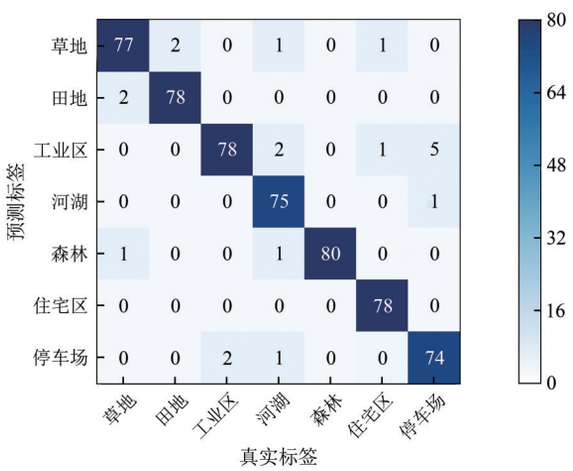


图6 ADC-CPANet方法在RSSCN7数据集上的混淆矩阵

Fig. 6 Confusion matrix of ADC-CPANet method on RSSCN7 dataset

ADC-CPANet在SIRI-WHU数据集中测试集上实验得到混淆矩阵如图7所示。由图7可见,在“农田”、“闲置土地”、“立交桥”和“公园”4种类别中,分类准确率均达到100%。“港口”、“居民区”、“草地”等类别对应的图像样本可能包含其他类别的语义信息,容易导致误判。尽管如此,ADC-CPANet也表现出十分优秀的分类性能,上述类别识别准确率均达到95%以上。此外,对于“池塘”、“河流”等场景类别之间高度相似的易混淆场景,ADC-CPANet也能取得较好的分类效果。由此可见,ADC-CPANet能够适应更具挑战性的SIRI-WHU数据集。

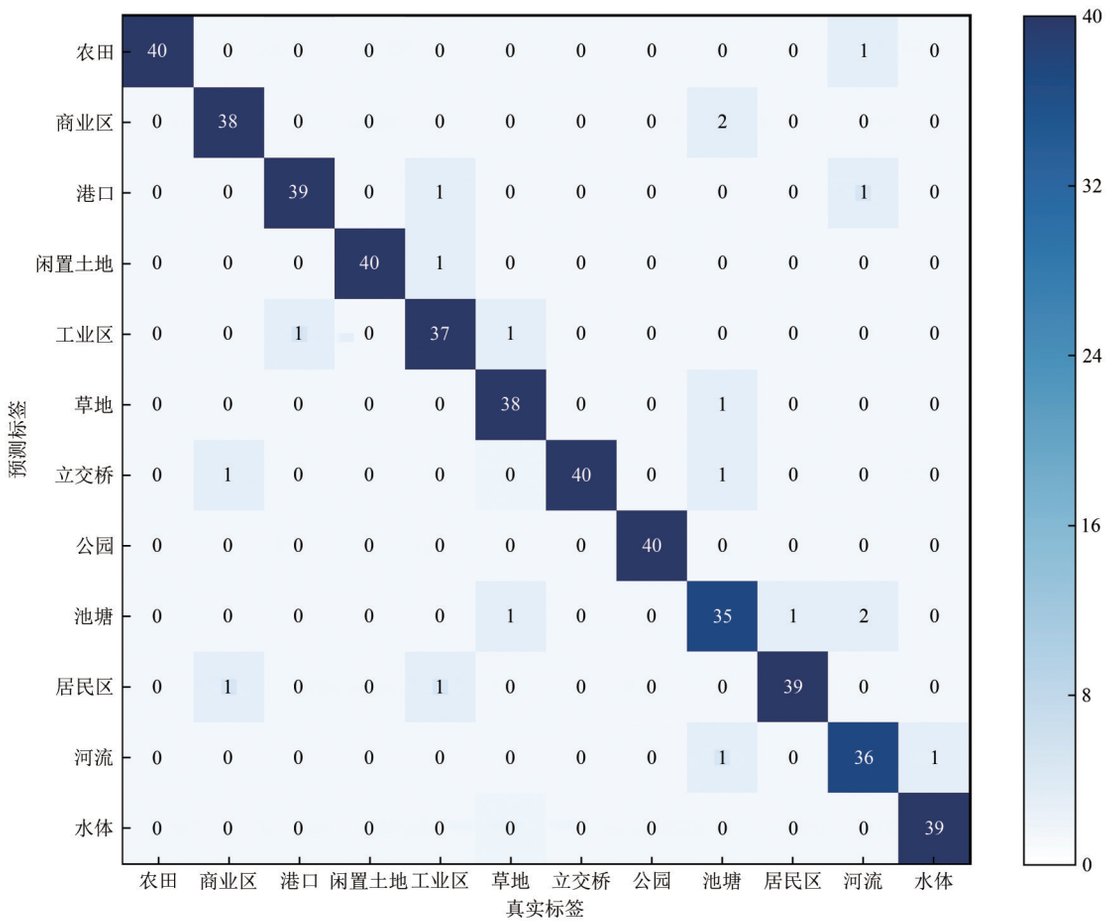
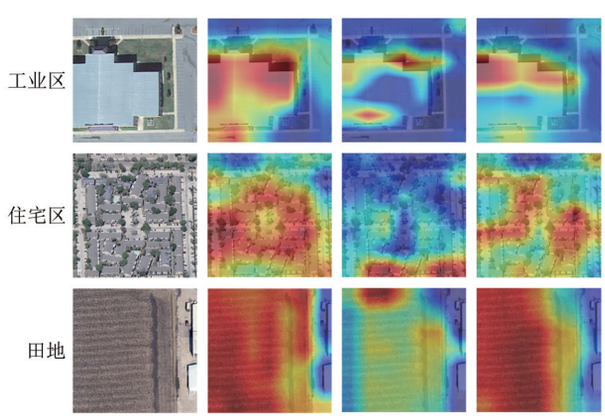


