引文格式:时永欣,周维勋,邵振峰.融合多尺度注意力的多视角遥感影像场景分类[J].武汉大学学报(信息科学版),2024,49(3):366-375.DOI:10.13203/j.whugis20220737

Citation: SHI Yongxin, ZHOU Weixun, SHAO Zhenfeng. Multi-view Remote Sensing Image Scene Classification by Fusing Multi-scale Attention[J]. Geomatics and Information Science of Wuhan University, 2024, 49(3): 366-375. DOI: 10.13203/j.whugis20220737

融合多尺度注意力的多视角遥感影像场景分类

时永欣[1] 周维勋[1,2] 邵振峰[3]

1 南京信息工程大学遥感与测绘工程学院,江苏 南京,210044

2 北京师范大学遥感科学国家重点实验室,北京,100875

3 武汉大学测绘遥感信息工程国家重点实验室,湖北 武汉,430079

摘 要:针对现有场景分类方法特征表征能力差以及单视角遥感影像分类精度难以提升的问题,提出一种融合多尺度注意力的多视角遥感影像场景分类方法。首先,将航空图像和地面图像构造成正负图像对,并划分为训练集、验证集和测试集;其次,构建融合多尺度注意力的卷积神经网络并训练,通过特征融合模块得到融合注意力且表征能力更强的特征,实现多尺度特征学习;然后,利用训练的多尺度注意力网络分别提取航空图像和地面图像特征并进行融合;最后,基于融合后的特征使用支持向量机进行场景分类。实验结果表明,相比现有方法,所提方法在两个公开数据集上均取得了更高的分类精度,改善了单视角场景分类效果,同时也证明了多视角所提供的补充信息能进一步提升遥感场景分类的准确性。

关键词:场景分类;多视角遥感图像;卷积神经网络;特征融合;视觉注意力

中图分类号:P237 **文献标识码**:A **收稿日期**:2023-02-24

DOI:10.13203/j.whugis20220737 **文章编号**:1671-8860(2024)03-0366-10

Multi-view Remote Sensing Image Scene Classification by Fusing Multi-scale Attention

SHI Yongxin[1] *ZHOU Weixun*[1,2] *SHAO Zhenfeng*[3]

1 School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China

2 State Key Laboratory of Remote Sensing Science, Beijing Normal University, Beijing 100875, China

3 State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China

Abstract: **Objectives**: Remote sensing scene classification provides new possibilities for the application of high-resolution images, and how to effectively realize scene recognition from high-resolution remote sensing images is still an important challenge. The existing scene classification methods only use remote sensing images from one viewpoint for scene recognition, which cannot accurately express the semantic information of complex high-resolution remote sensing images, and the accuracy of scene classification is difficult to be further improved. **Methods**: To solve this problem, a multi-view scene classification method for remote sensing images is proposed. First, the aerial image and ground image are constructed into a positive and negative image pair, and divided into training dataset, validation dataset and test dataset. Second, a convolutional neural network with fusion multi-scale attention is constructed, and features with fusion attention and strong representation ability are obtained through feature fusion module, so as to integrate different feature information and realize multi-scale feature learning. Third, the trained multi-scale attention network is used to extract features from aerial image and ground image, respectively. Finally, the

基金项目:国家自然科学基金(42001285);江苏省自然科学基金(BK20200813);遥感科学国家重点实验室开放基金(OFSLRSS202215);自然资源部国土卫星遥感应用重点实验室开放基金(KLSMNR-G202202)。

第一作者:时永欣,硕士生,研究方向为遥感信息智能提取。1033424423@qq.com

通讯作者:周维勋,博士,讲师。zhouwx@nuist.edu.cn

fused features are used to classify scenes based on the fused features using support vector machine. To demonstrate the performance of the proposed multi-scale attention network, we conduct experiments on two publicly available benchmark datasets — the AiRound and the CV-BrCT datasets. **Results:** The proposed method achieves remarkable performance, with the highest accuracy of 93.13% in the AiRound dataset and 85.18% in the CV-BrCT dataset, which improves the accuracy of single-view scene classification. **Conclusions:** The results demonstrate that the complementary information provided by multi-view images can further improve the performance of remote sensing scene classification.

Key words: scene classification; multi-view remote sensing image; convolutional neural network; feature fusion; visual attention

遥感影像场景分类是从高分辨率遥感影像中提取场景级语义信息,从而为影像分配一个语义类别[1-2]。在自然植被制图、土地利用、环境监测、国土资源调查等领域得到了广泛应用[3]。

现有的场景分类方法主要分为两类,包括基于手工特征和基于深度特征的方法[4]。基于手工特征的方法主要是提取遥感影像的光谱、纹理和形状等底层特征用于分类。例如文献[5]将纹理应用于森林地区的分类中,改善了高分辨率影像的分类性能。文献[6]将光谱与纹理相结合进行分类,提高了影像分类任务的准确性。基于深度特征的方法主要是利用卷积神经网络(convolutional neural network, CNN)提取特征进行分类[7]。CNN是一种典型的深度学习架构,具有强大的影像特征提取能力,完成特征从低层向高层的抽象化过程,形成图像的分层表达,在图像分类中取得了重要的突破。鉴于其优越性能,CNN在遥感场景分类领域也被广泛采用[8-12]。文献[9]提出一种基于多源数据的遥感知识感知与多尺度特征融合网络,高效挖掘多源遥感数据中的遥感知识信息,提高网络对地物多尺度特征的学习能力,细化最终的地物分类结果;文献[10]提出联合显著性和多层CNN的方法,利用深度CNN从样本集中提取高层次特征,更好地表达了场景信息;文献[11]提出端对端的多尺度联合CNN模型,通过对多个通道不同尺度的高层特征进行联合增强表达,实现了在小样本训练集上的高精度分类;文献[12]提出了用于遥感场景分类的双分支卷积神经网络,通过度量学习提高了分类精度。

