激光与光电子学进展 先进成像

融合 CNN 与 Transformer 结构的遥感图像分类方法

金传，童常青*
杭州电子科技大学理学院，浙江 杭州 310018

摘要 为解决高分辨率遥感图像所具有的类内差异大而类间差异小的特性导致的图像难分类问题，提出一种基于深度学习中卷积神经网络与 Transformer 优点的混合结构。对卷积层提取的特征信息使用两个带有空间位置信息的注意力机制，分别沿水平方向和垂直方向对每个通道进行特征聚集，以减少遥感场景特征的冗余映射，使网络能够提取更多与任务目标相关的信息。然后利用 Transformer 编码器结构对捕获的特征图进行编码操作，赋予特征图中感兴趣区域较大的权重。实验结果表明，与现有的基于深度学习的遥感图像分类方法相比，所提方法既降低了模型参数量，又提升了分类准确率，在遥感图像分类数据集 AID、NWPU-RESISC45 及 VGoogle 上均达到了最高的平均分类准确率，分别为 98.95%、96.00% 和 95.01%。

关键词 图像分类；卷积神经网络；Transformer；空间位置信息；注意力机制
中图分类号 TP751 **文献标志码** A **DOI**：10.3788/LOP223154

Remote Sensing Image Classification Method Based on Fusion of CNN and Transformer

Jin Chuan, Tong Changqing*
School of Sciences, Hangzhou Dianzi University, Hangzhou 310018, *Zhejiang, China*

Abstract To solve the difficult problem of the classification of high-resolution remote sensing images having large intraclass differences and small interclass differences, a hybrid structure using the advantages of convolutional neural networks and a Transformer in deep learning is proposed herein. Feature clustering is carried out for each channel along the horizontal and vertical directions using two attention mechanisms with spatial location information for the features extracted from the convolutional layer. This reduces the redundant mapping of remote sensing scene features and enables the network to extract more information relevant to the task object. Then, the captured feature maps are processed via encoding operations using the Transformer encoder structure to enable the allocation of greater weights to the regions of interest in the feature maps. The experimental results show that the proposed method reduces number of model parameters and increases the classification accuracy compared with the existing deep learning-based remote sensing image classification methods, achieving the highest average classification accuracy of 98.95%, 96.00%, and 95.01% on the remote sensing image classification datasets of AID, NWPU-RESISC45, and VGoogle, respectively.

Key words image classification; convolutional neural network; Transformer; spatial location information; attention mechanism

1 引 言

近年来，伴随着科学技术的快速发展，高分辨率的遥感图像被轻易获取，并广泛应用于城乡规划[1-2]、环境保护[3]、地理位置检索[4]和空间目标检测[5]等领域。遥感图像分类是遥感影像研究的重要方向，针对遥感图像类内差异大而类间差异小的特性，有效地区分不同场景尤为重要。因此，如何在复杂的特征纹理信息与空间尺度分布条件下提高模型的分类精度与泛化性能，具有重要的科学研究意义和实际应用价值。

早期遥感图像分类方法假定相同类别的场景应该共享相似的特征信息，根据人类视觉的特点直接设计特征本身并计算相似度。遥感图像分类任务主要面临类内差异大而类间差异小的问题，前者主要是视角差

收稿日期：2022-11-24；**修回日期**：2022-12-22；**录用日期**：2023-01-04；**网络首发日期**：2023-02-07
基金项目：国家社会科学基金(21BTJ071)
通信作者：*tongchangqing@hdu.edu.cn

