引用格式：WANG Jianan, GAO Yue, SHI Jun, et al. Scene Classification of Optical High-resolution Remote Sensing Images Using Vision Transformer and Graph Convolutional Network[J]. Acta Photonica Sinica, 2021, 50(11):1128002
王嘉楠,高越,史骏,等. 基于视觉转换器和图卷积网络的光学遥感场景分类[J]. 光子学报,2021,50(11):1128002

基于视觉转换器和图卷积网络的光学遥感场景分类

王嘉楠[1],高越[1],史骏[2],刘子琦[1]
(1 航天恒星科技有限公司, 北京 100095)
(2 合肥工业大学 软件学院, 合肥 230601)

摘　要:当前基于卷积神经网络的光学遥感图像场景分类方法大多是全局特征学习,忽略了场景局部特征,从而难以较好地解决类内差异大和类间相似性高的问题,因此,提出一种基于视觉转换器和图卷积网络双分支结构的光学遥感图像场景分类方法。该方法首先对场景图像进行分块,再利用位置编码和视觉转换器进行特征编码,从而挖掘图像内部的长距离依赖关系。另一方面,对遥感图像进行超像素分割,将每个超像素对应的卷积神经网络特征进行池化处理并作为图结构中的结点,利用图卷积网络对场景内部图结构进行建模,感知场景内部的空间拓扑关系。最终融合两个分支产生的特征形成场景内容的最终特征表示并用于分类。在光学遥感图像数据集上的实验验证了所提方法在遥感场景分类中的有效性。

关键词:遥感;场景分类;卷积神经网络;视觉转换器;图卷积网络
中图分类号:TP391　　**文献标识码**:A　　**doi**:10.3788/gzxb20215011.1128002

Scene Classification of Optical High-resolution Remote Sensing Images Using Vision Transformer and Graph Convolutional Network

WANG Jianan[1], GAO Yue[1], SHI Jun[2], LIU Ziqi[1]
(1 *Space Star Technology Co., Ltd., Beijing* 100095, *China*)
(2 *School of Software, Hefei University of Technology, Hefei* 230601, *China*)

Abstract: Most existing optical remote sensing scene classification methods based on convolutional neural network mainly perform global feature learning and fail to consider the local features in the scene, which cannot effectively address the large intraclass difference and high interclass similarity. Therefore, a novel remote sensing scene classification method based on two branches of vision transformer and graph convolution network is proposed. Firstly the scene image is divided into patches and the then positional encoding and vision transformer are used to encode the patches. Consequently, the long-range dependencies can be mined. On the other hand, the scene image is converted into superpixels. The convolutional neural networks features of each superpixel are pooled and used to represent the node of the graph structure. Then the graph convolutional network is applied to model the spatial topology relationships. Finally the final feature representation of the scene image are described by the features of the two branches. Experimental results on the optical remote sensing image datasets demonstrate the effectiveness of our method.

基金项目:安徽省自然科学基金(No. 1908085MF210)
第一作者:王嘉楠(1986—),女,工程师,硕士,主要研究方向为人工智能与遥感应用. Email: htwangjn@163.com
通讯作者:高越(1987—),男,高级工程师,硕士,主要研究方向为人工智能与遥感应用. Email: bjlguniversity@163.com
收稿日期:2021-05-25;**录用日期**:2021-07-26

Key words: Remote sensing; Scene classification; convolutional neural network; Vision transformer; graph convolutional network
OCIS Codes: 280.4788; 100.2960; 100.4996; 100.3008

0 引言

作为当前遥感对地观测技术领域的研究热点之一,高分辨率光学遥感图像场景分类旨在根据遥感场景图像内容将图像自动分类为一个特定的语义标签,为图像理解提供辅助参考。传统的场景分类方法通常使用低层或人工特征描述场景内容,难以表征高级语义信息,从而无法较好地应对遥感场景类内差异大和类间相似性高的问题。