图7 ADC-CPANet方法在SIRI-WHU数据集上的混淆矩阵

Fig. 7 Confusion matrix of ADC-CPANet method on SIRI-WHU dataset

为更清楚地了解网络关注的区域，本文通过GradCAM（Selvaraju等，2017）绘制出热力图来进行可视化解释。图8展示了ADC-CPANet与对比实验中分类效果较好的ResNet50（He等，2016）以及CoAtNet（Dai等，2021）在RSSCN7数据集上的可视化分析。其中，图8（a）是不同类别的原始遥感图像。图8（b）是ADC-CPANet对于目标类别的热力图成像。图8（c）是ResNet50（He等，2016）对于目标类别的热力图成像。图8（c）是CoAtNet（Dai等，2021）对于目标类别的热力图成像。从图8可以看出图8（c）对于局部信息更加关注，这与CNN的擅长捕捉局部信息的特点有关。图8（d）的关注点既有局部信息也有全局信息，因为CoAtNet（Dai等，2021）是CNN和Transformer相结合的模型，但相较于ADC-CPANet而言，ADC-CPANet对于全局特征和局部特征的提取更加全面和准确，并对不同特征信息进行融合，因此实现的效果更加优异。



(a) 原始遥感影像 (a) Original remote sensing images

(b) ADC-CPANet的热力图像 (a) Heat maps of ADC-CPANet

(c) ResNet50的热力图成像 (c) Heat maps of ResNet50

(a) CoAtNet热力图成像 (a) Heat maps of CoAtNet

图8 热力图可视化分析

Fig. 8 Visual analysis of thermal diagram

4 结 论

基于CNN和Transformer结构对不同尺度特征

的提取能力，针对遥感图像的特点，本文提出了局部特征提取模块ADC模块以及可实现局部特征与全局特征提取的CPA模块，并在CPA模块中设计了一种高效的多分组卷积头分解注意力。本文基于上述两种模块提出了一个遥感图像场景分类模型ADC-CPANet。ADC-CPANet的有效性在RSSCN7数据集和SIRI-WHU数据集上得到了验证。实验结果表明，ADC-CPANet的分类准确率分别达到96.43%和96.04%。与实验中的其他分类模型相比，ADC-CPANet获得了更高的图像分类准确率和各项评价指标，且具有更低的计算复杂度。下一步工作将引入门控机制来更有效地聚合多尺度信息，进一步提高模型的表征能力，从而提高遥感图像场景分类的精度。