上述遥感场景分类方法侧重于单视角(如卫星或航空)影像,研究表明,利用其他视角图像提供的补充信息能够进一步提高分类性能。例如文献[13]基于多个CNN模型,通过早期和晚期融合策略进行多视角的场景分类。文献[14]利用孪生网络提出互补信息学习模型,实现了航空和地面影像的多视角场景分类。航空-地面双视角图像作为一种特殊类型的多模态遥感数据,在遥感图像处理任务中被广泛使用,近年来,组合、利用多模态遥感数据进行遥感图像处理任务已成为遥感应用的一个重要发展方向。文献[15]提出多模态特征学习模型,将多模态遥感数据分解为模态共享表示和模态特定表示,建立多模态遥感数据集之间的映射关系,实现了土地覆盖分类任务。文献[16]提出基于多传感器融合和明确语义保存的深度哈希算法,通过融合多光谱影像来消除空间-光谱差异。文献[17]提出基于注意力的多尺度残差适应网络用于跨域场景分类,有效地解决了跨模态数据特征提取无法对齐等问题。文献[18]针对目标域中存在未知类别而影响分类精度的问题,采用分离机制区分目标域中已知类别和未知类别,并将已知类别用于跨域对齐和分类。文献[19]提出了一种通用的跨模态遥感信息关联学习方法,解决了多模态遥感信息之间的异质鸿沟问题,实现了更准确的跨模态检索。

将多视角图像所提供的互补信息组合起来进行场景分类任务,是提升遥感场景分类性能的一个重要分支。虽然文献[13]和文献[14]利用其他视角图像提供的互补信息改善了场景分类性能,但二者用于分类的特征均是通过简单的特征融合得到的,特征表征能力较差,从而导致分类性能提升受限。因此,本文提出了一种融合多尺度注意力的卷积神经网络用于多视角场景分类。该网络包括两个分支,分别输入影像对中的航空图像和地面图像,经过特征融合模块实现多尺度特征提取。为了进行多视角分类,将航空图像和地面图像的特征融合后利用支持向量机(support vector machine,SVM)进行分类,并在两个标准数据集上对分类精度进行了分析。

1 研究方法

1.1 网络结构

受文献[14]和文献[20]的启发,本文提出多尺度注意力网络(multi-scale attention network,MSAN)用于高分辨率遥感场景分类,该网络由两个完全相同的子网络和3个全连接层(FC_a、FC_g和FC_{ag})组成,两个子网络不共享权值,单个子网络由CNN网络分支和特征融合模块构成,具体如图1所示。