近年来,卷积神经网络(Convolutional Neural Networks,CNN)[1-7]受到广泛关注并被成功地应用于遥感图像场景分类。ZOU Qin等[8]将深度信念网络(Deep Belief Network,DBN)应用于遥感场景分类。XIA Guisong等[9]提出了AID航空影像数据集,验证了CNN在场景分类中的有效性。CHENG Gong等[10]提出了NWPU-RESISC45遥感图像场景数据集,比较了AlexNet[1]、VGG[2]和GoogLeNet[3]等经典CNN方法在该数据集上的分类性能。YU Yunlong等[11]构建了双分支深度特征融合框架,使用纹理和显着性深度学习结构对场景分类建模。HE Nanjun等[12]提出了多层堆叠协方差池化方法(Multilayer Stacked Covariance Pooling, MSCP),使用堆叠CNN特征图的协方差矩阵描述场景并实现分类。这些方法大多只是直接使用经典的CNN模型或融合来自不同层的特征,没有考虑场景的尺度变化。为此,BIAN Xiaoyong等[13]提出了多尺度多层的高斯编码(Multi-Scale Multi-Layer based Gaussian Coding, mSmL-Gcoding)并用于描述高维多尺度CNN特征分布。ZHANG Jun等[14]提出了多尺度深度特征表示(Multi-scale Deep Feature Representation, MDFR)方法,使用区域协方差细化CNN特征,再实现多尺度深度特征融合。另一方面,遥感场景图像的局部信息对于描述场景内容至关重要。为此,注意力机制[15-17]被应用于遥感场景分类,旨在特征学习的过程中关注场景内的局部显著区域。CAO Ran等[18]使用了基于自注意力机制的深度特征融合方法(Self-Attention-based Deep Feature Fusion, SAFF),从空间和通道两个维度细化CNN特征,提升了模型对场景内显著目标以及场景特征图(feature map)依赖关系的感知能力。边小勇等[19]提出了多尺度特征变换和注意力机制相融合的尺度注意力网络模型。尽管这些方法利用注意力机制增强了CNN特征表示能力,但是忽略了场景内部的长距离依赖关系以及潜在的空间拓扑关系,这些关系能够增强场景的语义表示能力,从而提高场景分类的准确性。

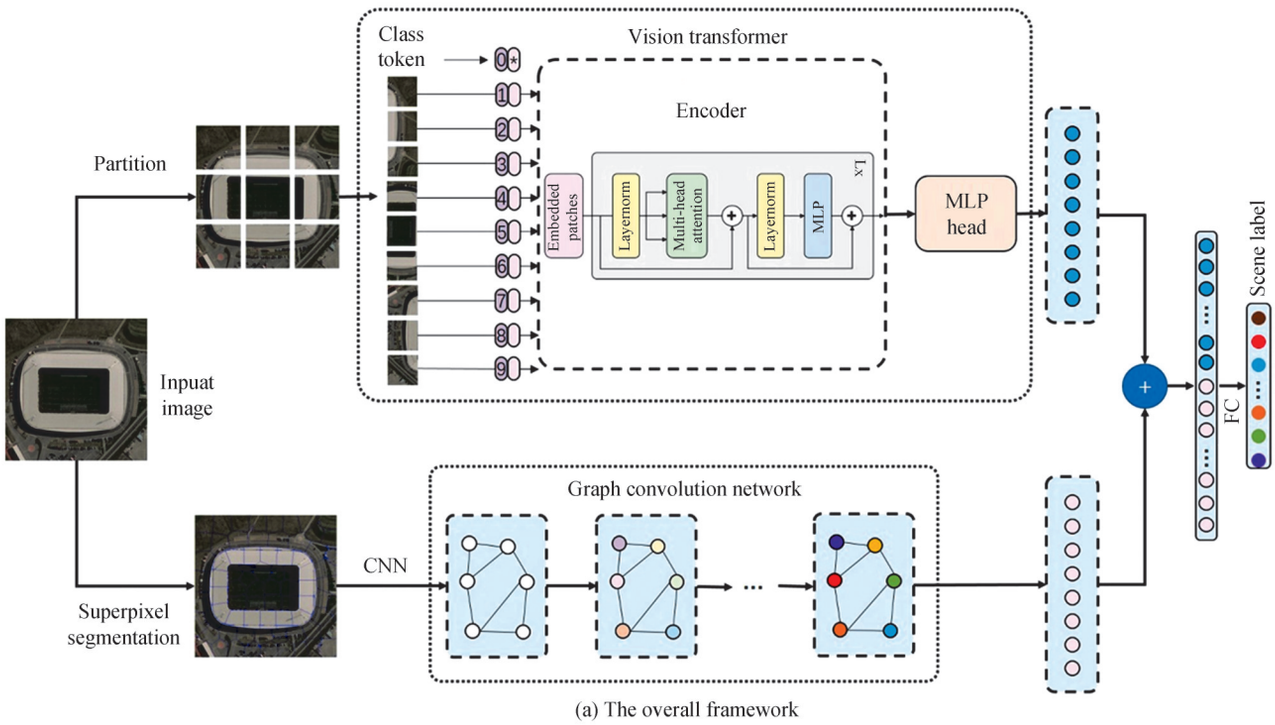
本文提出基于视觉转换器和图卷积网络双分支结构的遥感图像场景分类方法。首先使用视觉转换器(Vision Transformer, ViT)[20]对场景内部的长距离依赖关系进行挖掘,形成基于Transformer结构的遥感图像特征表示;其次,使用图卷积网络(Graph Convolutional Network, GCN)[21]对场景内部的空间拓扑关系进行建模,建立具有空间关系感知的遥感图像特征表示;最后,融合ViT和GCN双分支的特征表示,形成长距离依赖关系和空间拓扑关系融合感知的特征表示,增强整个遥感场景图像的特征表示能力。