异、目标形变、光照强弱及尺度变化等造成的；后者则主要是不同类别含有相同的底层特征，导致不同类别图像被分为同一类别造成的。为解决这个问题，朱国龙等[6]提出了一种适用于组合特征识别的遥感图像最近邻模糊分类器，对识别目标的组合特征与训练模板中的组合特征样本的平均值进行比较，克服了特征选择的不稳定性对分类结果产生的影响。赵蔷等[7]提出了一种对卫星遥感图像进行颜色特征提取的PCA-K-means算法，该算法去除了图像不同通道之间的相关性，在动态聚类的基础上，采用区域分类的空间一致性原则合并空间信息。冯逍等[8]根据高光谱遥感图像的特点和二维Gabor滤波器纹理分割的原理，提出了一种基于三维Gabor滤波器的高光谱遥感图像分类方法，该方法对高光谱遥感图像所有波段同时进行滤波，极大减少了高光谱遥感图像纹理信息提取的计算量。上述方法虽然在一定程度上可以提升遥感图像的分类准确率，但都是基于图像底层信息的，如尺度、形态、颜色和纹理特征等，缺乏语义信息，导致分类准确率难以满足实际应用需求。而与深度学习的分类方法[9-13]相结合后在很大程度上解决了图像特征信息提取问题，极大地提升了图像的分类准确率。特别是自AlexNet[14]被提出后，凭借卷积神经网络(CNN)精确捕获高层语义信息的能力，在计算机视觉领域获得广泛关注。但对于遥感图像分类任务，相关算法不但面临着内类差异大、类间差异小的问题，而且容易受到多类别标签和复杂背景环境的干扰，导致丢失空间位置信息。为解决此类问题，车思韬等[15]提出一种基于有监督对比学习的注意力机制和残差收缩单元算法，优化对待识别图像特征的提取。得益于Transformer模型[16]在自然语言处理领域取得的巨大成功，Dosovitskiy等[17]提出了视觉转换器(ViT)模型，凭借动态的、全局的感受野，模型在图像分类任务中表现优异，可以很好地学习图像不同位置的依赖关系。陈辉等[18]提出了一种基于Swin Transformer的多尺度混合光谱注意力网络(SMSaNet)，该网络有效地解决了不同特征区域之间无法对全局信息建立长距离依赖关系，从而导致遥感图像分类准确率不高的问题。Zhao等[19]针对图像中有限的感受野限制了CNN表示全局上下文和顺序属性，并且ViT容易丢失局部语义信息的问题，提出了一个分数阶傅里叶图像变换器(FrIT)作为骨干网络，以有效地提取全局和局部上下文。

综上所述，针对遥感图像分类任务，为使模型能够精确捕获高层语义信息的同时降低计算复杂度，考虑将CNN与Transformer结构相结合。其中CNN部分嵌入协调注意力(CA)机制，从通道和空间两个维度进行特征提取，使得特征信息的提取更加精确，从而保证在后续的模型训练中特征信息不易丢失，并且通过CNN进行特征提取能够有效缓解Transformer结构带来的高计算量问题。在CNN结构后接Transformer编码器则是为了使用其中的多头自注意力机制进行特征信息的重加权操作，使得模型在训练过程中能够精确捕获图像中感兴趣的特征区域。

2 模型构建

2.1 模型整体结构

模型卷积部分主要是基于MBConv结构的[13]，由于该结构中的Squeeze-and-Excitation(SE)模块忽略了位置信息的重要性，故使用带有空间位置信息的协调注意力(CA)机制[20]替换MBConv结构中的SE模块。替换后的MBConv结构如图1所示，其中k为卷积核大小，s为步距。

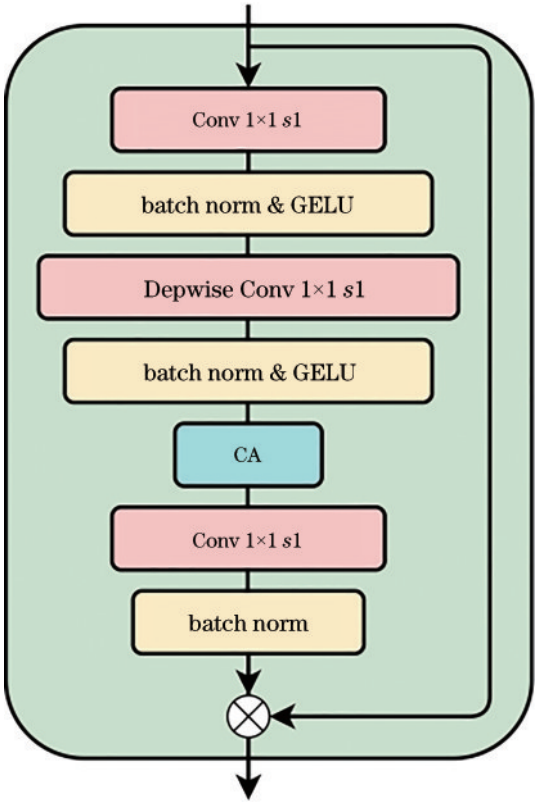

图1 嵌入CA机制的MBConv结构

Fig. 1 MBConv structure embedded in the CA mechanism

卷积操作使用的MBConv结构同后接的Transformer编码器均采用了“倒残差”结构的设计，该设计首先使用1×1卷积实现升维，再通过3×3的Depwise卷积提取特征，最后使用1×1卷积实现降维，以实现不同特征信息之间的相互融合。由于MBConv结构中的Hard-Swish激活函数计算成本较高，且函数只有在更深的网络层使用才能体现其优势，为了使网络更快更好地收敛，使用高斯误差线性单元(GELU)[21]作为网络的激活函数，表达式为

