引用格式：YU Tianwei, ZHENG Enrang, SHEN Junge, et al. Optical Remote Sensing Image Scene Classification Based on Multi-level Cross-layer Bilinear Fusion[J]. Acta Photonica Sinica, 2022, 51(2):0210007
余甜微,郑恩让,沈钧戈,等.基于多级别跨层双线性融合的光学遥感图像场景分类[J].光子学报,2022,51(2):0210007

基于多级别跨层双线性融合的光学遥感图像场景分类

余甜微[1],郑恩让[1],沈钧戈[2],王凯[3]
(1 陕西科技大学 电气与控制工程学院,西安 710021)
(2 西北工业大学 无人系统技术研究院,西安 710072)
(3 河南省水下智能装备重点实验室,郑州 450000)

摘　要:针对光学遥感场景图像存在由空间模式复杂、类间相似度大和同类多样性高导致的模型分类准确度受限的问题,提出一种基于多级别跨层双线性融合的光学遥感场景分类算法。首先从ResNet50模型中提取多层次特征信息,将膨胀卷积的扩张率设置为不同数值来提取多个空间尺度下的上下文特征,通过串行融合多尺度特征丰富特征信息的场景语义。为了充分利用低层、高层、全局上下文特征信息的互补优势,提出多级别注意力特征融合模块,有效增强模型的特征提取能力。最后采用跨层双线性融合方法对多级别特征进行分层融合,融合后的特征用于分类。通过在三个公开的遥感数据集UCM、AID和PatternNet上进行广泛试验,验证了所提方法的可行性,与其它先进的场景分类方法相比,该方法实现了更加优异的分类性能。

关键词:遥感;场景分类;膨胀卷积;多级别注意力;跨层双线性融合
中图分类号:TP751　　**文献标识码**:A　　**doi**:10.3788/gzxb20225102.0210007

0　引言

遥感图像场景分类是遥感图像分析与解译工作的关键组成部分,不同的语义场景根据每个地面区域的功能所确定,遥感图像场景分类旨在根据图像内容为遥感图像分配预定义的场景类别标签,如港口、机场、森林或居民区,在地质灾害监测、城市发展规划、军事目标探测、农业资源调查等方面具有相当广泛的应用[1]。随着卫星传感器与无人机等技术的不断提升,高分辨率遥感图像的获取成为了可能,遥感图像的空间纹理特征和场景语义信息得到了丰富,且存在同类别差异性大、不同类别相似性高的难题,因此更加有效的特征表示对提升遥感图像场景分类精度起着决定性作用。

目前应用于遥感图像场景分类的特征表示包括低层特征、中层特征、深度特征三种类型。早期传统的遥感图像场景分类方法大多通过手动提取遥感图像的低中层特征,低层特征侧重于设计图像的纹理、颜色和空间信息等局部或全局的浅层特征[2-4],手工特征在纹理整齐、空间分布均匀的遥感图像上表现良好,但难以刻画出复杂遥感场景的语义信息。对手工特征建模可获得中层特征,其中视觉词袋模型[5]的特征表示方法运用最为普遍。同手工特征相比,中层特征建立了手工特征与图像语义特征间的联系,但在实际应用中的性能本质上仍依赖于手工特征,缺乏对不同场景的灵活性。卷积神经网络(Convolutional Neural Networks,CNN)作为最成功的基于深度学习的网络,采用端对端的特征学习方法,近几年在场景分类领域表现出了更加优异的性能。大多数CNN方法基于迁移学习的思想,将训练完成的CNN模型进行微调或直接用于提取遥感图像的特征。例如,文献[6]提出利用预训练的CNN分别提取最后一个卷积层和全连接层

基金项目:国家自然科学基金(No. 61603233)
第一作者:余甜微(1997—),女,硕士研究生,主要研究方向为深度学习、计算机视觉。Email: twei_1yu024@163.com
导师(通讯作者):郑恩让(1962—),男,教授,博士,主要研究方向为智能信息处理。Email: zhenger@sust.edu.cn
收稿日期:2021-06-25;**录用日期**:2021-09-07

特征，得到图像的全局场景表示。文献[7]对六个不同的CNN模型进行微调后直接提取图像的全局特征用于场景分类。

尽管上述方法能够提高分类性能，但CNN在提取图像特征过程中，不同层次的信息被映射成为不同的特征表示，低层网络获取到蕴含空间细节信息的图像局部特征，局部特征通过抽象组合成为蕴含图像全局信息的高层语义特征。而仅仅利用CNN的中间层或全连接层特征作为图像的特征表示会忽略不同层级信息的互补优势，导致模型的表征能力不足。为了增强特征的表达能力，特征融合策略被引入到CNN中，以进一步提升模型的分类精度。文献[8]提出利用预训练的CNN模型获得多层卷积特征，采用级联或元素相加的方法得到多层特征的融合表示。文献[9]提出利用空间相似性策略重排由预训练的VGG19-Net[10]得到的局部特征，并将全局特征和局部特征相结合来增强图像表达效果。根据遥感图像的特点，也可以对现有的CNN结构进行改进，以提高特征的鲁棒性。文献[11]在CNN中集成了多扩张池化模块、反向残差模块以及通道注意力机制，提出一个轻量级的场景分类模型。然而，CNN模型采用单一尺寸的卷积核提取图像特征，感受野大小被固定在一定范围内，只能提取到单一尺度的图像局部细节信息，无法获取到较为丰富的不同尺度下的场景信息。此外，直接级联或元素相加的融合方法是基于提取图像的一阶特征，这类融合方法的特征表达能力有限，没有考虑到不同层次特征的相互作用，无法较好地学习到遥感图像上不同语义元素间的相关性。

为了更好地解决以上问题，本文提出基于多级别跨层双线性融合的光学遥感场景分类方法：1)针对CNN感受野大小有限导致提取出的特征无法包含丰富的场景语义信息，提出多尺度膨胀卷积模块(Multiscale Dilated Convolution Module，MDC)，通过构建多分支具有不同扩张率的膨胀卷积结构来捕获多尺度下的场景特征信息；2)针对以往基于CNN模型的场景分类方法忽略了不同层次特征间具有互补性的问题，提出多级别注意力特征融合模块(Multi-level Attention Feature Fusion Module，MAFF)，采用空间注意力抑制低层特征背景区域，根据低层特征、高层特征和全局语义特征的不同特性有效融合多种不同层次的信息，增强模型的特征提取能力；3)针对级联或元素相加的融合方式特征表达能力的不足，提出跨层双线性融合(Cross-layer Bilinear Fusion，CBF)，将来自不同尺度的特征通过哈达玛积进行逐元素汇合，捕获到同一网络不同层级间成对的二阶特征关系，提升模型的特征表达能力。

1 多级别跨层双线性融合模型

本文提出的多级别跨层双线性融合模型主要由MDC module、MAFF module、跨层双线性融合三个部分组成，具体流程如图1所示。该模型以ResNet50作为基本网络结构，其核心思想是首先使用预训练ResNet50网络对输入的场景图像提取到不同层次、不同分辨率的深度特征，然后利用MDC module对深度