1 本文方法

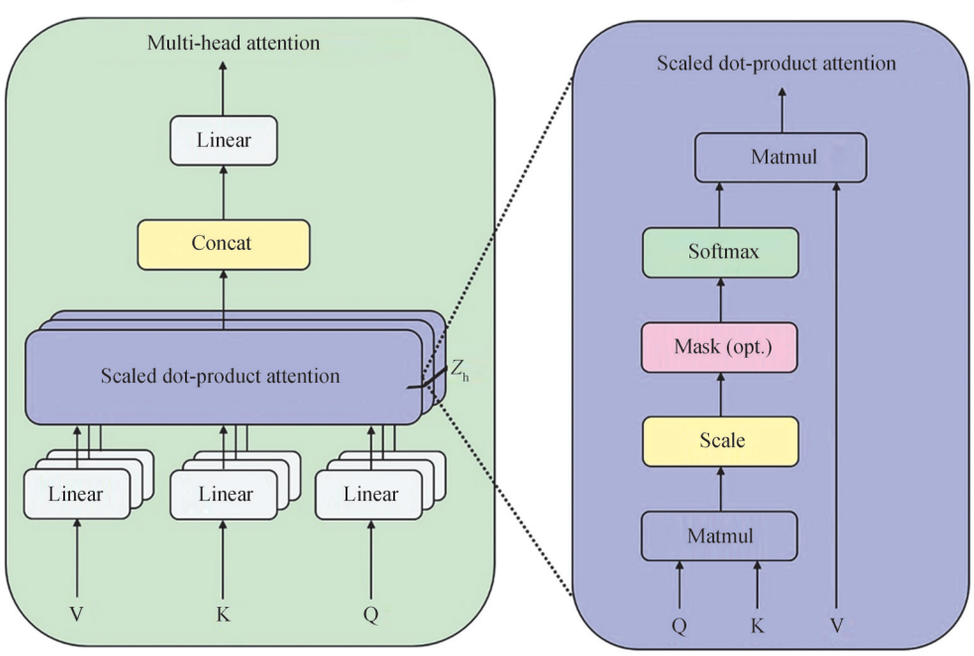
1.1 概述

提出的遥感场景分类方法如图1(a)所示,整个遥感场景图像处理由两个分支构成,即基于视觉转换器(ViT)的处理分支和基于图卷积网络(GCN)的处理分支。基于ViT的分支首先将图像分块,对分块图像及其对应的空间位置进行嵌入,利用L个ViT编码器(Encoder)对嵌入的分块图像进行处理,每个编码器由层归一化(LN)、多头注意力机制(MSA)以及多层感知器(MLP)构成,最后一个编码器输出的特征即为场景图像对应的ViT特征。另一方面,基于GCN的处理分支首先对场景进行超像素分割,同时利用预训练的CNN模型对该图像提取特征,再通过最大池化处理每个超像素的CNN特征,由此得到每个超像素所对应的CNN特征向量,将这些特征向量作为图结构中的结点,结点之间的边描述了结点间的相似关系,反映了超像素之间的空间拓扑关系。每个结点的特征和对应的邻接关系被送入一个两层的GCN模型,利用其消息传递机制获取各结点基于空间拓扑关系的GCN特征,最后计算各结点的GCN特征均值作为整个场景的GCN特征。由此,ViT特征和GCN特征通过串联融合的方式融合成一个特征向量,作为整个场景的特征表示。

最后,将其送入全连接层并通过交叉熵损失函数训练整个网络。



(a) The overall framework



(b) Multi-head attention

图1　所提方法处理流程

Fig. 1　The pipeline of the proposed method

1.2　基于视觉转换器的特征提取

给定一组包含N个遥感图像的集合$\boldsymbol{S}=\left\{\boldsymbol{X}_i,\boldsymbol{y}_i\right\}_{i=1}^{N}$,其中$\boldsymbol{X}_i\in\mathbb{R}^{C\times H\times W}$表示第$i$幅场景图像,$C$、$H$和$W$分别表示图像的通道数、高度和宽度,$\boldsymbol{y}_i$表示场景图像对应的类别。首先对每幅图像进行分块,则该图像可表示为一个包含m个图像块的序列$\left(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_m\right)$,每个图像块$\boldsymbol{x}_i\in\mathbb{R}^{C\times p\times p}$,其中$p$表示每个图像块的维度,且$m=HW/p^2$。ViT首先将图像块序列进行线性嵌入,利用可学习的嵌入矩阵$\boldsymbol{E}$将图像块投影成一个D维嵌入表示,该D维向量连带一个可以学习的分类标记$\boldsymbol{x}_{\text{class}}$被串联成最终的嵌入图像块序列,即[20]