$$\mathrm{GELU}(x)=xP(X\leqslant x)=\frac{x}{\sqrt{2\pi}\,\sigma}\int_{-\infty}^{x}\mathrm{e}^{-\frac{(X-\mu)^2}{2\sigma^2}}\mathrm{d}X,\tag{1}$$

式中：μ和σ分别为正态分布的均值和标准差。

此外，卷积操作和自注意力机制都可以表示为预定感受野的加权值之和。具体来说，卷积依赖于固定的卷积核从局部感受野捕获的图像的特征信息，既保证了学习到的卷积核对输入图像的局部特征具有最强的响应，又降低了模型的复杂度，但局部特征容易忽略图像不同区域之间的语义相关性。自注意力机制则可以直接捕获感受野的全局信息，并根据配对之间的相似性计算归一化权重，然而全局感受野使得模型计算量骤增。因此，在输入 Transformer 编码器结构之前，先利用 CNN 提取图像中感兴趣的特征区域，这有助于找出希望保留的特性信息同时降低模型的计算量。

鉴于上述比较，理想的模型应当能够结合 CNN 与 Transformer 的优势，整体结构如图 2 所示，该结构既能够充分提取图像中感兴趣的特征区域，又能够给予重要的特征信息较大的权重。因此，考虑在 Softmax 归一化操作之前，通过捷径分支对卷积操作与自注意力机制进行融合，表达式为

$$\boldsymbol{y}_i=\sum_{j\in\Theta}\frac{\exp\left(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j+\omega_{i-j}\right)}{\sum_{k\in\Theta}\exp\left(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_k+\omega_{i-k}\right)}\boldsymbol{x}_j,\tag{2}$$

式中：$\boldsymbol{x}_i,\boldsymbol{y}_i\in\mathbb{R}^D$ 分别表示位置 i 的输入与输出；ω_{i-j} 表示卷积操作中预定感受野对应的权值；Θ 表示全局空间。

2.2　CA 机制

关于 CA 机制的详细结构如图 3 所示，该结构充分考虑了位置信息。通过编码通道位置关系和长距离依存关系，即坐标信息嵌入和坐标位置生成，能够准确地显示网络感兴趣的区域。其次，为了获取注意力模块在空间上捕捉到的具有精确位置的特征信息，将全局池化核分解为一对一维特征，并进行编码操作。

具体来说，为了在卷积过程中将位置信息嵌入到通道注意力中，使用两个尺寸为 $(H,1)$ 和 $(1,W)$ 的池化核对输入 $\boldsymbol{X}$ 分别沿水平方向和垂直方向进行编码。因此，高度 h 处第 c 个通道的输出为