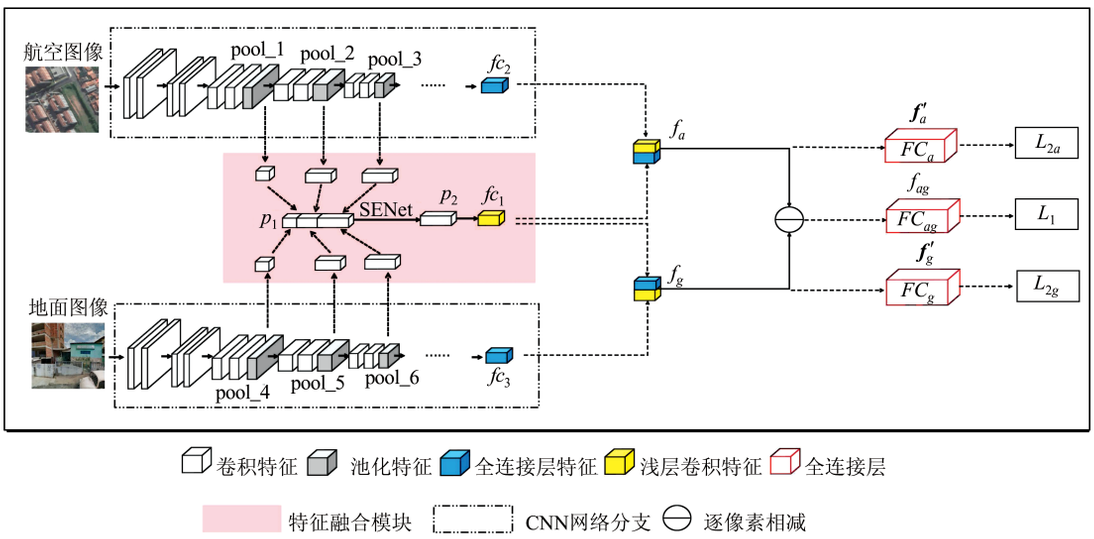

图1 多尺度注意力网络结构

Fig. 1 Architecture of Multi-scale Attention Network

MSAN以航空图像和地面图像构成的图像对作为输入。具体分为4个步骤:(1)将航空图像和地面图像分别输入到两个子网络,子网络的输出(f_a和f_g)通过逐像素相减操作后输入到包含单个神经元的全连接层FC_{ag};(2)FC_{ag}的输出值使用Sigmoid函数映射到[0,1],该步骤能够使FC_{ag}的输出值衡量航空图像和地面图像的匹配程度,当航空图像和地面图像来自同一类别时,该数值趋向于1,反之数值趋于0;(3)子网络的输出(f_a和f_g)分别输入全连接层FC_a和FC_g,使其转化为N维特征向量($\boldsymbol{f}_a'$和$\boldsymbol{f}_g'$),其中,N为场景类别数,该步骤能进一步增强模型的鲁棒性,加快模型的收敛速度;(4)为了充分利用两个视角图像所提供的补充信息,在训练结束后,提取子网络的输出(f_a和f_g)作为学习的特征,并将f_a和f_g融合后用于训练SVM分类器,进行多视角场景分类。

1.2 特征融合模块和损失函数

1.2.1 注意力特征融合模块

CNN不同层次的卷积特征包含不同的遥感场景信息,浅层卷积特征中的结构信息比较丰富,而深层卷积特征中的语义信息比较丰富。为了充分利用CNN不同层特征的互补性,设计了如图1所示的特征融合模块。该模块以CNN网络分支的最后3个池化层作为输入,池化层融合后的特征p_1经过注意力模块SENet(squeeze and excitation network)学习特征通道之间的相关性,得到融合注意力[21]且表征能力更强的特征p_2,经过池化运算后得到卷积特征fc_1。

在特征融合的过程中,需要统一池化层的尺寸。如果将3个池化层统一到最后一个池化层的尺寸,由于前两个池化层需要较大的步长,会丢失网络浅层中的结构信息,导致无法充分地利用CNN网络分支的浅层信息[20]。为解决这一问题,在特征融合模块中,本文将不同池化层统一成中间池化层(图1的pool_2与pool_5)相同的宽度和高度。具体来说,利用平均池化对第一层池化层(图1的pool_1与pool_4)进行下采样,利用转置卷积运算对第三层池化层(图1的pool_3与pool_6)进行上采样。此外,在统一尺寸的过程中,利用标准化将统一尺寸的特征图在通道维度归一化至[0,1],计算式为:

$$\hat{\boldsymbol{x}}^i_{H,W,C}=\frac{\boldsymbol{x}^i_{H,W,C}}{\sqrt{\sum_{j=1}^{C}\left(\boldsymbol{x}^i_{H,W,j}\right)^2+\varepsilon}} \tag{1}$$

式中,$\boldsymbol{x}^i_{H,W,C}$为图1中CNN网络分支中三层池化层经过尺寸统一处理后的特征图,$i=1,2,3$;$\hat{\boldsymbol{x}}^i_{H,W,C}$为标准化后的特征图;$H$、$W$、$C$分别为特征图的高度、宽度、通道数。

将图 1 中 CNN 网络分支中的第一层池化层(图 1 的 pool_1 与 pool_4)、第二层池化层(图 1 的 pool_2 与 pool_5)、第三层池化层(图 1 的 pool_3 与 pool_6)统一尺寸后分别进行式(1)中的标准化,然后通过级联聚合运算聚合其通道数,串联生成图 1 中的 p_1 特征。级联聚合计算式为:

$$\tilde{\boldsymbol{x}}=\mathrm{cat}(\hat{\boldsymbol{x}}^1,\hat{\boldsymbol{x}}^2,\hat{\boldsymbol{x}}^3) \tag{2}$$

式中,$\hat{\boldsymbol{x}}^i\in R^{H\times W\times C_i}$、$\tilde{\boldsymbol{x}}\in R^{H\times W\times(C_1+C_2+C_3)}$ 为特征图,$i=1,2,3$;cat 代表聚合操作。

对于注意力模块 SENet,其作用是通过学习的方式来自动获取到每个特征通道的重要程度,将重要的浅层卷积特征增强,不重要的浅层卷积特征减弱。SENet 模块主要通过压缩、激励和重定权重 3 个步骤实现特征的重新标定[21],如图 2 所示。

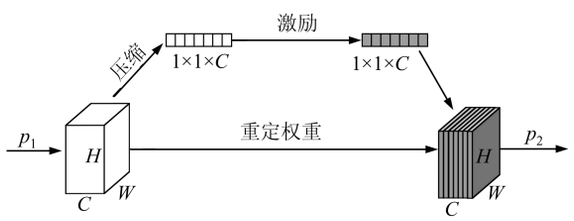

图 2　SENet 模块

Fig. 2　SENet Module

1)压缩部分主要目的是进行全局信息的嵌入,通过平均池化运算将特征层的长宽进行压缩并留下通道维度的信息,计算式为:

$$F_{sq}(\boldsymbol{P}_C)=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{P}_C(i,j) \tag{3}$$

式中,$\boldsymbol{P}_C$ 为 p_1 特征;压缩过程将 $H\times W\times C$ 的 p_1 特征图转化为 $1\times 1\times C$ 的 $F_{sq}(\boldsymbol{P}_C)$ 特征图。