$$\boldsymbol{z}_0=\left[\boldsymbol{x}_{\text{class}};\ \boldsymbol{x}_1\boldsymbol{E};\ \boldsymbol{x}_2\boldsymbol{E};\ \cdots;\ \boldsymbol{x}_m\boldsymbol{E}\right]+\boldsymbol{E}_{\text{pos}},\boldsymbol{E}\in\mathbb{R}^{\left(p^2C\right)\times D},\boldsymbol{E}_{\text{pos}}\in\mathbb{R}^{(m+1)\times D}\tag{1}$$

式中，$\boldsymbol{E}_{\text{pos}}$用于表示各图像块在原始图像中的空间位置并被编码到嵌入表示中。经过嵌入表示后的图像块被送入ViT的编码器(Encoder)中，编码器由层归一化(LayerNorm)、多头注意力机制(Multiheaded Self-Attention，MSA)以及多层感知器(Multilayer Perceptron，MLP)顺序处理[17]，如式(2)和(3)所示。其中层归一化用在每个编码器模块前和残差连接后，多层感知器为2层且其激活函数为GELU函数。对于最后一个编码器的输出$\boldsymbol{z}_L$，对其$m+1$个特征求均值，从而作为场景图像对应的ViT特征表示。

$$\boldsymbol{z}_l'=\text{MSA}\left(\text{LN}\left(\boldsymbol{z}_{l-1}\right)\right)+\boldsymbol{z}_{l-1}\quad l=1,\cdots,L\tag{2}$$

$$\boldsymbol{z}_l=\text{MLP}\left(\text{LN}\left(\boldsymbol{z}_l'\right)\right)+\boldsymbol{z}_l'\quad l=1,\cdots,L\tag{3}$$

如图1(b)和(c)所示，多头注意力MSA可以表示为

$$\text{MSA}\left(\boldsymbol{z}\right)=\text{Concat}\left(\boldsymbol{H}_1,\boldsymbol{H}_2,\cdots,\boldsymbol{H}_h\right)\boldsymbol{W}^o\tag{4}$$

式中，$\boldsymbol{H}_i\left(i=1,\cdots,h\right)$可表示为

$$\boldsymbol{H}_i=\text{Attention}\left(Q,K,V\right)=\text{Attention}\left(\boldsymbol{z}\boldsymbol{W}_i^Q,\boldsymbol{z}\boldsymbol{W}_i^K,\boldsymbol{z}\boldsymbol{W}_i^V\right)=\text{softmax}\left(\frac{\left(\boldsymbol{z}\boldsymbol{W}_i^Q\right)\left(\boldsymbol{z}\boldsymbol{W}_i^K\right)^{\text{T}}}{\sqrt{d_k}}\right)\left(\boldsymbol{z}\boldsymbol{W}_i^V\right)\tag{5}$$

式中，$\boldsymbol{W}_i^Q,\boldsymbol{W}_i^K\in\mathbb{R}^{D\times d_k}$，$\boldsymbol{W}_i^V\in\mathbb{R}^{D\times d_v}$，$\boldsymbol{W}^o\in\mathbb{R}^{hd_v\times D}$，且$d_k=d_v=D/h$，$\text{Concat}\left(\cdot\right)$表示以列向量堆叠，$\text{Attention}\left(\cdot\right)$表示多头注意力处理。

1.3 基于图卷积网络的特征提取

由于遥感场景图像内容丰富，如图1中的遥感场景图像所示，场景中除了球场本身，还有周围的草地，球场和草地之间存在一定的空间拓扑关系。如果将场景分解成多个局部处理单元，通过挖掘这些处理单元的局部空间拓扑关系，有助于提升场景整体特征表示的鉴别能力。因此，针对遥感场景图像潜在的空间拓扑关系，利用SLIC算法[22]对场景图像进行过分割，产生场景内的超像素，这些超像素可以视为场景内的局部处理单元。同时，使用预训练的残差网络ResNet[4]对整个图像提取CNN特征，在生成最后一组特征图(feature map)后，根据之前得到的超像素，利用最大池化(max pooling)方式对每个超像素的CNN特征进行处理，从而得到每个超像素所对应的CNN特征表示，将这些超像素的CNN特征作为图结构的结点，记为$\boldsymbol{B}\in\mathbb{R}^{s\times t}$，其中$s$表示超像素的数目，$t$为超像素对应的CNN特征维数，由此构造邻接矩阵$\boldsymbol{A}\in\mathbb{R}^{s\times s}$为