$$z_c^h=\frac{1}{W}\sum_{0\leqslant i\leqslant W}x_c(h,i),\tag{3}$$

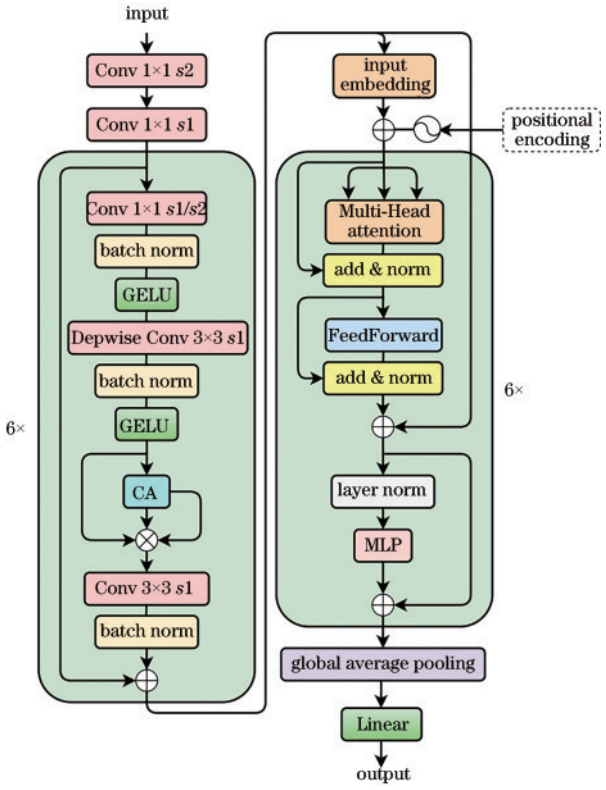

图 2　模型整体结构

Fig. 2　Overall structure of the proposed model

类似地，宽度 w 处的第 c 个通道的输出为

$$z_c^w=\frac{1}{H}\sum_{0\leqslant j\leqslant H}x_c(j,w)。\tag{4}$$

与产生单个特征向量的 SE 模块不同，通过上述转换方式，模型沿两个方向进行特征聚集，一个沿空间方向获取图像中不同物体之间的相互依存关系，另一个沿空间方向保留图像中单个物体的精确位置信息，生成的特征图具备方向感知能力。此外，这种转换使得注意力机制能够有效地捕获通道间关系，有助于模型更准确地定位图像中感兴趣的区域。

给定式(3)和式(4)生成的关于输入 $\boldsymbol{X}$ 的特征映射，首先对它们在深度方向进行拼接，然后对它们进行

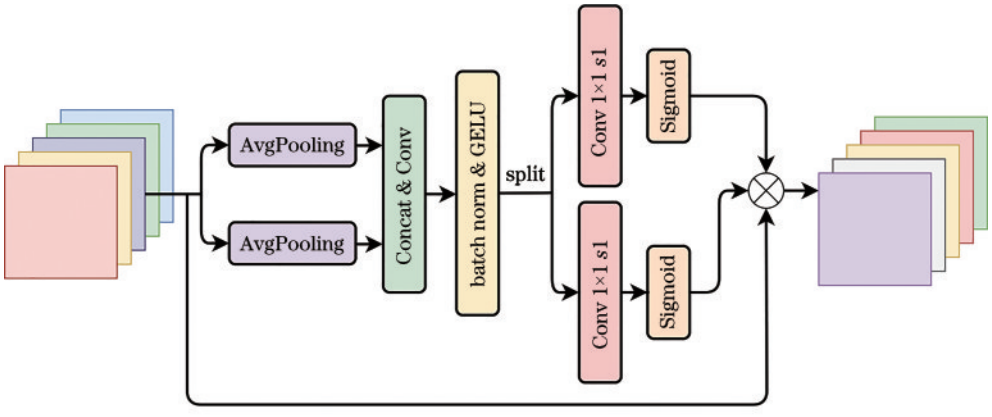

图 3　CA 结构

Fig. 3　Structure of the CA

权重共享的1×1卷积F_1操作，公式为

$$\boldsymbol{f}=\delta\left(F_1\left[\boldsymbol{z}^h,\boldsymbol{z}^w\right]\right),\tag{5}$$

式中：$[\cdot,\cdot]$表示沿深度方向的拼接操作；δ为激活函数；$\boldsymbol{f}\in\mathbb{R}^{C/r\times(H+W)}$是在水平方向和垂直方向上的中间特征图；$r$控制缩放比例。然后，将$\boldsymbol{f}$沿深度方向拆分为两个独立的张量，即$\boldsymbol{f}^h\in\mathbb{R}^{C/r\times H}$和$\boldsymbol{f}^w\in\mathbb{R}^{C/r\times W}$，再利用两个1×1卷积$F_h$和$F_w$操作对$\boldsymbol{f}^h$和$\boldsymbol{f}^w$进行处理，分别生成同输入$\boldsymbol{X}$具有相同通道数的张量，公式为

$$\boldsymbol{g}^h=\sigma\left[F_h\left(\boldsymbol{f}^h\right)\right],\boldsymbol{g}^w=\sigma\left[F_w\left(\boldsymbol{f}^w\right)\right],\tag{6}$$

分别用作水平方向和垂直方向的注意力权重。最后，输出表示为

$$y_c(i,j)=x_c(i,j)\times g_c^h(i)\times g_c^w(j)。\tag{7}$$

相较于仅关注注意力机制在不同通道间进行重加权操作的SE模块，CA机制还考虑了位置信息。即沿水平和垂直方向的注意力同时应用于输入$\boldsymbol{X}$，两个注意力图中的每个元素都反映了感兴趣的对象是否存在于相应的行和列中，这种编码过程能够更准确地定位图像中感兴趣对象的准确位置。