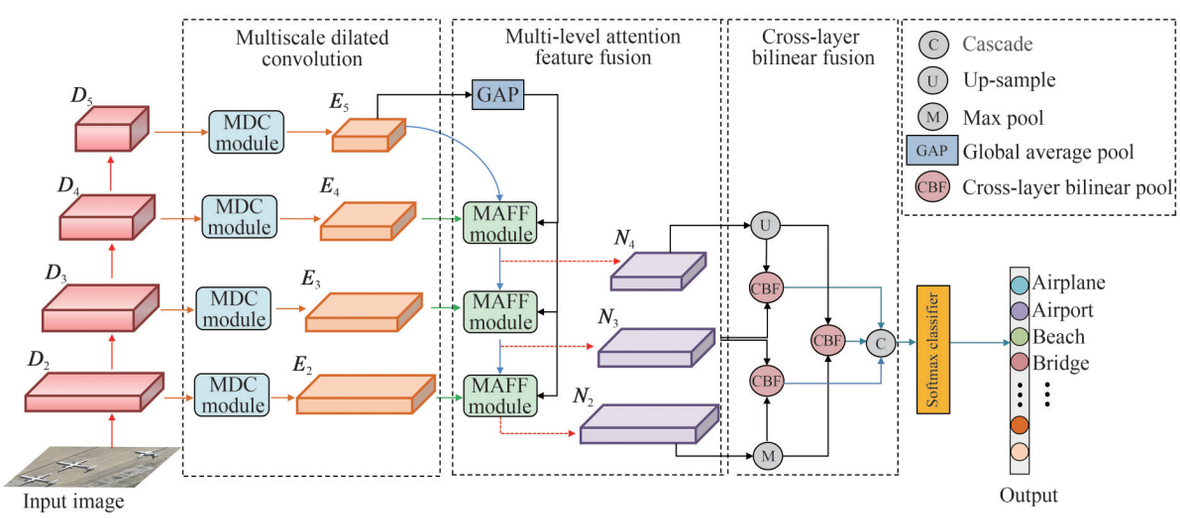

图1 本文算法整体流程

Fig.1 Overall flowchart of the proposed method

特征的上下文场景信息进行多尺度语义增强，再通过MAFF module逐步获得了将低、高层以及全局语义特征进行优势互补的多级别特征，最后通过跨层双线性融合将多级别特征进行集成，将最终融合得到的特征表示馈送到softmax分类器实现预测分类，输出场景图像的预测标签。

1.1 多尺度膨胀卷积模块

ResNet50模型由多个带有快捷连接的残差学习块堆叠组成，避免了网络过拟合的情况且加深了网络深度，具有更强大的特征提取能力。通过ResNet50模型分别提取到conv2_x、conv3_x、conv4_x、conv5_x层的特征，作为具有56×56、28×28、14×14、7×7不同分辨率的来自4个层次的深度特征$\{D_2, D_3, D_4, D_5\}$。

膨胀卷积与普通卷积相比，通过对扩张率的设置控制在卷积核中填充0的个数且不增加额外学习参数。对于尺寸为$k \times k$的卷积核，扩张率为r的膨胀卷积在卷积核的任意相邻像素值之间引入了$r-1$个0值，有效地将卷积核大小扩大到$k_d = k + (k-1)(r-1)$。图2所示为三个具有不同扩张率的3×3膨胀卷积核。融合不同扩张率的膨胀卷积所感知到的多个空间尺度下的场景信息，可提升模型对不同尺度特征的获取能力，有助于场景类别的推断。但是随着r增大，使用膨胀卷积存在棋盘效应问题，即膨胀卷积计算覆盖的实际区域类似于国际象棋棋盘状，输入图像经过连续计算后的数据间依赖性降低，容易忽略掉一些位置的信息，不利于图像关键特征信息的提取。为了避免棋盘效应，基于混合膨胀策略[12]提出了MDC module。混合膨胀策略的关键是一组内进行堆叠的膨胀卷积的扩张率不能有大于1的公约数，这样下一次的卷积运算都确保弥补了经过上次卷积运算后像素间存在的空洞，最终可获得一个没有孔洞的完整感受野。因此分别使用扩张率为1、2、3的膨胀卷积，将一组r由低到高的膨胀卷积进行叠加，既扩展了卷积运算的感受野，又有效弥补了棋盘效应带来的弊端。

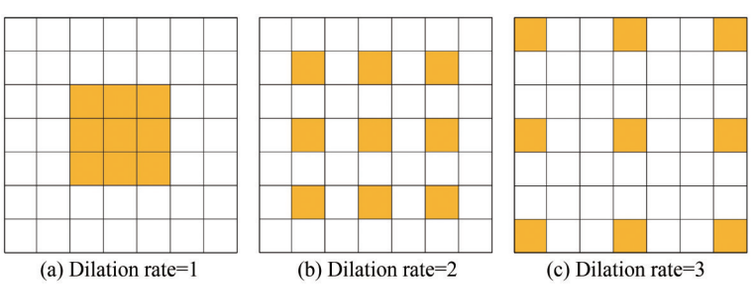

图2 不同扩张率的膨胀卷积

Fig.2 Dilated convolution with different dilation rate

MDC module的结构如图3所示。首先，输入为经过ResNet50提取到的不同层次的特征图，分别采用卷积核尺寸均为3×3，通道数为256，扩张率设置为1、2、3的膨胀卷积对其进行卷积操作，令padding和扩张率

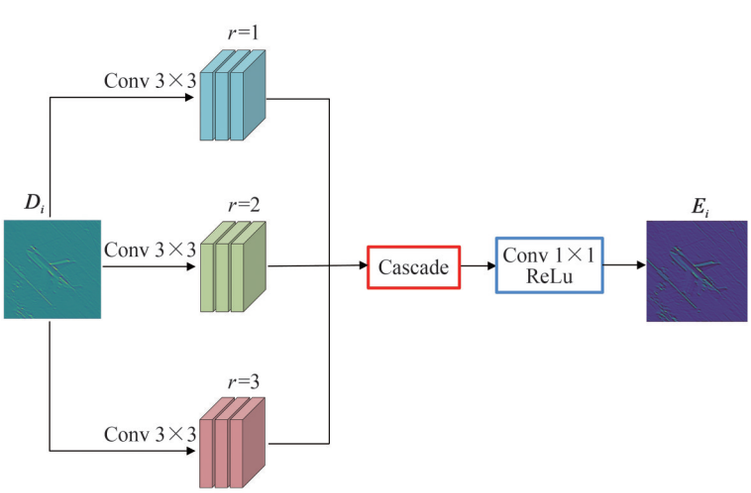

图3 多尺度膨胀卷积模块

Fig.3 Multiscale dilated convolution module

相等保持卷积前后特征图的尺寸一致，获取到不同尺度空间的特征表示，将不同尺度的特征信息进行跨通道级联后采用1×1的深度卷积对通道数进行聚合降维，将通道维数还原为256，加强不同通道间的信息交流，最终执行线性整流函数（Rectified Linear Unit，ReLu）得到特征图$\boldsymbol{E}_i$，具体表示为

$$\boldsymbol{E}_i=\sigma\left\{\mathrm{Cov}\left\{\mathrm{Cat}\left(D\mathrm{Cov}\left(\boldsymbol{D}_i,1\right),D\mathrm{Cov}\left(\boldsymbol{D}_i,2\right),D\mathrm{Cov}\left(\boldsymbol{D}_i,3\right)\right)\right\}\right\}\quad i\in\{2,3,4,5\}\tag{1}$$

式中，$\boldsymbol{D}_i$表示利用ResNet50模型提取到的多层次特征图；$\sigma(\cdot)$表示ReLu激活映射，$\mathrm{Cov}(\cdot)$表示1×1卷积层；$D\mathrm{Cov}\left(\boldsymbol{D}_i,r\right)$表示利用扩张率为$r$的膨胀卷积对特征图$\boldsymbol{D}_i$采样；$\mathrm{Cat}(\cdot)$为通道拼接操作；$\boldsymbol{E}_i$表示第$i$个层次对应的多尺度语义增强的特征图。