$$\boldsymbol{A}_{ij}=\begin{cases}1 & \text{if } \boldsymbol{B}_i\text{belongs to Spatial } 4-\text{neighborhood of } \boldsymbol{B}_j \text{ or } \boldsymbol{B}_j\text{belongs to Spatial } 4-\text{neighborhood of } \boldsymbol{B}_i\\ 0 & \text{otherwise}\end{cases}\tag{6}$$

对各个超像素$\boldsymbol{B}_i\left(i=1,\cdots,s\right)$建立空间4邻域关系，实现了对场景内部空间拓扑关系的构造，由此，矩阵$\boldsymbol{A}$描述了各超像素间的空间拓扑结构。根据图卷积网络的分层传播规则[21]，将所构造的空间拓扑关系和各超像素的CNN特征进行消息传递，将空间拓扑关系嵌入至CNN特征，获取对应的关系感知表示，即

$$\boldsymbol{H}^{(l+1)}=\sigma\left(\hat{\boldsymbol{A}}\boldsymbol{H}^{(l)}\boldsymbol{W}^{(l)}\right)\tag{7}$$

式中，$\boldsymbol{H}^{(l)}\in\mathbb{R}^{s\times d_l}$是第$l$层结点的特征，是第$l$层特征的维数；$\boldsymbol{H}^{(l+1)}\in\mathbb{R}^{N\times d_{l+1}}$表示结点更新后的特征。$\boldsymbol{H}^{(0)}=\boldsymbol{X}$即各超像素的CNN特征和$\hat{\boldsymbol{A}}\in\mathbb{R}^{s\times s}$是具有自连接的规范化邻接矩阵，即$\hat{\boldsymbol{A}}=\boldsymbol{D}^{-\frac{1}{2}}\left(\boldsymbol{A}+\boldsymbol{I}\right)\boldsymbol{D}^{-\frac{1}{2}}$，$\boldsymbol{I}$是单位矩阵。对角度矩阵$\boldsymbol{D}$为$\boldsymbol{D}_{ii}=\sum_j\boldsymbol{A}_{ij}$；$\boldsymbol{W}^{(l)}\in\mathbb{R}^{d_l\times d_{l+1}}$是可训练权重；$\sigma\left(\cdot\right)$是激活函数。由此，通过GCN的消息传递机制，实现了基于空间拓扑关系的特征嵌入。

2 实验结果与分析

2.1 数据集

实验使用了国际公开的AID[9]和NWPU-RESISC45[10]遥感场景图像数据集用于评估所提方法的遥感场景分类性能。AID数据集包含来自30个场景类别的10 000张图像。每个图像的尺寸为600像素×600像素。每个类别中的图像数量变化从220到420，空间分辨率变化从0.5 m到8 m。此外，每个类别的图像都是在不同的成像条件下采集的，导致类内差异很大。NWPU-RESISC45数据集包含45个场景类别，每个场景

类别包含700个场景图像，图像大小为256像素×256像素。空间分辨率变化从0.2 m到30 m。与AID相比，NWPU-RESISC45数据集具有大规模、图像丰富、类内多样性高和类间相似度高等显着特征，从而增加了场景分类的难度。

2.2 实验设置

为了公平比较，根据文献[9-10]为每个数据集设置了相同的训练测试比率，即从AID数据集中每个类别的图像中随机选择20%进行训练，其余80%作为测试。此外，还将AID的训练测试比率设置为50%：50%。对于NWPU-RESISC45数据集，分别将图像的训练比率设置为10%、20%，将其余90%、80%也分别进行测试。实验采用总体准确率（Overall Accuracy，OA）和标准差（Standard Deviation，Std）定量评估分类结果。OA可以看作是正确分类的图像与所有预测图像的比率，Std用于度量OA的变化程度。对每个训练集随机重复实验10次，并将测试集上分类结果的总体准确率和标准差作为比较的最终结果。

本文使用PyTorch来实现算法流程，在ViT分支上将图像大小调整为224×224，并将图像分成16×16图像块，多头注意力模块（MSA）的头数Head设置为12，编码器（Encoder）数量L设置为12，使用ViT在ImageNet-21k上的预训练权重ViT-Base[20]进行训练。基于GCN的特征提取部分，使用ResNet-50先提取场景图像的CNN特征，再构造两层GCN模型，每层的特征维数设置为256，将ReLu用作GCN中的激活函数。整个实验的运行环境为Intel Core i7-9700X CPU（3.60 GHz）以及NVIDIA GTX 2080Ti GPU。