2.3　Transformer编码器结构

Transformer结构主要由多头自注意力机制和前馈神经网络组成。多头自注意力机制由多个自注意力机制组成，如图4所示。其中自注意力机制的输入是CNN输出的特征映射，即查询矩阵$\boldsymbol{Q}$、键矩阵$\boldsymbol{K}$和值矩阵$\boldsymbol{V}$。自注意力机制是注意力机制的一种特殊情况，即序列与自身匹配，以提取其部分之间的语义依赖关系。每个自注意力机制的计算公式为

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V},\tag{8}$$

式中：d_k为通道维度。查询矩阵$\boldsymbol{Q}$与键矩阵$\boldsymbol{K}$使用内积匹配，结果是一个注意力矩阵，其值刻画了$\boldsymbol{Q}$和$\boldsymbol{K}$在语义上的相关程度。多头自注意力机制并行使用多个自注意力，学习不同类型数据之间的相互依赖关系，公式为

$$\text{Multihead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Concat}\left(\boldsymbol{y}_{\text{head }1},\cdots,\boldsymbol{y}_{\text{head }h}\right)\boldsymbol{W}^o,\tag{9}$$

式中：$\boldsymbol{y}_{\text{head }i}=\text{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^Q,\boldsymbol{K}\boldsymbol{W}_i^K,\boldsymbol{V}\boldsymbol{W}_i^V)$，$\boldsymbol{W}_i^Q\in\mathbb{R}^{d_{\text{model}}\times d_k}$，$\boldsymbol{W}_i^K\in\mathbb{R}^{d_{\text{model}}\times d_k}$，$\boldsymbol{W}_i^V\in\mathbb{R}^{d_{\text{model}}\times d_v}$，$\boldsymbol{W}_i^O\in\mathbb{R}^{hd_v\times d_{\text{model}}}$，选定$h=8$，即由8个自注意力机制构成。

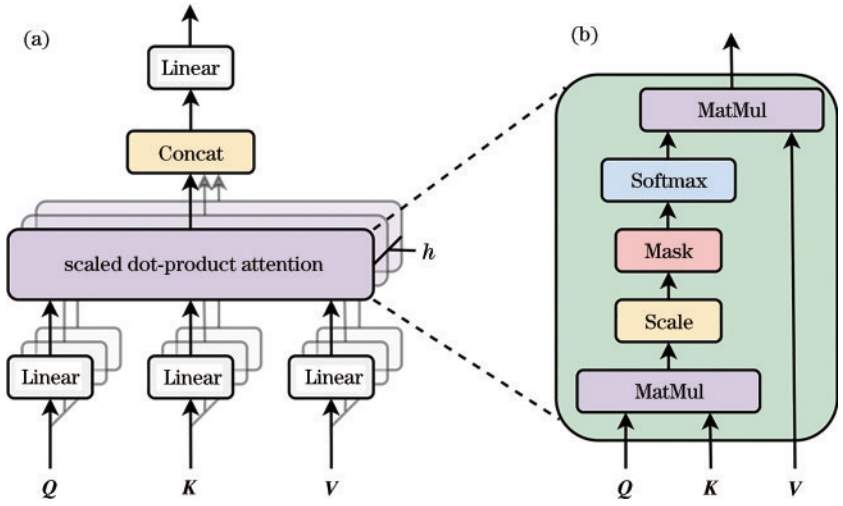

图4　多头自注意力机制的结构。(a) 多头自注意力机制；(b) 自注意力机制

Fig. 4　Multihead self-attention mechanism structure. (a) Mutilhead self -attention mechanism; (b) self-attention mechanism

Transformer编码器结构中的每一层都包含一个全连接的前馈神经网络，分别相同地应用于每个位置。该结构包括两个线性变换，中间为ReLU激活函数，公式为

$$\text{FFN}(x)=\max(0,xW_1+b_1)W_2+b_2,\tag{10}$$

虽然线性变换在不同位置上是相同的，但它们在层与层之间使用的参数却不相同。

在式(8)中，由于缺少位置信息，自注意力层的位置是不可知的，考虑使用位置自注意力代替自注意力，嵌入可学习的相对位置编码[22]。每个注意力头使用一个可训练的相对位置编码，其值仅取决于像素与像素之间的距离，将自注意力机制表示为

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}+B\right)\boldsymbol{V},\tag{11}$$