2)激励部分主要是对压缩后的特征重新标定,计算式为:

$$F_{ex}(F_{sq}(\boldsymbol{P}_C),\boldsymbol{Q})=\sigma(g(F_{sq}(\boldsymbol{P}_C),\boldsymbol{Q})=\sigma(\boldsymbol{W}_2\delta(\boldsymbol{W}_1(F_{sq}(\boldsymbol{P}_C)))) \tag{4}$$

式中,$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 为相应特征通道所生成的权重参数矩阵($\boldsymbol{W}_1$、$\boldsymbol{W}_2\in\boldsymbol{Q}$);$\boldsymbol{Q}$ 是权重矩阵集合;$\delta(\cdot)$ 和 $\sigma(\cdot)$ 分别表示 ReLU 和 Sigmoid 激活函数。

激励部分通过两个全连接层实现,其中第一个全连接将特征的 C 个通道压缩至 C/r 个通道,即式(4)中的 $\boldsymbol{W}_1(F_{sq}(\boldsymbol{P}_C))$ 部分(后连接 ReLU 函数),其中 r 是指压缩比例,文献[21]证明 $r=16$ 时整体性能和计算量最平衡。第二个全连接作用是再将其特征恢复为 C 个通道,即式(4)中的 $\boldsymbol{W}_2\delta(\boldsymbol{W}_1(F_{sq}(\boldsymbol{P}_C)))$ 部分(后连接 Sigmoid 函数)。

3)重定权重是将激励后的特征与最原始的特征 p_1 相乘,并输出一个新的特征 p_2。整个过程通过显式的建模通道之间的互相依赖关系,自适应地重新校正通道的特征响应。

$$F_{\mathrm{scale}}(\boldsymbol{P}_C,F_{ex}(F_{sq}(\boldsymbol{P}_C),\boldsymbol{Q}))=\boldsymbol{P}_C\cdot F_{ex}(F_{sq}(\boldsymbol{P}_C),\boldsymbol{Q}) \tag{5}$$

融合后的特征 fc_1 与子网络的特征 fc_2、fc_3 经式(1)中的级联聚合后分别得到多尺度特征 f_a 和 f_g。综上所述,特征融合模块以 CNN 网络分支的最后 3 个池化层特征作为输入,其中图 1 的 pool_1、pool_2 和 pool_3 为航空图像 CNN 网络分支的最后 3 个池化层特征,图 1 的 pool_4、pool_5 和 pool_6 为地面图像 CNN 网络分支的最后 3 个池化层特征,整个流程主要分为 5 部分。首先将 3 个池化层的特征统一成中间池化层尺寸大小,即 pool_1 与 pool_4 通过平均池化进行下采样,pool_3 与 pool_6 通过转置卷积进行上采样,池化和转置卷积步长均为 2;然后将航空图像和地面图像统一尺寸后的三层池化特征图分别经标准化串联,得到融合特征 p1;其次是将 p1 特征经过 SENet 模块进行更新得到特征 p2;接着将特征 p2 经过平均池化运算得到新的卷积特征 fc_1;最后将特征融合模块的输出特征 fc_1 与网络分支的输出特征 fc_2、fc_3 融合后得到可用于分类的特征 f_a 和 f_g。

1.2.2　损失函数

网络采用二元交叉熵损失函数 L_1 和多元交叉熵损失函数 L_2 组合而成的混合损失函数进行网络训练,计算式为:

$$L=\lambda_1 L_1+\lambda_2 L_2 \tag{6}$$

式中,λ_1 和 λ_2 表示平衡两个损失函数之间的权重,通过对比不同数值的收敛情况将 λ_1 的值设置为 1,λ_2 的值设置为 0.5。L_1 的计算式为:

$$L_1=-\frac{1}{N}\sum_{i=1}^{N}\left[y_i\log(f(y_i))+(1-y_i)\log(1-f(y_i))\right] \tag{7}$$

式中,N 表示航空图像和地面图像所构成的图像对数目;y_i 表示图像对 i 真实标签;$f(y_i)$ 表示图像对 i 的预测标签,$f(y_i)=\mathrm{Sigmoid}(f_{ag})$,$f_{ag}$ 表示 FC_{ag} 层的输出值。

多元交叉熵损失函数 L_2 由用于航空视角图像分类的 L_{2a} 和用于地面视角图像分类的 L_{2g} 组成,计算式分别为:

$$L_{2a}=-\sum_{i_a=1}^{N_a}y_i^a\log(f(y_{i_a}^a)) \tag{8}$$

$$L_{2g}=-\sum_{i_g=1}^{N_g} y_i^g \log\left(f\left(y_{i_g}^g\right)\right) \tag{9}$$

式中,N_a表示地面图像数目;$y_{i_a}^a$表示地面图像i_a的真实标签;$f(y_{i_a}^a)$表示地面图像i_a的预测标签,$f(y_{i_a}^a)=\mathrm{softmax}(f_a')$,$f_a'$为$FC_a$层的输出值;$N_g$表示航空图像数目;$y_{i_g}^g$是航空图像$i_g$的真实标签;$f(y_{i_g}^g)$是航空图像$i$的预测标签,$f(y_{i_g}^g)=\mathrm{softmax}(f_g')$,$f_g'$为$FC_a$层的输出值。