2.3 分类性能的比较

对AID和NWPU-RESISC45数据集进行遥感场景分类的实验，并将所提出的方法与代表性的场景分类方法进行了比较，包括CaffeNet[9-10]、VGG-16[9-10]、GoogLeNet[9-10]、双分支深度特征融合框架（two-stream deep feature fusion）[11]、多层堆叠协方差池化方法（MSCP）[12]、基于多尺度多层的高斯编码（mSmL-Gcoding）[13]、尺度深度特征表示（MDFR）方法[14]、基于自注意力机制的深度特征融合方法（SAFF）[18]以及尺度注意力网络方法（scale-attention network）[19]。

表1列出了所有方法在AID数据集上的实验结果。可以看出作为特征融合的分类，双分支深度特征融合方法[11]和MSCP[12]产生的分类性能在训练比例分别为20%和50%的条件下明显优于基本的CaffeNet[9]、VGG-16[9]、GoogLeNet[9]。对于基于多尺度的方法，如mSmL-Gcoding[13]和MDFR[14]，由于使用了多个尺度的特征表示，产生了较好的分类性能。另一方面，基于注意的方法，如SAFF[18]和尺度注意力网络方法[19]，由于侧重于场景中有助于分类的局部辨别部分，因此具有相对稳定的性能。对比上述方法，本文方法具有最高的总体准确率，分别为94.52%和96.80%，表明该方法通过融合视觉转换器和图卷积网络的特征，挖掘了图像内部的长距离依赖关系和场景内部的空间拓扑关系，获取了更具有鉴别性能的特征表示。

表1 AID数据集中20%和50%训练比例下不同方法的标准差和总体准确率

Table 1 Stds and overall accuracies of different methods with 20% and 50% training ratio in the AID dataset

Method	20% training ratio/%	50% training ratio/%
CaffeNet[9]	86.86±0.47	89.53±0.31
VGG－VD－16[9]	86.59±0.29	89.64±0.36
GoogLeNet[9]	83.44±0.40	86.39±0.55
Two－stream deep feature fusion[11]	94.09±0.34	95.99±0.35
MSCP[12]	91.52±0.21	94.42±0.17
mSmL－Gcoding[13]	91.69±0.36	95.61±0.28
MDFR[14]	90.62±0.27	93.37±0.29
SAFF[18]	90.25±0.29	93.83±0.28
Scale－attention network[19]	92.53±0.33	95.72±0.27
Proposed method	**94.52±0.25**	**96.80±0.17**

与AID数据集相比，NWPU-RESISC45数据集具有数据量大、场景变化大、类内多样性和类间相似性高等特点，对遥感场景分类提出了更高的挑战。表2显示了不同训练比例下不同方法在NWPU-RESISC45上的总体分类精度和标准差。显然，基于融合的方法、基于多尺度的方法和基于注意的方法都优于经典的

CaffeNet[10]、VGG-16[10]和GoogLeNet[10]。值得注意的是，基于特征融合的方法，双分支深度特征融合[11]和MSCP[12]产的分类性能不如其在AID数据集上的表现，这很大程度上是由于NWPU-RESISC45数据集较大的图像变化所引起的。对于基于多尺度的方法，MDFR[14]不稳定，而mSmL-Gcoding[13]由于嵌入了高维CNN特征分布，显示出更好的分类性能。作为基于注意力机制的方法，尺度注意力网络方法[19]由于融合了多尺度特征变换和注意力机制，其分类性能优于SAFF[18]。对比以上方法，本文方法取得了较优的识别性能，在训练比例为10%和20%的情况下，分别取得了90.50%和93.31%的识别率。说明本文方法将视觉转换器和图卷积网络相结合，充分考虑了图像内部的局部特征和空间拓扑关系，增强了场景高层语义特征的表示能力，因此取得了较好的分类识别能力。

表2 NWPU-RESISC45数据集中10%和20%训练比例下不同方法的标准差和总体准确率

Table 2 Stds and overall accuracies of different methods with 10% and 20% training ratio in the NWPU-RESISC45 dataset