1.2 多级别注意力特征融合模块

图像的全局上下文信息能够拥有全局感受野，综合考虑全局信息可以有效推断出场景类别，淡化背景细节干扰。由于遥感场景图像具有地物信息复杂的特点，在注意力机制的基础上提出多级别注意力特征融合模块，设计空间注意力机制学习图像上不同区域的重要性，滤除与遥感图像无关的背景干扰信息，并在低层局部特征和高层语义特征融合时引入全局上下文信息，实现各个层级的特征信息互补，提升网络场景分类的正确率。

图4为提出的多级别注意力特征融合模块。首先，为了更加关注场景图像的关键位置信息，提出空间注意力机制用于增强图像的边缘信息和目标区域。如图5所示，在注意力机制中，对于根据式（2）得到的低层多尺度增强特征图$\boldsymbol{E}_l$，分别在通道维度上采用全局平均池化和全局最大池化，将二者的结果逐元素求和来激活图像的关键目标区域。为了获取更加丰富的图像边缘信息，将尺寸为3×3的卷积核替换为两个尺寸分别为3×1和1×3的小卷积核，从不同方向上提取特征信息的同时减小计算量。将两支路的特征利用对应元素相乘的方法进行汇合，使用由Sigmoid激活函数归一化的特征映射应用于$\boldsymbol{E}_l$得到空间注意力的输出$\boldsymbol{F}_l$。r表示通道降维系数，实验中设置r为8。具体过程可描述为

$$\begin{cases}\boldsymbol{F}_l=\boldsymbol{E}_l\times\left(\sigma\left(\boldsymbol{F}_{l_1}\times\boldsymbol{F}_{l_2}\right)\right)\\ \boldsymbol{F}_{l_1}=\mathrm{Conv}_{1\times3}\left(\theta\left(\mathrm{Conv}_{3\times1}\left(\boldsymbol{E}_l\right)\right)\right)\\ \boldsymbol{F}_{l_2}=\theta\left(\mathrm{GAP}\left(\boldsymbol{E}_l\right)+\mathrm{GMP}\left(\boldsymbol{E}_l\right)\right)\end{cases}\tag{2}$$

式中，$\boldsymbol{F}_{l_1}$和$\boldsymbol{F}_{l_2}$分别表示图5中上下两条支路得到的特征，$\sigma(\cdot)$表示Sigmoid激活函数，$\theta(\cdot)$表示批归一化和ReLu非线性变换，$\mathrm{GAP}(\cdot)$和$\mathrm{GMP}(\cdot)$表示全局平均池化和全局最大池化。

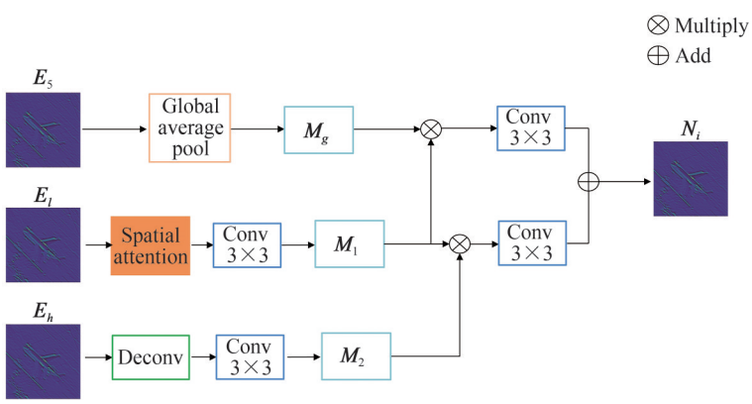

图4 多级别注意力特征融合模块

Fig.4 Multi-level attention feature fusion module

将最高层次的多尺度增强特征$\boldsymbol{E}_5$经过全局平均池化获取到场景图像的全局上下文信息$\boldsymbol{M}_g$；采用3×3的卷积核对$\boldsymbol{F}_l$进行特征提取得到特征图$\boldsymbol{M}_1$，与全局上下文信息元素$\boldsymbol{M}_g$相乘进行特征融合，再经过3×3的卷积计算进一步增强融合特征的泛化能力，补充低维特征高层语义信息的缺失，同时，抑制低维特征的背景噪声。类似地，对于高层多尺度增强特征$\boldsymbol{E}_h$首先进行反卷积处理，将$\boldsymbol{E}_h$与低层多尺度增强特征图$\boldsymbol{E}_l$转换成相

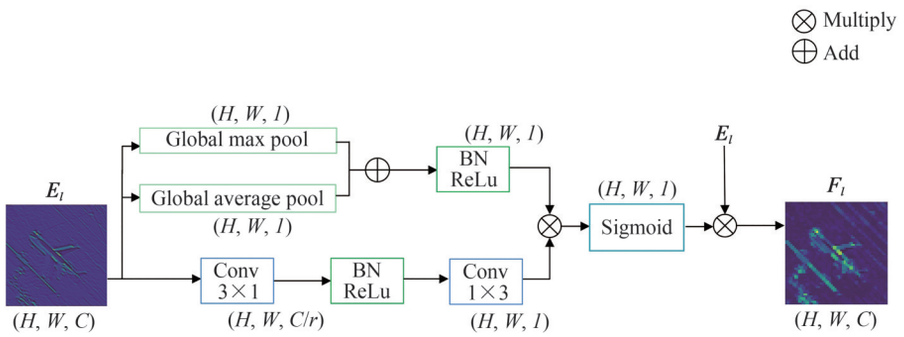

图5　空间注意力模块

Fig.5　Spatial attention module

互匹配的尺寸。特征图$\boldsymbol{M}_2$由卷积运算得到，与特征图$\boldsymbol{M}_1$进行逐元素相乘后再采用3×3卷积层得到融合特征图的输出；以此实现低层局部特征和高层语义特征的信息互补，结合低层特征的空间细节信息和高层特征的语义信息获得更有利于场景分类的融合特征。将这些融合的特征信息逐元素累加得到最终的特征图，以此作为输入进入下一阶段的融合过程。由此经过阶段级特征融合递进的方式逐步生成更具表征能力的特征表示$\boldsymbol{N}_i$。上述过程可表示为

$$\boldsymbol{N}_i = C\left(C(\boldsymbol{F}_1)\times \mathrm{GAP}(\boldsymbol{E}_5)\right) + C\left(C(\boldsymbol{F}_1)\times C\left(D(\boldsymbol{E}_{\mathrm{h}})\right)\right) \qquad i\in(2,3,4) \tag{3}$$

式中，$C(\cdot)$表示包括卷积、批归一化以及ReLu激活操作，$D(\cdot)$表示反卷积操作。

1.3　跨层双线性融合

双线性池化模型通过特征的外积或点积可获得图像的二阶信息，比一阶特征更具鲁棒性和区分性，在细粒度的视觉分类中获得了优异的性能[13]。细粒度视觉分类任务是指将同一基本类别的图像划分为更详细的子类别，例如鸟的品种分类，然而不同类别的图像间高度相似，同一类别的图像间差异很大，这同样符合遥感图像场景分类任务的特点。受到细粒度视觉分类的启发，引入双线性池化模型，提出一种跨层双线性融合的方法来有效融合由上述MAFF module生成的不同层次的特征$\{\boldsymbol{N}_2,\boldsymbol{N}_3,\boldsymbol{N}_4\}$，通过哈达玛积运算对任意两个不同层级特征提取二阶双线性信息，利用这种跨层建模捕获到成对的局部特征间的关联性，能够实现分层的特征交互和高效信息集成，充分聚合了不同层次特征图中包含的图像深层语义和浅层纹理信息，为场景分类任务提供了显著优势。最重要的是，与传统双线性池化方法相比，哈达玛积运算为两个矩阵的对应元素乘积，并不改变矩阵维数，可解决外积运算引起的维数爆炸问题，并且大大减少了计算参数。