2 实验结果与分析

2.1 实验数据

本文使用的数据集是由文献[13]提供的两个多视角数据集。第一个数据集是AiRound数据集,由11类、11 753幅图像构成,包括机场、桥、教堂、森林、湖泊、河流、摩天大楼、体育场、雕像、塔和公园。第二个数据集是CV-BrCT数据集,由9类、24 000对图像构成,包括公寓、医院、住宅、工业、停车场、宗教、学校、商店和空地。两个数据集的每个类别都包含航空视角和地面视角图像。图3给出了两种数据集的部分图像对。

图3 两种数据集的部分航空图像和地面图像

Fig. 3 Aerial and Ground Example Images of Two Datasets

训练时,将图像对的大小分别调整为AlexNet网络需要的227×227和VGG16网络需要的224×224。利用Adam优化器优化损失函数,其中梯度衰减和平方梯度衰减因子分别设为0.9和0.99。损失函数λ_1设置为1,λ_2设置为0.5。具体训练参数如表1所示。

表1 训练参数

Tab. 1 Training Parameters

数据集	网络模型	批量大小	学习率/10^{-5}	迭代次数
AiRound	MSAN-AlexNet	80	2	500
	MSAN-VGG16	6	0.2	2 500
CV-BrCT	MSAN-AlexNet	80	2	3 000
	MSAN-VGG16	6	0.2	15 000

2.2 实验设置

将AiRound和CV-BrCT数据集中各类别按照6:2:2随机划分为训练集、验证集和测试集。为了与现有方法[12,14,22-23]进行公平比较,分别选取预训练网络AlexNet[7]和VGG16[24]作为图1中的CNN网络分支,并移除每个网络分支的最后一个全连接层、softmax层和分类输出图层,其中MSAN-AlexNet、MSAN-VGG16分别代表图1中CNN网络分支使用的AlexNet、VGG16网络。

2.3 结果分析

2.3.1 多尺度注意力网络分类结果

利用训练后的MSAN分别提取多视角和单视角航空和地面图像的特征,并将两种视角的特征融合后进行SVM分类,结果见表2和表3。研究表明,使用SVM分类器进行场景分类任务性能优于softmax分类器[25]。

表2 多视角分类结果/%

Tab. 2 Multi-view Classification Results/%

方法	AiRound数据集	CV-BrCT数据集
MSAN-AlexNet	92.27	84.92
MSAN-VGG16	93.13	85.19

表 3　单视角分类结果/%

Tab. 3　Single-View Classification Results/%

方法	AiRound 数据集		CV-BrCT 数据集	
	航空图像	地面图像	航空图像	地面图像
MSAN-AlexNet	81.97	82.83	78.49	64.63
MSAN-VGG16	81.87	84.12	78.84	67.25

由表 2 可以看出,在两种数据集上,VGG16 作为网络分支的分类精度比 AlexNet 作为网络分支的精度稍高,因为 VGG16 网络更深,特征表征能力更强。当前,多数用于遥感场景分类的影像为航空视角图像,通过航空遥感图像的整体空间表征进行遥感场景分类。尽管航空视角遥感图像有明显优势,但其视角特殊性会使航空图像背景复杂度高,场景中的地物也易受到植被的覆盖、云层的遮挡等影响,无法详细地探究地面所包含的地物信息。地面视角图像在一定程度上能弥补航空视角图像的缺陷,将地面视角图像和航空视角图像中的互补信息结合能极大程度上解决单一视角带来的问题。为了进一步分析多视角图像的补充信息对分类性能的提升,利用多尺度注意力网络提取单视角图像的特征进行 SVM 分类。由表 2 和表 3 可以看出,对于两种数据集,多视角分类显著提升了单视角分类的精度。

此外,图 4 给出了表 2 和表 3 中航空视角图像、地面视角图像以及组合两个视角图像的分类结果示例。由图 4 可以看出,多视角的补充信息能极大提高图像分类精度。以图 4 中机场为例,航空图像和地面图像的单视角分类结果均是错误的,但同时使用两个视角的图像能够被正确识别。

图 4　单视角和多视角图像分类示例

Fig. 4　Example of Single-View and Multi-view Image Classification

图 5 给出了 MSAN-VGG16 在两个数据集上分类的混淆矩阵。由混淆矩阵可知,在 AiRound 数据集中,湖泊类别分类精度为 55.56%,在各类别中最低。这是因为分类结果为湖泊的图像中有 44.44% 的真实类别是河流,二者混淆严重。

图 6 为 AiRound 数据集中河流分类正确的图像,以及部分湖泊被错误分类为河流的图像。由图 6 可以看出,分类错误的湖泊和河流相似性较高,很难辨别。对于 CV-BrCT 数据集,住宅准确率最高,医院类别准确率最低。这是因为分类时医院和公寓、学校、商店、宗教存在混淆。

图 7 为 CV-BrCT 数据集中医院和空地分类错误的图像,其中无论医院的航空图像还是地面图像都很难辨认为医院,此外,空地很大一部分被错误分类为住宅。图 7 中空地被分类错误的航空和地面图像中大部分都包含住宅,这是导致类别混淆的主要原因之一。

2.3.2　消融实验

为了进一步验证 MSAN 的性能,采用相同的参数设置并围绕是否特征融合、损失函数与权值是否共享进行了多组消融实验。对于特征融合消融实验,不进行特征融合表示直接使用 CNN 网络分支的卷积特征,即图 1 中的 fc_2,进行后续分类。消融实验的具体结果见表 4,其中 $M1$ 表示网络使用二元交叉熵损失函数并且进行特征融合,$M2$ 表示网络使用混合损失函数但不进行特征融合,$M3$ 表示网络使用混合损失函数并且进行特征融合。从表 4 可以看出,对于多视角分类,相比共享权值,不共享权值的网络整体效果更好。这是由于航空图像和地面图像属于两种不同角度拍摄的图像且差异较大,分别训练各自的子网络能够更好地学习不同视角的特征,使得网络性能更好。