参考文献(References)

Bashmal L, Bazi Y and Rahhal M A. 2021. Deep vision transformers for remote sensing scene classification//Proceedings of 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. Brussels: IEEE: 2815-2818 [DOI: 10.1109/IGARSS47720.2021.9553684]

Dai Z H, Liu H X, Le Q V and Tan M X. 2021. CoAtNet: marrying convolution and attention for all data sizes//Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc.: 303

Dalal N and Triggs B. 2005. Histograms of oriented gradients for human detection//Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego: IEEE: 886-893 [DOI: 10.1109/CVPR.2005.177]

Deng P F, Xu K J and Huang H. 2021. CNN-GCN-based dual-stream network for scene classification of remote sensing images. National Remote Sensing Bulletin, 25(11): 2270-2282 (邓培芳, 徐科杰, 黄鸿. 2021. 基于CNN-GCN双流网络的高分辨率遥感影像场景分类. 遥感学报, 25(11): 2270-2282) [DOI: 10.11834/jrs.20210587]

Ding X H, Guo Y C, Ding G G and Han J G. 2019. ACNet: strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE: 1911-1920 [DOI: 10.1109/ICCV.2019.00200]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. 2021. An image is worth 16x16 words: transformers for image recognition at scale//9th International Conference on Learning Representations. [s.l.]: OpenReview.net

Guo J Y, Han K, Wu H, Tang Y H, Chen X H, Wang Y H and Xu C. 2022. CMT: convolutional neural networks meet vision transformers//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE: 12165-12175 [DOI: 10.1109/CVPR52688.2022.01186]

Guo M H, Lu C Z, Liu Z N, Cheng M M and Hu S M. 2023. Visual attention network. Computational Visual Media, 9(4): 733-752 [DOI: 10.1007/s41095-023-0364-2]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Hendrycks D and Gimpel K. 2023. Gaussian error linear units (GELUs)[EB/OL]. [2022-11-23]. https://arxiv.org/abs/1606.08415

Hu J, Shen L and Sun G. 2018. Squeeze-and-excitation networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE: 7132-7141 [DOI: 10.1109/CVPR.2018.00745]

Huang T, Huang L, You S, Wang F, Qian C and Xu C. 2022. LightViT: towards light-weight convolution-free vision transformers[EB/OL]. [2022-11-23]. https://arxiv.org/abs/2207.05557

Ioffe S and Szegedy C. 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift//Proceedings of the 32nd International Conference on Machine Learning. Lille: JMLR.org: 448-456

Krizhevsky A, Sutskever I and Hinton G E. 2012. ImageNet classification with deep convolutional neural networks//Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc.: 1097-1105

Li L L, Tian T, Li H and Wang L Z. 2020. SE-HRNet: a deep high-resolution network with attention for remote sensing scene classification//Proceedings of 2020 IEEE International Geoscience and Remote Sensing Symposium. Waikoloa: IEEE: 533-536 [DOI: 10.1109/IGARSS39084.2020.9324633]

Li M T, Ma J J, Tang X, Han X, Zhu C and Jiao L C. 2022. Resformer: bridging residual network and transformer for remote sensing scene classification//Proceedings of 2022 IEEE International Geoscience and Remote Sensing Symposium. Kuala Lumpur: IEEE: 3147-3150 [DOI: 10.1109/IGARSS46834.2022.9883041]

Liu K, Zhou Z, Li S Y, Liu Y F, Wan X, Liu Z W, Tan H and Zhang W F. 2020. Scene classification dataset using the Tiangong-1 hyperspectral remote sensing imagery and its applications. Journal of Remote Sensing (in Chinese), 24(9): 1077-1087 (刘康, 周壮, 李盛阳, 刘云飞, 万雪, 刘志文, 谭洪, 张万峰. 2020. 天宫一号高光谱遥感场景分类数据集及应用. 遥感学报, 24(9): 1077-1087) [DOI: 10.11834/jrs.20209323]

Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, Lin S and Guo B N. 2021. Swin transformer: hierarchical vision transformer using shifted windows//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE: 9992-10002 [DOI: 10.1109/ICCV48922.2021.00986]

Liu Z, Mao H Z, Wu C Y, Feichtenhofer C, Darrell T and Xie S N. 2022. A ConvNet for the 2020s//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE: 11966-11976 [DOI: 10.1109/CVPR52688.2022.01167]

Lowe D G. 1999. Object recognition from local scale-invariant fea-

tures//Proceedings of the 7th IEEE International Conference on Computer Vision. Kerkyra: IEEE: 1150-1157 [DOI: 10.1109/ICCV.1999.790410]

Ouyang S B, Chen W T, Li X J, Dong Y S and Wang L Z. 2022. Geomorphological scene classification dataset of high-resolution remote sensing imagery in vegetation-covered areas. National Remote Sensing Bulletin, 26(4): 606-619 (欧阳淑冰, 陈伟涛, 李显巨, 董玉森, 王力哲. 2022. 植被覆盖区高精度遥感地貌场景分类数据集. 遥感学报, 26(4): 606-619) [DOI: 10.11834/jrs.20221385]

Park N and Kim S. 2022. How do vision transformers work?//Proceedings of the 10th International Conference on Learning Representations. [s.l.]: OpenReview.net

Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D and Batra D. 2017. Grad-CAM: visual explanations from deep networks via gradient-based localization//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice: IEEE: 618-626 [DOI: 10.1109/ICCV.2017.74]

Srinivas A, Lin T Y, Parmar N, Shlens J, Abbeel P and Vaswani A. 2021. Bottleneck transformers for visual recognition//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE: 16514-16524 [DOI: 10.1109/CVPR46437.2021.01625]

Sun K, Xiao B, Liu D and Wang J D. 2019. Deep high-resolution representation learning for human pose estimation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE: 5686-5696 [DOI: 10.1109/CVPR.2019.00584]

Tan M X and Le Q V. 2021. EfficientNetV2: smaller models and faster training//Proceedings of the 38th International Conference on Machine Learning. [s.l.]: PMLR: 10096-10106

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc.: 6000-6010

Xu C A, Lü Y F, Zhang X H, Liu Y, Cui C H and Gu X Q. 2021. A discriminative feature representation method based on dual attention mechanism for remote sensing image scene classification. Journal of Electronics and Information Technology, 43(3): 683-691 (徐从安, 吕亚飞, 张筱晗, 刘瑜, 崔晨浩, 顾祥岐. 2021. 基于双重注意力机制的遥感图像场景分类特征表示方法. 电子与信息学报, 43(3): 683-691) [DOI: 10.11999/JEIT200568]

Xu K J, Deng P F and Huang H. 2021. HSRS-SC: a hyperspectral image dataset for remote sensing scene classification. Journal of Image and Graphics, 26(8): 1809-1822 (徐科杰, 邓培芳, 黄鸿. 2021. HSRS-SC: 面向遥感场景分类的高光谱图像数据集. 中国图象图形学报, 26(8): 1809-1822) [DOI: 10.11834/jig.200835]

Yu D H, Zhang B M, Zhao C, Guo H T and Lu J. 2020. Scene classification of remote sensing image using ensemble convolutional neural network. Journal of Remote Sensing (in Chinese), 24(6): 717-727 (余东行, 张保明, 赵传, 郭海涛, 卢俊. 2020. 联合卷积神经网络与集成学习的遥感影像场景分类. 遥感学报, 24(6): 717-727) [DOI: 10.11834/jrs.20208273]

Yu W H, Luo M, Zhou P, Si C Y, Zhou Y C, Wang X C, Feng J S and Yan S C. 2022. MetaFormer is actually what you need for vision//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE: 10809-10819 [DOI: 10.1109/CVPR52688.2022.01055]