首先对$\boldsymbol{N}_4$采用双线性插值生成更高分辨率的特征图，对$\boldsymbol{N}_2$使用最大池化，将$\{\boldsymbol{N}_2,\boldsymbol{N}_3,\boldsymbol{N}_4\}$三个具有不同分辨率的卷积特征匹配为同一空间维度。在进行跨层双线性融合前，通过1×1卷积层将不同层次的特征在不损失特征图分辨率的前提下映射到高维空间，卷积计算后使用ReLu函数大幅增强特征非线性，从而提升特征的表达能力。对于任意两个来自不同层次的卷积特征，在每个空间位置上通过哈达玛积运算得到维度不变的双线性特征，在其空间维度上采取求和池化(Sum pooling)操作，用于保存综合信息，最后执行符号平方根变换和L2归一化操作以获得融合特征$\boldsymbol{z}$。双线性池化的过程表示为

$$\boldsymbol{b}_{AB} = \boldsymbol{N}_A \circ \boldsymbol{N}_B \tag{4}$$

$$\boldsymbol{\xi}_{AB} = \sum_{j=1}^{C} \boldsymbol{b}_{AB}(1{:}H\times W, j) \tag{5}$$

$$\boldsymbol{y} = \mathrm{sign}(\boldsymbol{\xi}_{AB})\sqrt{\left|\boldsymbol{\xi}_{AB}\right|} \tag{6}$$

$$\boldsymbol{z} = \boldsymbol{y}/\left\|\boldsymbol{y}\right\|_2 \tag{7}$$

式中，$\boldsymbol{b}_{AB}\in\boldsymbol{R}^{H\times W\times C}$表示对成对层级特征经过哈达玛积运算后得到的双线性特征，$\boldsymbol{N}_A\in\boldsymbol{R}^{H\times W\times C}$和$\boldsymbol{N}_B\in\boldsymbol{R}^{H\times W\times C}$表示分辨率进行统一后的两个不同卷积特征，特征图的尺寸为$H\times W\times C$，$\circ$为哈达玛积，$\boldsymbol{\xi}_{AB}\in\boldsymbol{R}^{1\times C}$表示对双线性特征$\boldsymbol{b}_{AB}$各个通道上所有位置的元素求和，$\boldsymbol{y}$表示对$\boldsymbol{\xi}_{AB}$进行符号平方根运算的结果。

将不同层级的特征$\{\boldsymbol{N}_2,\boldsymbol{N}_3,\boldsymbol{N}_4\}$两两之间交互得到其对应的融合特征$\boldsymbol{z}$，对提取到的三组融合特征进行

级联聚合，得到最终的图像特征表示z'定义为

$$z'=z_{23}\cup z_{34}\cup z_{24} \tag{8}$$

式中，z_{ij}表示N_i和N_j两个不同层级的特征通过双线性池化后的融合双线性特征，∪表示沿通道维度的拼接操作。最后把特征表示z'输入到softmax分类器，用于预测输入图像的类别标签。

2 试验结果分析

2.1 试验数据集

为了验证本文方法的可靠性，选取3个公开可用的数据集UC Merced Land-Use Data Set (UCM)[5]、Aerial Image Data Set(AID)[14]和PatternNet[15]进行试验。如表1所示，对3个数据集的基本信息进行了对比。UCM数据集包含2 100张256×256像素、0.3 m空间分辨率的RGB图像，涵盖21种土地利用类别，每个类别包含100张场景图像。AID数据集总共具有10 000张像素大小为600×600的航空场景图像，分辨率约为1～8 m，该数据集包含30个场景类别，每个类别的图像数量220～420张不等。PatternNet数据集被划分为38类，每类场景含有800张图像，共30 400张图像，大小为256×256像素，分辨率约为0.062～0.493 m。

表1 场景数据集信息

Table 1 Scene dataset information

Datasets	Images per class	Total images	Images size	Scene classes	Spatial resolution/m
UCM	100	2 100	256×256	21	0.3
AID	220～420	10 000	600×600	30	1～8
PatternNet	800	30 400	256×256	38	0.062～0.493

2.2 评价指标

采用总体分类准确率(Overall Accuracy，OA)和混淆矩阵(Confusion Matrix，CM)来评估提出算法的性能。

1) OA定义为正确分类的样本数占测试集总样本数的比例，它反映了数据集总体的分类情况，计算公式为

$$\mathrm{OA}=\frac{S}{N}\times 100\% \tag{9}$$

式中，S是测试集中正确分类的样本数，N是测试集的总样本数。

2) CM通过矩阵形式的表达，更为直观地呈现出类别间的错误情况。混淆矩阵的对角线元素表示该类别的分类正确率，其余元素表示来自第i类的图像被误分为第j类的概率。

2.3 实验设置

1) 环境设置：使用基于ImageNet的预训练模型来初始化ResNet50的网络参数。在一台带有i7-8700 CPU和11 GB NVIDIA GeForce GTX 1080Ti GPU的服务器上进行了基于Pytorch深度学习框架的试验。对于模型采用的随机梯度下降优化方法，学习率l_r设置为10^{-3}，每100个迭代轮次l_r变为$l_r=0.1\cdot l_r$，动量因子设置为0.9，权重衰减设置为0.009。

2) 训练设置：为了方便与其他场景分类算法进行综合评估，根据相关参考文献设定了训练比例，因此三个数据集训练比率的设置与对比的参考文献保持一致。其中，对于UCM数据集，将训练比率设置为50%和80%。而对于AID和PatternNet数据集，训练比率被固定为20%和50%。所有输入图像的尺寸统一固定为224×224。此外，为了减轻模型过拟合问题，采用按比例缩放和垂直翻转等数据增强技术。

表2 参数d对AID数据集分类性能的影响

Table 2 Influence of parameter d on classification performance of AID dataset

d	OA (20%)
128	93.02
256	93.08
512	93.11
1 024	**93.70**
2 048	93.23

2.4 参数讨论

由MAFF module生成的多级别特征在进行跨层双线性融合之前，需要利用3个1×1卷积层将多级别特征映射到高维空间。d是高维空间的映射维数，设置合适的d，可以显著增强多级别特征的判别性。在AID数据集上按20%的训练率进行试验，如表2所示，不同的映射维数对模型的分类性能有一定的影响。当映射维数由128上升到1 024时，模型的分类准确率也随之增大。当映射维数为1 024时，模型取得最高的分类准确率，映射维数增加到2 048时，分类准确率反而有所下降。

2.5 消融试验

本文网络模型主要包括3个模块，即MDC、MAFF和CBF模块，为了分析各个模块的有效性，利用提出的3个模块设计出4种不同的结构，并在AID数据集上按20%的训练比率进行消融试验，结果如表3。对于每种结构试验时设置相同的参数，并且每次只删除一个模块。

表3 20%训练比率下AID数据集上的消融实验

Table 3 Ablation experiment on AID dataset with a training ratio of 20%

Architecture	Methods	OA (20%)
1	Without MDC module	93.28±0.33
2	Without MAFF module	91.23±0.25
3	Without CBF	92.91±0.22
4	ResNet50+MDC+MAFF+CBF (Ours)	**93.70±0.11**