通过对比 3 种方法可知,无论网络分支是

AlexNet还是VGG16,$M3$的分类精度均是最高的,而$M1$的分类精度均是最低的。结果表明,混合损失函数和特征融合都能有效地提高网络分类精度。

	机场	桥	教堂	森林	湖泊	河流	摩天大楼	体育场	雕像	塔	公园
机场	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
桥	0.00%	91.30%	0.00%	0.00%	0.00%	4.35%	0.00%	0.00%	0.00%	0.00%	4.35%
教堂	0.00%	0.00%	90.91%	0.00%	0.00%	0.00%	9.09%	0.00%	0.00%	0.00%	0.00%
森林	0.00%	0.00%	0.00%	81.82%	0.00%	4.55%	0.00%	0.00%	0.00%	0.00%	13.64%
湖泊	0.00%	0.00%	0.00%	0.00%	55.56%	44.44%	0.00%	0.00%	0.00%	0.00%	0.00%
河流	0.00%	0.00%	0.00%	0.00%	3.57%	96.43%	0.00%	0.00%	0.00%	0.00%	0.00%
摩天大楼	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%
体育场	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%
雕像	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	5.56%	94.44%	0.00%	0.00%
塔	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%
公园	0.00%	0.00%	0.00%	9.09%	0.00%	0.00%	0.00%	0.00%	4.55%	0.00%	86.36%

(a) AiRound数据集

	公寓	医院	住宅	工业	停车场	宗教	学校	商店	空地
公寓	91.14%	0.10%	4.67%	0.10%	0.00%	0.76%	0.95%	2.29%	0.00%
医院	27.63%	21.05%	0.00%	2.63%	0.00%	7.89%	31.58%	9.21%	0.00%
住宅	2.43%	0.00%	95.27%	0.33%	0.00%	0.53%	0.72%	0.59%	0.13%
工业	0.68%	0.17%	2.54%	86.44%	0.17%	1.02%	4.92%	4.07%	0.00%
停车场	3.67%	0.00%	0.00%	1.22%	93.88%	0.00%	0.41%	0.82%	0.00%
宗教	8.33%	0.57%	11.49%	2.87%	0.00%	56.61%	12.07%	8.05%	0.00%
学校	3.81%	0.19%	3.43%	5.33%	0.00%	3.43%	81.33%	2.48%	0.00%
商店	9.87%	0.26%	4.94%	5.45%	0.00%	5.45%	3.90%	70.13%	0.00%
空地	1.06%	0.00%	31.91%	1.06%	0.00%	0.00%	0.00%	0.00%	65.96%

(b) CV-BrCT数据集

图5 两种数据集分类结果混淆矩阵

Fig. 5 Confusion Matrix for Two Datasets

图6 AiRound数据集河流和湖泊部分分类结果

Fig. 6 Classification Results of River and Lake in AiRound Dataset

2.3.3 对比实验

表5给出了本文的MSAN与其他6种方法的对比结果,6种对比方法均使用航空图像和地面图像进行多视角的场景分类。其中,预训练CNN方法和微调CNN方法均不使用特征融合模块。预训练CNN方法是直接使用预训练网络AlexNet和VGG16进行多视角遥感场景分类。微调CNN方法是提前根据数据集中的地面图像和航空图像微调AlexNet和VGG16网络的所有参数,然后将微调后的网络用于多视角遥感场景分类。此外,NBCL模型是文献[22]提出的基于邻域的协同学习遥感场景分类方法,融合多级特

征以增强图像表达的鉴别能力;VGG_VD16+SAFF模型是文献[23]提出的基于自注意力的深度特征融合方法,用于融合遥感图像的深层特征;Siamese ResNet_50模型是文献[12]提出的用于遥感场景分类的孪生卷积神经网络;CILM模型是文献[14]提出的一种互补信息学习模型。