式中：B表示相对位置编码，为网络训练过程中可学习的参数。

3　模型训练与结果分析

3.1　实验环境

实验通过连接远程服务器进行训练，操作系统为Ubuntu 18.04.3、CPU为2.60 GHz Intel Xeon Gold 6240、GPU为NVIDIA Tesla V100、显存为32G。实验环境为Python 3.8和CUDA 11.0，深度学习模型框架为PyTorch 1.8.1，模型对比调用timm库进行训练。

3.2 实验数据与预处理

为了验证所提模型的有效性和适用性，使用AID[23]、NWPU-RESISC45[24]及VGoogle[25]s3个公共数据集进行验证。其中，AID是遥感图像场景分类任务中常用的数据集；NWPU-RESISC45与AID相比，场景类别丰富，具有更高的类间相似性；VGoogle类内多样性和类间相似性较高，对遥感图像场景分类具有更高的挑战性。3个数据集的特征如表1所示，部分场景实例如图5所示。为防止过拟合，实验中采用的数据增强方式有：随机旋转，对图像进行0°～360°的随机旋转，模拟图像获取过程中角度的随机性；随机缩放，保持图像原始尺寸不变的条件下，对图像进行随机比例的缩放，模拟检测过程中受距离影响造成的物体尺度不同的情况；随机裁剪，对图像任意区域进行裁剪，模拟检测过程中物体部分缺失以及遮挡的情况；饱和度和色度的调整，模拟检测过程中物体受光线影响的情况。

表1 数据集特征

Table 1 Characteristics of the datasets

Dataset	Number of classes	Number of images	Total number of images	Resolution /m	Image size	Year
AID	30	220–420	10000	0.5–8	600×600	2017
NWPU-RESISC45	45	700	31500	0.2–30	256×256	2016
VGoogle	38	1502–1847	59404	0.075–9.555	256×256	2019

图5 三种数据集的场景实例。(a) 机场；(b) 海滩；(c) 中心；(d) 公园；(e) 山脉；(f) 桥梁；(g) 教堂；(h) 港口；(i) 立交桥；(j) 云朵；(k) 河流；(l) 路口；(m) 飞机；(n) 灌木丛；(o) 储存罐

Fig. 5 Scene instances of three datasets. (a) Airport; (b) beach; (c) center; (d) park; (e) mountain; (f) bridge; (g) church; (h) harbor; (i) overpass; (j) cloud; (k) river; (l) intersection; (m) aeroplane; (n) chaparral; (o) storage tank

3.3 模型训练

尽管随机梯度下降(SGD)算法在模型训练中表现优异，但是，由于学习率被分散使用在模型训练中的各个参数上，模型在训练过程中收敛较慢，需要花费较多的时间进行调整。此外，SGD在误差函数的优化过程中容易陷入局部最优解。因此，网络采用带有权重衰减的AdamW优化器。学习率衰减方法为余弦退火，输入网络的批量大小为64，模型训练迭代次数为100。考虑到刚开始训练时模型的权重是随机初始化的，此时若选择一个较大的学习率，可能导致模型的不稳定，因此选择Warmup预热学习率的方式。模型训练过程中具体参数设置如表2所示。

表2 模型训练的参数设置

Table 2 Parameter settings for model training

Parameter	Value	Parameter	Value
Epoch	100	Drop rate	0.2
Batch_size	64	Optimiser	AdamW
Learning rate	0.000005	Warmup	10
Weight decay	0.0005	Random seed	42

实验过程中采用总体精度(OA)作为评价指标，即被正确分类的样本数占该类别样本总数的比例。利用AID、NWPU-RESISC45和VGoogle三种公开数据集进行验证实验。为方便同相关方法进行比较，训练比例与文献[26-28]中的数据集训练比例设置保持一

致，随机从三个数据集中挑选部分数据进行训练和测试。在AID数据集中，随机选取20%和50%场景图像作为训练集，其余作为测试集；在NWPU-RESISC45与VGoogle数据集中，随机选取10%和20%场景图像作为训练集，其余作为测试集。所对比的相关模型包括VGG-16、GoogLeNet、EfficientNet-B0、ResNet-50、LGRIN、ViT、PVT、Swin-T、TRS和TSTNet遥感图像分类模型。在相同的实验环境、图像预处理方式设置条件下，所有实验均重复5次，评估不同模型在3种数据集上取得的准确率，结果如表3所示。