1) MDC module的有效性。结构1中省略了MDC module，将ResNet50网络提取到不同层次的特征直接通过MAFF module进行逐级递进融合。尽管结构1与结构2、3相比，获得了更好的性能，但与本文方法(结构4)相比，在20%的训练比率下，整体准确率下降了0.42%，证明了MDC module的优越性。

2) MAFF module的有效性。从表3看出，结构2中省略了MAFF module，将四个通过MDC module的多尺度语义增强特征直接进行跨层双线性融合，然而取得了最低的整体准确率。结构1、3、4中都包含MAFF module，相比于结构2，在20%的训练率下，它们的整体准确率分别提升了2.05%、1.68%、2.47%。结果表明MAFF module通过对不同层次的特征进行有效融合能够显著提升模型的分类精度。

3) CBF的有效性。在结构3中，将通过MAFF module后的$\{\boldsymbol{N}_2, \boldsymbol{N}_3, \boldsymbol{N}_4\}$采用全局平均池化后直接级联的方法取代CBF。结构4与其相比，实现了更加优异的分类性能，在20%的训练率下，整体准确率提升了0.79%，验证了跨层双线性融合方法有助于提高遥感图像场景分类的准确率。

最后，根据在AID数据集上评估的试验结果，当同时引入本文提出的MDC、MAFF和CBP三个模块时，模型取得的整体准确率最高，充分体现了所提场景分类模型的有效性。

2.6 与其他先进方法的对比分析

2.6.1 UCM数据集上的对比

其它先进的场景分类方法与本文方法在UCM数据集上的性能比较如表4所示。UCM数据集的场景类别数最少，类别间的差异也更明显。在80%和50%两种训练比率下，本文方法均获得了最高的分类准确率。ARCNet[17]融合VGG16模型提取的特征进行场景分类，与ARCNet相比较，以80%的样本进行训练时，准确率提升0.2%，20%样本进行训练时，准确率提升1.32%。Siamese ResNet50[19]使用由两个相同的ResNet50模型组成的孪生网络进行场景分类，本文同样采用ResNet50作为基础特征提取网络，在训练样本

表4 在UCM数据集上的分类结果比较

Table 4 Classification result comparison on UCM dataset

Methods	OA (80%)	OA (50%)
GBNet[16]	98.57±0.48	97.05±0.19
ARCNet[17]	99.12±0.40	96.81±0.14
Fusion by Addition[18]	94.72±1.79	
Fine-tuned ResNet50[19]	91.90	89.43
Siamese ResNet50[19]	94.29	90.95
Two-stream fusion[20]	98.02±1.03	96.97±0.75
CNN-CapsNet[14]	99.05±0.24	97.59±0.16
Ours	**99.32±0.20**	**98.13±0.18**

数量为80%时分类准确率比Siamese ResNet50高5.03%，在训练样本数量为50%时高7.18%。证明了本文方法能够进一步提升遥感场景分类的准确率。

UCM数据集上生成的混淆矩阵如图6所示。可以看出，21个场景类别中有19个类别都实现了100%的分类精度。如图7所示，仅森林(forest)和停车场(parking lot)两个类别发生了混淆，被划分到错误的类别中。其中，森林类别的分类精度达到了90%，有10%的图像被误判为高尔夫球场(golf course)类别，因为这

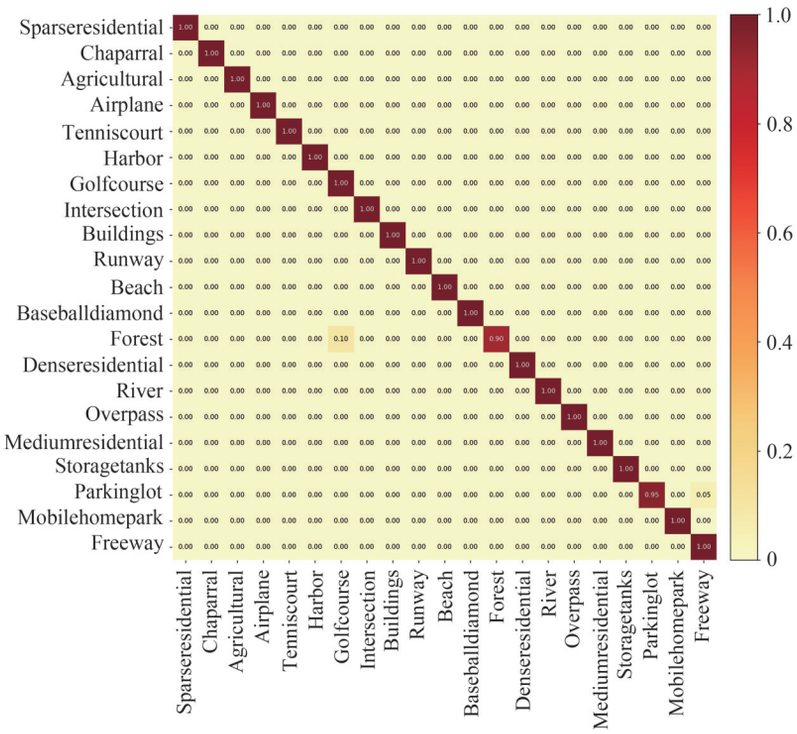

图6 80%训练比率下UCM数据集的混淆矩阵

Fig.6 CM on UCM dataset with a training ratio of 80%

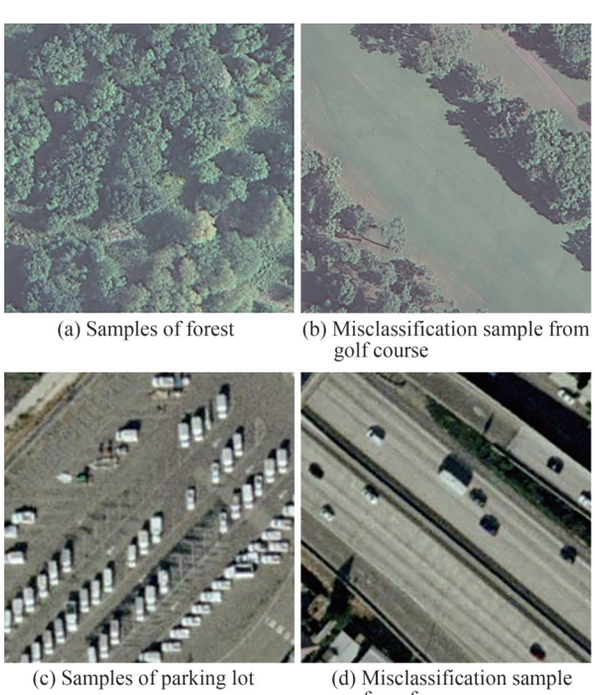

图7 森林和停车场类别典型样本及其被错分的样本

Fig.7 Typical samples of forest and parking lot categories and misclassification samples

两类图像都包含大面积的树木。停车场类别的分类精度为95%，有5%的场景图像被误判为高速公路(freeway)类别，两类场景图像都包含汽车这一相似对象。一些类间差异较小的场景，例如网球场(tennis court)、棒球场(baseball diamond)、建筑物(buildings)，通常会增加模型的分类难度，而利用本文方法实现了这些场景的高精度分类，分类准确率都达到了1，充分证实了所提方法的有效性。

2.6.2 AID数据集上的对比

AID数据集与UCM数据集相比，场景类别的数量扩展到30个，共有10 000张场景图像，像素分辨率变化范围大，进一步增加了AID数据集的分类难度。将该数据集上的训练比率分别设定为50%和20%进行试验，结果如表5。