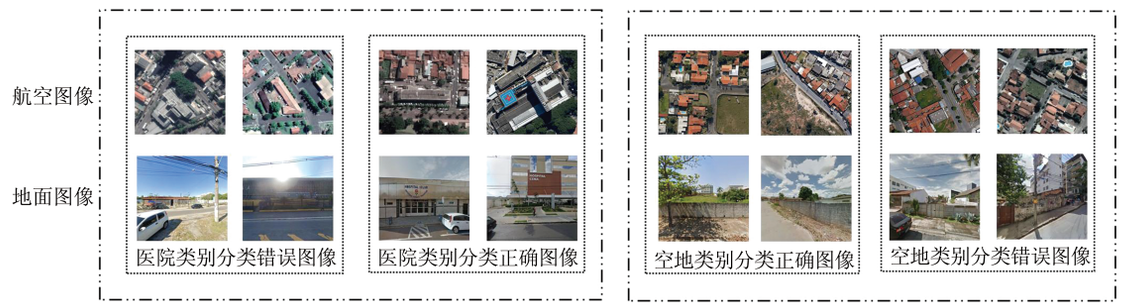

图7　CV-BrCT数据集医院和空地部分分类结果

Fig. 7　Classification Results of Hospital and Open Space in CV-BrCT Dataset

表4　多视角分类消融实验结果/%

Tab. 4　Results of Multi-view Classification Ablation Experiments/%

CNN网络分支	权值	方法	AiRound数据集	CV-BrCT数据集
AlexNet	共享权值	*M*1	88.41	79.70
		*M*2	89.27	82.27
		*M*3	90.12	83.25
	不共享权值	*M*1	89.69	82.27
		*M*2	90.12	84.60
		*M*3	92.27	84.91
VGG16	共享权值	*M*1	88.84	80.96
		*M*2	90.12	82.56
		*M*3	91.41	83.45
	不共享权值	*M*1	90.12	82.56
		*M*2	90.55	84.62
		*M*3	93.13	85.18

表5　多尺度注意力网络与其他方法对比结果/%

Tab. 5　Comparison Results of Different Methods/%

网络模型	CNN网络分支	AiRound数据集	CV-BrCT数据集
NBCL[22]		81.12	80.29
VGG_VD16+SAFF[23]		82.83	74.74
Siamese ResNet_50[12]		87.55	80.80
微调CNN	AlexNet	90.12	84.60
	VGG16	90.56	84.62
预训练CNN	AlexNet	89.27	74.95
	VGG16	90.55	74.31
CILM[14]	AlexNet	90.55	84.79
	VGG16	91.41	84.86
MSAN	AlexNet	92.27	84.91
MSAN	VGG16	93.13	85.18

由表5可知,在6种对比分类方法中,对于AiRound数据集,NBCL模型用于多视角遥感场景分类精度为81.12%,分类精度最低;CILM的VGG16模型分类精度最高,达到91.41%;本文的MSAN-AlexNet模型精度为92.27%,MSAN-VGG16模型精度达到93.13%,相对于CILM模型精度提高1.72%。此外,在AiRound数据集中,MSAN-VGG16模型对于桥、湖泊、雕像、塔和公园5个类别的分类精度有明显提升,其中MSAN-VGG16模型对于塔的分类精度达到100%,这是其他6种分类方法无法达到的,而对于最易混淆的湖泊类别,CILM模型、MSAN-VGG16模型的精度分别为33.33%、55.56%,MSAN-VGG16模型比CILM模型的精度提升了22.23%。对于CV-BrCT数据集,预训练CNN方法分类精度最低,为74.31%;CILM的基于VGG16模型的分类精度最高,为84.86%;MSAN-AlexNet方法精度为84.91%,MSAN-VGG16方法的精度为85.18%,相对于CILM模型精度提高了0.32%。此外,在CV-BrCT数据集中,MSAN-VGG16模型对于公寓、医院、停车场、空地和商店5个类别的分类精度有明显提升,而对于最容易混淆的医院类别,CILM模型的精度为17.11%,MSAN-VGG16模型的分类精度为21.05%,相比CILM模型提升了3.94%。本文所提MSAN相较于其他6种方法性能较优,原因主要是本文所设计的注意力特征融合模块能够对卷积神经网络所提取的多层次特征进行筛选和融合,深度探究有利于多视角场景分类的图像特征信息,实现多尺度特征学习的同时提高多视角遥感场景分类的精度。

3 结 语

本文提出一种融合多尺度注意力的卷积神经网络用于多视角场景分类,与传统的场景分类方法不同,本文方法结合航空图像和地面图像之间的互补信息进行多视角场景分类,能获得具有更强鉴别能力的特征。网络包括两个分支,分别输入影像对中的航空图像和地面图像,经过特征融合模块实现多尺度特征提取。利用训练完成的多尺度注意力网络分别提取多视角(航空和地面)图像的特征,并将两个视角的特征融合后进行SVM分类。在两个公开数据集上的实验表明,本文提出的多尺度注意力网络在AiRound数据集和CV-BrCT数据集中最高精度分别达到93.13%和85.18%,改善了遥感影像场景分类结果。未来的研究工作将从网络特征融合方面进行优化,以适应更复杂的遥感场景,进一步提升遥感场景分类性能。

参 考 文 献

[1] Li E Z, Xia J S, Du P J, et al. Integrating Multilayer Features of Convolutional Neural Networks for Remote Sensing Scene Classification [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2017, 55(10): 5653-5665.

[2] Bian X Y, Chen C, Tian L, et al. Fusing Local and Global Features for High-Resolution Scene Classification[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2017, 10(6): 2889-2901.

[3] Du Peijun, Zhang Peng, Guo Shanchuan, et al. A Semantic Segmentation Model for Mapping Plastic Greenhouse Based on Spectral Index and High-Resolution Imagery [J]. *Geomatics and Information Science of Wuhan University*, 2023, 48(10): 1670-1683. [3](杜培军, 张鹏, 郭山川, 等. 融合光谱指数与高分影像的塑料大棚语义分割模型[J]. 武汉大学学报(信息科学版), 2023, 48(10): 1670-1683.)

[4] Chaib S, Liu H, Gu Y F, et al. Deep Feature Fusion for VHR Remote Sensing Scene Classification[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2017, 55(8): 4775-4784.