Zhang J R, Zhao H W and Li J. 2021. TRS: transformers for remote sensing scene classification. Remote Sensing, 13(20): 4143 [DOI: 10.3390/rs13204143]

Zhao B, Zhong Y F, Xia G S and Zhang L P. 2016. Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing, 54(4): 2108-2123 [DOI: 10.1109/TGRS.2015.2496185]

Zhu Q Q, Deng W H, Zheng Z, Zhong Y F, Guan Q F, Lin W H, Zhang L P and Li D R. 2022a. A spectral-spatial-dependent global learning framework for insufficient and imbalanced hyperspectral image classification. IEEE Transactions on Cybernetics, 52(11): 11709-11723 [DOI: 10.1109/TCYB.2021.3070577]

Zhu Q Q, Lei Y, Sun X L, Guan Q F, Zhong Y F, Zhang L P and Li D R. 2022b. Knowledge-guided land pattern depiction for urban land use mapping: a case study of Chinese cities. Remote Sensing of Environment, 272: 112916 [DOI: 10.1016/j.rse.2022.112916]

Zou Q, Ni L H, Zhang T and Wang Q. 2015. Deep learning based feature selection for remote sensing scene classification. IEEE Geoscience and Remote Sensing Letters, 12(11): 2321-2325 [DOI: 10.1109/LGRS.2015.2475299]

ADC-CPANet：A remote sensing image classification method based on local-global feature fusion

WANG Wei，LI Xijie，WANG Xin

School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410114, China

Abstract：The rapid development of remote sensing technologies, such as satellites and unmanned aerial vehicles, has led to a surge in the amount and types of high-resolution remote sensing images. This advancement marks the onset of the “era of remote sensing big data.” Compared with low-resolution ones, high-resolution remote sensing images provide richer texture, detailed information, and a more

complex structure, making them crucial for applications like urban planning. However, images within the same category can vary substantially, whereas images from different categories may appear similar. Therefore, multi-scale feature extraction is important for remote sensing image scene classification. Current methods for remote sensing image scene classification can be divided into two categories according to the feature representation: those based on manual design features and those based on deep learning. Those based manual design features cover scale-invariant feature transformation and gradient scale histogram. They can achieve good results for simple classification tasks, but the feature information they extract may be incomplete or redundant, so the accuracy of classification in complex scenes remains low. By contrast, the methods based on deep learning have made incredible progress in scene classification owing to their powerful feature extraction ability. Compared with traditional methods, Convolution Neural Networks (CNNs) are commonly used in visual tasks, particularly those that involve more complex connections and diverse convolution forms. CNNs are effective at extracting local features, but they struggle with capturing long-distance dependencies among features. The Transformer architecture, which has recently been applied to computer vision, addresses this limitation through its self-attention layer that enables global feature extraction. Recent studies show that hybrid architectures combining CNNs and Transformers can utilize their advantages. This study proposes an Aggregation Depth-wise Convolution (ADC) module and a Convolution Parallel Attention (CPA) module. The ADC module effectively extracts local feature information and enhances the robustness of the model to image flipping and rotation. The CPA module integrates global and local feature extraction, with a multi-group convolution head decomposition designed to expand the receptive field and enhance feature extraction capacity. A remote sensing image scene classification model called ADC-CPANet is designed on the basis of two modules. The ADC and CPA modules are stacked at each stage of the model, improving its ability to extract global and local features. The effectiveness of ADC-CPANet is validated using the RSSCN7 and Google Image datasets. Experimental results demonstrate that ADC-CPANet achieves classification accuracies of 96.43% on the RSSCN7 dataset and 96.04% on the Google Image dataset, outperforming other advanced models. ADC-CPANet excels in extracting global and local features, achieving competitive scene classification accuracy.

Key words: remote sensing image, scene classification, convolutional neural network, Transformer, Multi-Gconv Head Decomposition Attention, ADC-CPANet model

Supported by National Defense Science and Technology Innovation Special Zone Project of China (No. 2019XXX00701); Key Research and Development Projects of Hunan Province, China (No. 2020SK2134); Natural Science Foundation of Hunan Province, China (No. 2022JJ30625)