表5 在AID数据集上的分类结果比较

Table 5 Classification result comparison on AID dataset

Methods	OA (50%)	OA (20%)
ResNet50[21]	91.31±0.58	88.23±0.70
GBNet[16]	94.58±0.12	92.20±0.23
ARCNet[17]	93.10±0.55	88.75±0.40
Two-stream fusion[20]	94.58±0.25	92.32±0.41
Ours	**95.84±0.26**	**93.70±0.11**

从表5中看出，当训练样本数量为50%和20%时，本文方法都超过了其他场景分类方法。本文方法是通过对不同层级特征进行跨层双线性融合实现了分层特征交互，GBNet[16]也使用了分层特征进行特征融合，与其相比能够证明所提的跨层双线性融合方法的有效性，在50%的样本训练时，准确率提升1.26%，在20%的样本训练时，准确率提升1.5%。Two-Stream Fusion[20]采用两种不同的融合策略来融合不同类型的深度卷积特征，使用50%和20%的样本进行训练时，本文方法与其相比，准确率分别提升了1.26%和1.38%。以上对比结果证明本文方法获得了最佳分类性能，并显著提高了AID数据集的分类精度。

图8显示了训练比率为50%时该数据集上生成的混淆矩阵。可以看出，30个场景类别中有21个类别的

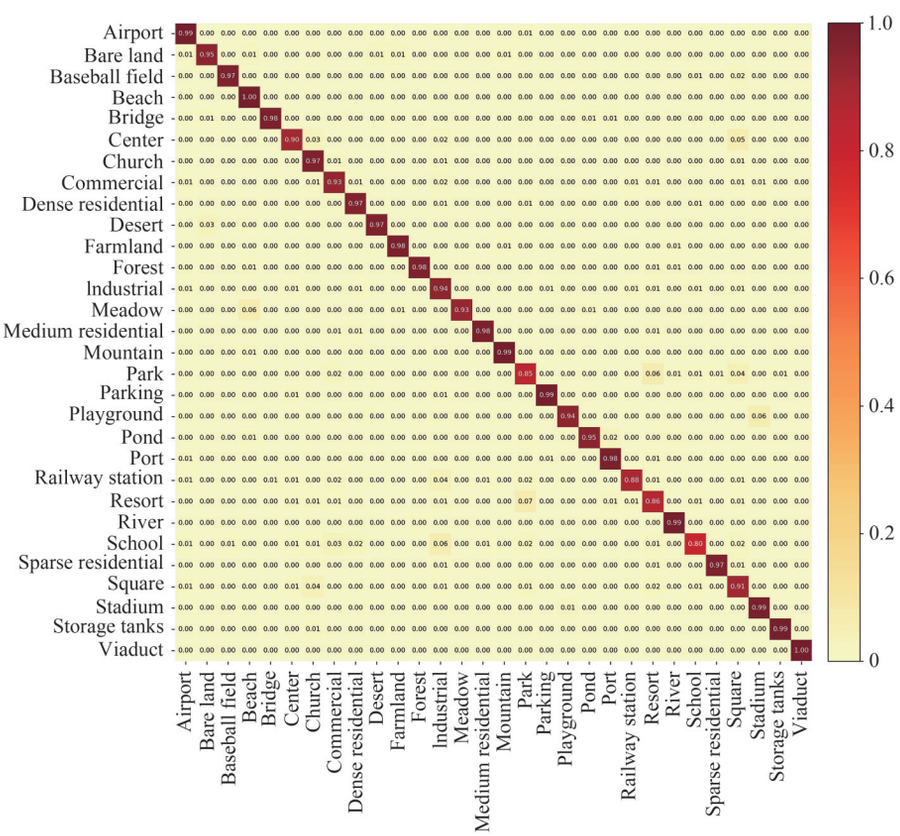

图8 50%训练比率下AID数据集的混淆矩阵

Fig.8 CM on AID dataset with a training ratio of 50%

分类准确率超过95%，仅公园（park）、火车站（railway station）、度假胜地（resort）、学校（school）这四个类别的准确率低于90%。分类准确率最低的场景为学校类别，达到了90%的分类精度，其中大多数该类别样本被错分为了工业（industrial）类别。而一些由于共享类似结构信息难以区分的场景，如稠密住宅区（dense residential）、稀疏住宅区（sparse residential）和中等住宅区（medium residential），分类准确率分别达到97%、97%和98%，可以被模型准确地分类。此外，如图9所示，河流、港口和桥梁具有相似的图像纹理和相同的空间分布，也实现了99%、98%和98%的高分类精度，证实了本文方法的优越性。

图9 AID数据集中河流、港口和桥梁场景类别的部分样本

Fig.9 Samples from river, port and bridge scene categories of AID dataset

2.6.3 PatternNet数据集上的对比

PatternNet数据集与其它两个数据集相比，涵盖了最多的场景类别数。表6为PatternNet数据集上的分类结果对比。对于该数据集，本文方法也实现了最佳性能，整体准确率分别达到99.42%和99.60%。本文方法利用到注意力机制，所以有必要和其他基于注意力机制的场景分类算法进行对比。SDAResNet[23]同时引入空间注意力和通道注意力来提取显著性场景信息，它的分类准确率超过了其它所有的方法，在此基础上，本文方法又将分类结果在50%和20%的训练比率下分别提高了0.02%和0.12%。由此，在该数据集上的表现证实了本文方法能够显著提高遥感图像的场景分类精度。

表6 在PatternNet数据集上的分类结果比较

Table 6 Classification result comparison on PatternNet dataset

Methods	OA (50%)	OA (20%)
GLANet (SVM)[22]	99.40±0.21	98.91±0.19
SDAResNet[23]	99.58±0.10	99.30±0.08
VGGNet (SVM)[24]		97.5±0.02
ResNet101 (SVM)[24]		98.6±0.02
Inception-V3 (SVM)[24]		97±0.02
Ours	**99.60±0.06**	**99.42±0.05**

训练比率为20%时该数据集上生成的混淆矩阵如图10所示，38个场景类别中有34个类别都实现了99%以上的分类精度，绝大部分的场景类别达到了100%的分类准确率，仅养老院（nursing home）、十字路口（overpass）、灌木丛（chaparral）、飞机（airplane）四个场景类别的准确率不到99%。其中，养老院在所有场景类别中的分类精度最低，但也已经取得了95%的准确率。本文算法与SDAResNet[23]相比，在十字路口场景类别上的分类准确率得到了很大的提升，由在SDAResNet模型上进行试验得到的93%提升到了100%，证明了本文方法能够普遍区分复杂度较高的遥感场景类别。