[5] Franklin S E, Hall R J, Moskal L M, et al. Incorporating Texture into Classification of Forest Species Composition from Airborne Multispectral Images [J]. *International Journal of Remote Sensing*, 2000, 21(1): 61-79.

[6] Zhang L P, Huang X, Huang B, et al. A Pixel Shape Index Coupled with Spectral Information for Classification of High Spatial Resolution Remotely Sensed Imagery [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2006, 44(10): 2950-2961.

[7] Krizhevsky A, Sutskever I, Hinton G E. ImageNet Classification with Deep Convolutional Neural Networks[C]// The 25th International Conference on Neural Information Processing Systems, Lake Tahoe, Nevada, 2012.

[8] Men Jilin, Liu Yueyan, Zhang Bin, et al. Land Use Classification Based on Multi-structure Convolution Neural Network Features Cascading[J]. *Geomatics and Information Science of Wuhan University*, 2019, 44(12): 1841-1848. (门计林, 刘越岩, 张斌, 等. 多结构卷积神经网络特征级联的高分影像土地利用分类[J]. 武汉大学学报(信息科学版), 2019, 44(12): 1841-1848.)

[9] Gong Jianya, Zhang Zhan, Jia Haowei, et al. Multi-source Data Ground Object Extraction Based on Knowledge-aware and Multi-scale Feature Fusion Network[J]. *Geomatics and Information Science of Wuhan University*, 2022, 47(10): 1546-1554. (龚健雅, 张展, 贾浩巍, 等. 面向多源数据地物提取的遥感知识感知与多尺度特征融合网络[J]. 武汉大学学报(信息科学版), 2022, 47(10): 1546-1554.)

[10] Zhang Yuan, Wang Dong, Wang Xiaohua, et al. Urban Building Change Detection Using Multi-scale Siamese Atrous Convolutional Neural Network[J]. *Journal of Geomatics*, 2023, 48(4): 30-34. (张缘, 王冬, 王晓华, 等. 多尺度空洞卷积网络城市建筑物变化检测应用[J]. 测绘地理信息, 2023, 48(4): 30-34.)

[11] Zheng Zhuo, Fang Fang, Liu Yuanyuan, et al. Joint Multi-scale Convolution Neural Network for Scene Classification of High Resolution Remote Sensing Imagery [J]. *Acta Geodaetica et Cartographica Sinica*, 2018, 47(5): 620-630. (郑卓, 方芳, 刘袁缘, 等. 高分辨率遥感影像场景的多尺度神经网络分类法[J]. 测绘学报, 2018, 47(5): 620-630.)

[12] Liu X N, Zhou Y, Zhao J Q, et al. Siamese Convolutional Neural Networks for Remote Sensing Scene Classification [J]. *IEEE Geoscience and Remote Sensing Letters*, 2019, 16(8): 1200-1204.

[13] Machado G, Ferreira E, Nogueira K, et al. AiRound and CV-BrCT: Novel Multi-view Datasets for Scene Classification[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote*

Sensing, 2021, 14: 488-503.

[14] Geng W X, Zhou W X, Jin S G. Multi-view Urban Scene Classification with a Complementary-Information Learning Model[J]. *Photogrammetric Engineering & Remote Sensing*, 2022, 88(1): 65-72.

[15] Hong D F, Hu J L, Yao J, et al. Multimodal Remote Sensing Benchmark Datasets for Land Cover Classification with a Shared and Specific Feature Learning Model[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2021, 178: 68-80.

[16] Sun Y X, Feng S S, Ye Y M, et al. Multisensor Fusion and Explicit Semantic Preserving-Based Deep Hashing for Cross-Modal Remote Sensing Image Retrieval[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2021, 60: 5219614.

[17] Zhu S H, Du B, Zhang L P, et al. Attention-Based Multiscale Residual Adaptation Network for Cross-scene Classification[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 1809, 60: 5400715.

[18] Gong T F, Zheng X T, Lu X Q. Cross-Domain Scene Classification by Integrating Multiple Incomplete Sources[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2021, 59(12): 10035-10046.

[19] Lü Yafei, Xiong Wei, Zhang Xiaohan. A General Cross-Modal Correlation Learning Method for Remote Sensing[J]. *Geomatics and Information Science of Wuhan University*, 2022, 47(11): 1887-1895. (吕亚飞, 熊伟, 张筱晗. 一种通用的跨模态遥感信息关联学习方法[J]. 武汉大学学报(信息科学版), 2022, 47(11): 1887-1895.)

[20] Lu X Q, Sun H, Zheng X T. A Feature Aggregation Convolutional Neural Network for Remote Sensing Scene Classification[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2019, 57(10): 7894-7906.

[21] Hu J, Shen L, Sun G. Squeeze and Excitation Networks[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018.

[22] Muhammad U, Hoque M Z, Wang W Q, et al. Patch-Based Discriminative Learning for Remote Sensing Scene Classification[J]. *Remote Sensing*, 2022, 14(23): 5913.

[23] Cao R, Fang L Y, Lu T, et al. Self-Attention-Based Deep Feature Fusion for Remote Sensing Scene Classification[J]. *IEEE Geoscience and Remote Sensing Letters*, 2021, 18(1): 43-47.

[24] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-scale Image Recognition[EB/OL]. [2014-12-20]. https://arxiv.org/abs/1409.1556.pdf.

[25] Xia G S, Hu J W, Hu F, et al. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2017, 55(7): 3965-3981.