Method	10% training ratio/%	20% training ratio/%
CaffeNet[10]	81.22±0.19	85.16±0.18
VGG-VD-16[10]	87.15±0.45	90.36±0.18
GoogLeNet[10]	82.57±0.12	86.02±0.18
Two-stream deep feature fusion[11]	85.02±0.25	87.01±0.19
MSCP[12]	85.33±0.17	88.93±0.14
mSmL-Gcoding[13]	89.34±0.48	91.64±0.26
MDFR[14]	83.37±0.26	86.89±0.17
SAFF[18]	84.38±0.19	87.86±0.14
Scale-attention network[19]	88.92±0.29	92.25±0.18
Proposed method	**90.50±0.26**	**93.31±0.15**

2.4 消融实验

所提方法通过ViT和GCN两个处理分支，实现了遥感场景图像内部长距离依赖关系和空间拓扑关系的建模，并将这些关系嵌入到场景图像的特征表示中。对这两个处理分支进行消融实验，比较所提方法及其在不用ViT（w/o ViT）和不用GCN（w/o GCN）的情况下在AID和NWPU-RESISC45数据集上的总体分类准确率和标准差，如表3所示。可以看到，proposed method w/o GCN优于proposed method w/o ViT，表明与GCN单独分支结构相比，transformer分支结构产生了较好的分类识别性能，同时对比表1和表2中基于CNN的场景分类方法，proposed method w/o GCN也具有可比的分类效果。这说明transformer学习长依赖的能力在遥感场景分类中展现出了良好的应用前景。此外，通过融合ViT和GCN两个分支的特征表示，所提方法具有较优的分类准确率。

表3 AID和NWPU-RESISC45数据集上所提方法的消融实验对比

Table 3 Ablation experiments of the proposed method in the AID and NWPU-RESISC45 datasets

Method	AID	NWPU-RESISC45
	50% training ratio/%	20% training ratio/%
Proposed method w/o ViT	89.90±0.50	88.10±0.21
Proposed method w/o GCN	94.20±0.22	91.70±0.18
Proposed method	**96.80±0.17**	**93.31±0.15**

2.5 GID数据集的遥感场景分类实验

为了验证所提方法对于真实遥感场景图像分类的有效性，对比了DenseNet-121、所提方法及其在不用ViT和不用GCN情况下在GID[23]数据集上的分类性能。GID数据集来源于我国高分2号卫星，空间分辨率为4 m，包含了5个场景大类（建筑物、耕地、森林、草地和水体）共150张图像，大小为6 800×7 200。为方便场景细粒度分类研究，该数据集在5个大类的基础上进一步分为15个场景子类（稻田、灌溉地、旱耕地、花园地、乔木林地、灌木林地、天然草地、人工草地、工业用地、城市住区、农村住区、交通用地、河流、湖泊和池塘）每类2 000张图像，图像大小包含3个尺度：56×56、112×112、224×224。针对GID数据集15个场景子类进

行实验，选取每类的20%用于训练，每类剩下的80%用于测试。表4列出了三种方法在GID数据集上的分类性能。可以发现，仅使用transformer结构的proposed method w/o GCN优于代表性的CNN分类方法DenseNet。而仅使用GCN分支的proposed method w/o ViT分类准确率低于DenseNet，这可能是由于GID数据集较AID和NWPU-RESISC45两个数据集类内变化大，类间差异小，导致过分割后所构造的图结构不能很好地反映场景内的局部拓扑关系，从而影响识别性能。而通过ViT和GCN两个分支的融合，所提方法依然优于其他方法，表明遥感场景图像内部长距离依赖关系和空间拓扑关系的联合建模有助于增强整个场景特征表示的鉴别能力。

表4　GID数据集20%训练比例下不同方法的总体准确率

Table 4　Overall accuracies of different methods with 20% training ratio in the GID dataset

Method	DenseNet-121	Proposed method w/o ViT	Proposed method w/o GCN	Proposed method
OA/%	69.10	68.50	70.20	**72.30**

2.6　分类效率的测试

遥感场景图像分类应用中，分类效率也是需要考虑的重要因素。对代表性的CNN分类方法DenseNet和所提方法在AID、NWPU-RESISC45以及GID三个数据集上的每秒帧率（Frame Per Second，FPS）进行了实验对比，如表5所示。可以看到，所提方法在三个数据集上的FPS明显优于DenseNet-121，这说明相比较基于CNN的方法，所提方法使用了transformer结构和GCN结构的分支，计算效率得到明显的提升，同时也表明transformer在计算效率和分类效果上均具有较好的应用前景。

表5　AID、NWPU-RESISC45和AID数据集上所提方法的效率比较

Table 5　FPS comparison of proposed method in the AID，NWPU-RESISC45 and GID datasets

Method	AID 50% training ratio/%	NWPU-RESISC45 20% training ratio/%	GID 20% training ratio/%
DenseNet-121	63.4	63.5	24.4
Proposed method	**135.6**	**138.1**	**139.3**