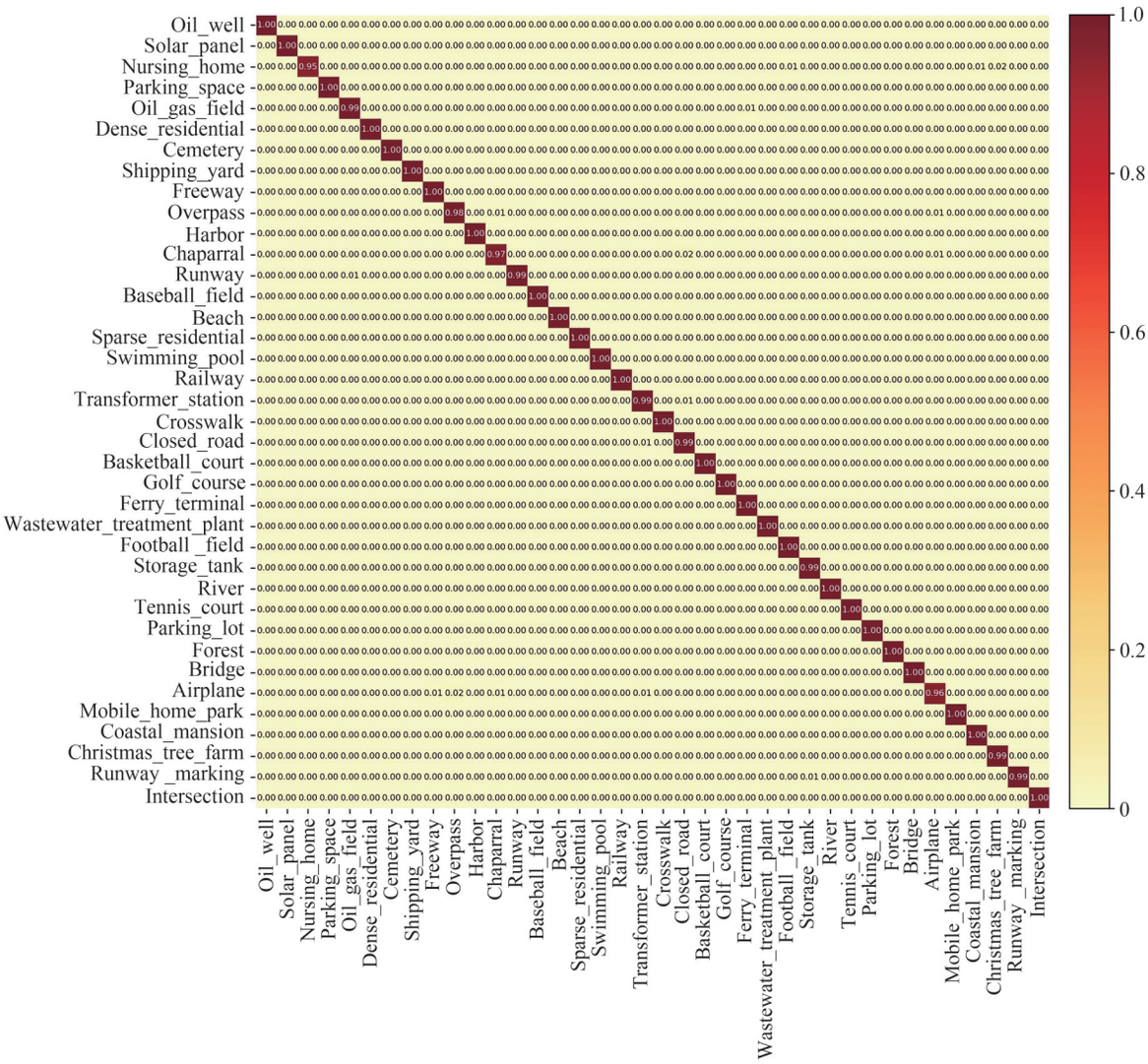

图10　20%训练比率下PatternNet数据集的混淆矩阵

Fig.10　CM on PatternNet dataset with a training ratio of 20%

3　结论

本文提出了一种新的多级别跨层双线性融合网络模型。整体网络模型以ResNet50作为特征提取器，首先获取到多尺度多层次的遥感图像语义特征。多尺度膨胀卷积模块的引入可对膨胀卷积的扩张率进行分支调整，将多个分支提取到的不同空间尺度信息进行融合，在增大感受野的同时丰富了遥感特征的场景信息，有效克服了CNN采用单一固定尺寸卷积核的不足。提出的多级别注意力融合模块，不仅实现了低层、高层、全局上下文特征的信息互补，而且能够利用空间注意力机制加强模型对图像重点区域的关注，避免了冗余背景细节对分类造成的干扰。受到细粒度视觉分类任务的启发，采用跨层双线性融合方法对多级别特征获取二阶双线性信息后进行分层融合，以捕获不同层级间特征的相关性，与基于提取一阶信息的融合方法相比，得到了更具区分性和鲁棒性的融合特征表示。最后，通过在UCM、AID和PatternNet三个广泛使用的数据集上进行训练和测试，达到了比现有方法更加优异的表现，实验结果表明本文方法更有利于遥感图像的场景分类任务。但是由于训练集数量大，网络的训练速度较慢。未来将更加关注于设计轻量级和高精度的网络模型，从提高网络训练速度方面来提高场景分类任务的性能。

参考文献

[1]　XU Suhui, MU Xiaodong, ZHAO Peng, et al. Scene classification of remote sensing image based on multi-scale feature and deep neural network[J]. Acta Geodaetica et Cartographica Sinica, 2016, 45(7): 834-840.

许夙晖，慕晓冬，赵鹏，等. 利用多尺度特征与深度网络对遥感影像进行场景分类[J]. 测绘学报，2016，45(7)：834-840.

[2] REN Jianfeng，JIANG Xudong，YUAN Junsong. Learning LBP structure by maximizing the conditional mutual information[J]. Pattern Recognition，2015，48(10)：3180-3190.

[3] LUO Bin，JIANG Shujing，ZHANG Liangpei. Indexing of remote sensing images with different resolutions by multiple features[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing，2013，6(4)：1899-1912.

[4] CHENG Gong，ZHOU Peicheng，HAN Junwei，et al. Auto-encoder-based shared mid-level visual dictionary learning for scene classification using very high resolution remote sensing images[J]. IET Computer Vision，2015，9(5)：639-647.

[5] YANG Yi，NEWSAM S. Bag-of-visual-words and spatial extensions for land-use classification[C]. Proceedings of the 18th Sigspatial International Conference on Advances in Geographic Information Systems，2010：270-279.

[6] HU Fan，XIA Guisong，HU Jingwen，et al. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery[J]. Remote Sensing，2015，7(11)：14680-14707.

[7] NOGUEIRA K，PENATTI O A B，SANTOS J A. Towards better exploiting convolutional neural networks for remote sensing scene classification[J]. Pattern Recognition，2017，61：539-556.

[8] MA Chenhui，MU Xiaodong，SHA Dexuan. Multi-layers feature fusion of convolutional neural network for scene classification of remote sensing[J]. IEEE Access，2019，99：1-1.

[9] YUAN Yuan，FANG Jie，LU Xiaoqiang，et al. Remote sensing image scene classification using rearranged local features[J]. IEEE Transactions on Geoscience and Remote Sensing，2019，57(3)：1779-1792.

[10] SIMONYAN K，ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv：1409.1556，2014.

[11] ZHANG Bin，ZHANG Yongjun，WANG Shugen. A lightweight and discriminative model for remote sensing scene classification with multidilation pooling module[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing，2019，12(8)：2636-2653.

[12] WANG Panqu，CHEN Pengfei，YUAN Ye，et al. Understanding convolution for semantic segmentation[C]. Proceedings of the 2018 IEEE Winter Conference Applications of Computer Vision，2018：1451-1460.

[13] LIN T Y，ROYCHOWDHURY A，MAJI S. Bilinear cnn models for fine-grained visual recognition[J]. Proceedings of the IEEE International Conference on Computer Vision，2015：1449-1457.

[14] XIA Guisong，HU Jingwen，HU Fan，et al. AID：A benchmark dataset for performance evaluation of aerial scene classification[J]. IEEE Transactions on Geoscience and Remote Sensing，2017，55(7)：3965-3981.

[15] ZHOU Weixun，NEWSAM S，LI Congmin，et al. PatternNet：a benchmark dataset for performance evaluation of remote sensing image retrieval[J]. ISPRS Journal Photogram Remote Sensing，2018，145：97-209.

[16] SUN Hao，LI Siyuan，ZHENG Xiangtao，et al. Remote sensing scene classification by gated bidirectional network[J]. IEEE Transactions on Geoscience and Remote Sensing，2020，58(1)：82-96.

[17] WANG Qi，LIU Shaoteng，CHANUSSOT J，et al. Scene classification with recurrent attention of VHR remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing，2019，57(2)：1155-1167.

[18] CHAIB S，LIU Huang，GU Yanfeng，et al. Deep feature fusion for VHR remote sensing scene classification[J]. IEEE Transactions on Geoscience and Remote Sensing，2017，55(8)：4775-4784.

[19] LIU Xuning，ZHOU Yong，ZHAO Jiaqi，et al. Siamese convolutional neural networks for remote sensing scene classification[J]. IEEE Geoscience and Remote Sensing Letters，2019，16(8)：1200-1204.

[20] YU Yulong，LIU Fuxian. A two-stream deep fusion frame work for high-resolution aerial scene classification[J]. Computational Intelligence and Neuroscience，2018，2018：1-13.

[21] QIAN Xiaoliang，LI Jia，CHENG Gong，et al. Evaluation of the effect of feature extraction strategy on the performance of high-resolution sensing image classification[J]. Journal of Remote Sensing，2018，22(5)：758-776.
钱晓亮，李佳，程塨，等. 特征提取策略对高分辨率遥感图像场景分类性能影响的评估[J]. 遥感学报，2018，22(5)：758-776.

[22] GUO Yiyou，JI Jinsheng，LU Xiankai，et al. Global-local attention network for aerial scene classification[J]. IEEE Access，2019，7：67200-67212.

[23] GUO Dongen，XIA Yang，LUO Xiaobo. Scene classification of remote sensing images based on saliency dual attention residual network[J]. IEEE Access，2020，8:1-1.

[24] SHAFAEY M A，SALEM A M，EBEID H M，et al. Comparison of CNNs for remote sensing scene classification[C]. Proceedings of the International Conference on Computer Engineering and Systems，2018：27-32.

Optical Remote Sensing Image Scene Classification Based on Multi-level Cross-layer Bilinear Fusion

YU Tianwei[1], ZHENG Enrang[1], SHEN Junge[2], WANG Kai[3]
(1 *School of Electrical and Control Engineering, Shaanxi University of Science and Technology, Xi'an* 710021, *China*)
(2 *Unmanned System Research Institute, Northwestern Polytechnical University, Xi'an* 710072, *China*)
(3 *Henan Key Laboratory of Underwater Intelligent Equipment, Zhengzhou* 450000, *China*)

Abstract: Remote sensing, a kind of detection technology, provides non-contact surface observation through sensor platform. With the rapid development of unmanned aerial vehicle, remote sensing and satellites technology, quantitative remote sensing images with higher resolution can be generated. Compared with medium and low-resolution remote sensing images, these high-resolution remote sensing images contain richer ground objects and spatial details, which can express the spatial structure and texture features of ground object more clearly, providing good conditions and foundation for remote sensing image interpretation and analysis. Therefore, high-resolution remote sensing images have become an important data source for fine earth observation. The scene classification of high-resolution remote sensing image refers to the analysis of extracted remote sensing image information, dividing the scene image of interest to different categories, such as forest, river, railway, etc., and is widely applied in environmental monitoring, urban planning, military object detection, global climate change research and other fields. Unlike general natural images, the geometry structure and space pattern of remote sensing images are highly complex, and there are also problems such as complex background and many types, which is a great challenge for effectively describing remote sensing image content. In addition, as a result of the complexity and diversity of remote sensing image scenes, different scenes may contain almost the same ground object targets, or the same scene may contain different ground object targets. At this regard, how to design discriminative feature representation to describe the image directly affects the quality of scene classification. In the past few decades, many approaches have been proposed, and most of these methods can be divided into two main categories. The traditional scene classification methods, such as Scale Invariant Feature Transform (SIFT), Histogram of Oriented Gradients (HOG) and Color Histogram (CH), mainly use hand-crafted feature but highly depend on the priori knowledge of the designer, resulting in the features with low-level semantics and limited representational capacity. By contrast, convolutional neural network has been successfully applied in remote sensing scene classification as its excellent feature self-learning ability. It can learn features directly from data without the need of priori knowledge of the designer. However, the accuracy of the scene classification approach based on CNN largely depends on the network structure and due to the complex spatial patterns, large inter-class similarity and high intra-class diversity of remote sensing scene images, the scene classification accuracy is severely limited. To address above issues, a novel remote sensing image scene classification algorithm via multilevel cross-layer bilinear fusion is proposed. Firstly, ResNet50 avoids the issues of model overfitting and gradient vanishing. It is employed to extract the remote sensing image multi-level features. In this way, the four multi-scale multi-level features of conv2_x, conv3_x, conv4_x and conv5_x layers were extracted by ResNet50 model. The dilated convolution with different expansion rates can perceive scene information at multiple spatial scales, promoting the network to acquire features at different scales. The context features at multiple spatial scales are extracted by setting the expansion rate of dilated convolution to different values. Then, the scene semantics of the feature information is enriched by serial fusion of multi-scale features. Since features at different levels contain different types of information, the high-level features provide global semantic information, which is helpful to identify and locate objects in the image. On the contrary, the low-level features contain rich local spatial information to refine and enrich the internal structure of salient objects. Such features can help high-level features to complement their loss of spatial information, which is beneficial for classification. The global context information of an image has a global receptive field. Considering the global information, the scene category can be inferred and the interference of background details can be filtered. By taking the advantages of low-level, high-level, and global context features, a multilevel attention feature fusion module is presented, which can effectively enhance the feature extraction capability of the model. The spatial attention is designed to focus on the key location of the scene image, which adaptively learn the importance of different image regions, depressing the irrelevant information of

background. The global context information is integrated into the feature fusion process of low-level local features and high-level semantics features to realize the complementary feature information of each level, resulting in pleasing scene classification accuracy. Finally, inspired by fine-grained visual classification, a cross-layer bilinear fusion method is utilized to perform layered fusion of multilevel features, and the fused features are used for classification. Hadamard product operation at any two different levels is utilized to extract second-order bilinear information. Based on this cross-layer modeling to capture the association between local features, the hierarchical feature interaction and efficient information integration can be achieved, and the deep semantic information and shallow texture information contained in different hierarchical features are fully aggregated. Moreover, compared with the traditional bilinear pooling method, the Hadamard product is the product of two matrices′ corresponding elements, which does not change the dimension of the matrix, effectively solved the dimension explosion caused by the outer product operation. Through extensive experiments conducted on the UCM, AID and PatternNet datasets, the effectiveness of the proposed method is verified. Compared with other advanced approaches, the proposed method achieves more excellent classification performance. On the UCM dataset, for training with 80% data, the overall accuracy reached 99.32%, and the classification accuracy is increased by 0.75% compared with GBNet. On the AID dataset, the proposed method achieved 95.84% accuracy in 50% of training samples, with an improvement of 2.74% compared with ARCNet. On the PatternNet dataset, 50% of the samples are trained, and the overall accuracy is 99.6%, that has increased by 0.02% compared with SDAResNet.

Key words: Remote sensing; Scene classification; Dilated convolution; Multi-level attention; Cross-layer bilinear fusion

OCIS Codes: 100.4996; 100.1830; 150.0155; 100.3008

Foundation item: National Natural Science Foundation of China (No. 61603233)