3　结论

本文提出了一种光学遥感图像场景分类方法，该方法分别利用视觉转换器和图卷积网络分别对遥感场景图像进行建模，借助视觉转换器挖掘了图像内部的长距离依赖关系，通过图卷积网络感知遥感场景潜在的空间拓扑关系，在这两种关系表示的基础上，生成更具有鉴别能力的特征表示并被应用于遥感图像场景分类。在国际公开的AID、NWPU-RESISC45和GID遥感图像场景分类数据集上的实验验证了所提方法在遥感场景分类中具有较优的分类识别能力，同时表明遥感场景图像内部潜在的长距离依赖关系和空间拓扑关系的挖掘有助于提升特征表示的鉴别能力。

参考文献

[1] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]. Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2012: 1097-1105.

[2] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint, arXiv:1911.04129, 2014.

[3] SZEGEDY C, LIU Wei, JIA Yangqing, et al. Going deeper with convolutions[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 1-9.

[4] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 770-778.

[5] CHENG Gong, YANG Ceyuan, YAO Xiwen, et al. When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs[J]. IEEE Transactions on Geoscience and Remote Sensing, 2018, 56(5): 2811-2821.

[6] CHENG Gong, XIE Xingxing, HAN Junwei, et al. Remote sensing image scene classification meets deep learning: challenges, methods, benchmarks, and opportunities[J]. IEEE Journal of Selected Topics in Applied Earth Observations

and Remote Sensing, 2020, 13: 3735-3756.
[7] CHENG Gong, SUN Xuxiang, LI Ke, et al. Perturbation-seeking generative adversarial networks: A defense framework for remote sensing image scene classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, DOI: 10.1109/TGRS.2021.3081421.
[8] ZOU Qin, NI Lihao, ZHANG Tong, et al. Deep learning based feature selection for remote sensing scene classification[J]. IEEE Geoscience and Remote Sensing Letters, 2015, 12(11): 2321-2325.
[9] XIA Guisong, HU Jingwen, HU Fan, 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.
[10] CHENG Gong, HAN Junwei, LU Xiaoqiang. Remote sensing image scene classification: Benchmark and state of the art [J]. Proceedings of the IEEE, 2017, 105(10): 1865-1883.
[11] YU Yunlong, LIU Fuxian. Dense connectivity based two-stream deep feature fusion framework for aerial scene classification[J]. Remote Sensing, 2018, 10(7):1158-1183.
[12] HE Nanjun, FANG Leyuan, LI Shutao, et al. Remote sensing scene classification using multilayer stacked covariance pooling[J]. IEEE Transactions on Geoscience and Remote Sensing, 2018, 56(12): 6899-6910.
[13] BIAN Xiaoyong, CHEN Chunfang, DENG Chunhua, et al. Hierarchical deep feature representation for high-resolution scene classification[C]. Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2019: 517-520.
[14] ZHANG Jun, ZHANG Min, SHI Lukui, et al. A multi-scale approach for remote sensing scene classification based on feature maps selection and region representation[J]. Remote Sensing, 2019, 11(21):2504-2523.
[15] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]. Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2017: 6000-6010.
[16] WANG Fei, JIANG Mengqing, QIAN Chen, et al. Residual attention network for image classification[C]. Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR), 2017: 3156-3164.
[17] HU Jie, SHEN Li, SUN Gang. Squeeze-and-excitation networks[C]. Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR), 2018: 7132-7141.
[18] CAO Ran, FANG Leyuan, LU Ting, et al. Self-attention-based deep feature fusion for remote sensing scene classification[J]. IEEE Geoscience and Remote Sensing Letters, 2020, 18(1): 43-47.
[19] BIAN Xiaoyong, FEI Xiongjun, MU Nan. Remote sensing image scene classification based on scale-attention network [J]. Journal of Computer Applications, 2020, 40(3): 872-877.
边小勇, 费雄君, 穆楠. 基于尺度注意力网络的遥感图像场景分类[J]. 计算机应用, 2020, 40(3): 872-877.
[20] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16 × 16 words: Transformers for image recognition at scale[J]. arXiv preprint, arXiv:2010.11929, 2020.
[21] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint, arXiv: 1609.02907, 2017.
[22] ACHANTA R, SHAJI A, SMITH K, et al. SLIC superpixels compared to state-of-the-art superpixel methods[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(11): 2274-2282.
[23] TONG Xinyi, XIA Guisong, LU Qikai, et al. Land-cover classification with high-resolution remote sensing images using transferable deep models[J]. Remote Sensing of Environment, 2020, 237: 111322.

Foundation item: Anhui Provincial Natural Science Foundation (No. 1908085MF210)