表3 不同模型在3种数据集上的准确率

Table 3 Accuracy of different models on three datasets units: %

Method	Number of parameters/10^6	AID		NWPU-RESISC45		VGoogle	
		20% training data	50% training data	10% training data	20% training data	10% training data	20% training data
VGG-16	134.4	86.59±0.29	89.64±0.36	76.47±0.18	79.79±0.65	72.41±0.22	76.74±0.16
GoogLeNet	54.4	83.44±0.40	86.39±0.55	76.19±0.38	78.48±0.26	77.33±0.57	86.79±0.47
EfficientNet-B0	4.1	83.69±0.11	86.17±0.16	79.96±0.27	82.89±0.16	78.30±0.26	88.38±0.29
ResNet-50	23.6	92.39±0.15	94.96±0.19	86.23±0.41	88.93±0.12	88.02±0.15	92.99±0.10
LGRIN	4.6	94.74±0.23	97.65±0.25	91.91±0.15	94.43±0.16		
ViT-Base	85.8	91.16±0.41	94.44±0.28	87.59±0.21	90.87±0.17	86.22±0.33	91.42±0.17
PVT-Medium	43.3	92.84±0.19	95.93±0.17	90.51±0.13	92.66±0.14	86.60±0.14	92.32±0.22
Swin-Base	86.8	94.86±0.22	97.80±0.15	91.80±0.16	94.04±0.11	88.48±0.12	93.19±0.13
TRS	46.3	95.54±0.18	98.48±0.06	93.06±0.11	95.56±0.20		
TSTNet	173.0	97.20±0.22	98.70±0.12	94.08±0.24	95.70±0.10		
Proposed method	20.4	97.81±0.08	98.95±0.06	94.82±0.04	96.00±0.07	91.27±0.02	95.01±0.14

由表3可知：相比单一结构模型，所提方法在3种数据集的不同训练比例下的分类准确率都达到了最佳。在AID数据集20%和50%训练比例下，所提方法所达到的平均准确率分别为97.81%和98.95%；在NWPU-RESISC45数据集10%和20%训练比例下，所达到的平均准确率分别为94.82%和96.00%；在VGoogle数据集10%和20%训练比例下，所达到的平均准确率分别为91.27%和95.01%，均超过了其他现有方法。

为进一步验证模型设计的合理性，对模型生成的特征图截取反向传播过程中获得的梯度信息，在此基础上使用类别激活热力图(Grad-CAM)[29]可视化模型，如图6所示。显然，构造的模型可以精确地定位图像中感兴趣的对象。

3.4 消融实验

将卷积操作、CA机制和Transformer编码器三部分交互考虑，从而研究模型准确率提升的主要原因，分别对比CA机制与Transformer编码器结构对实验结果的影响。在相同的实验环境、图像预处理和网络超参数设置条件下进行实验，结果如表4所示。

由表4可知，当使用CA机制和Transformer编码器相结合的模型时，在3个数据集上的准确率均有较大提升。在AID数据集不同训练比例下，相比原始模型，CA机制和Transformer编码器相结合的模型的准确率分别提升12.64个百分点和5.98个百分点；在NWPU-RESISC45数据集不同训练比例下，准确率分别提升17.01个百分点和9.45个百分点；在VGoogle数据集不同训练比例下，准确率分别提升4.95个百分点和2.18个百分点。通过实验结果可知：相比原始模型，当添加CA机制时，模型的分类准确率有小幅度提升，在AID数据集不同训练比例下，分别提升1.4个百分点和0.45个百分点；NWPU-RESISC45数据集不同训练比例下，分别提升1.44个百分点和0.69个百分

表4 消融实验的准确率

Table 4 Accuracy of ablation experiments unit: %

Method	AID		NWPU-RESISC45		VGoogle	
	20% training data	50% training data	10% training data	20% training data	10% training data	20% training data
Without CA+Transformer	85.17±0.57	92.97±0.09	77.81±0.16	86.55±0.13	86.32±0.16	92.83±0.10
With CA	86.57±0.85	93.42±0.09	79.25±0.25	87.24±0.09	86.91±0.20	92.79±0.20
With Transformer	77.79±0.17	90.45±0.32	72.06±0.32	83.88±0.27	76.97±0.30	86.65±0.22
With CA+Transformer	97.81±0.08	98.95±0.06	94.82±0.04	96.00±0.07	91.27±0.02	95.01±0.14

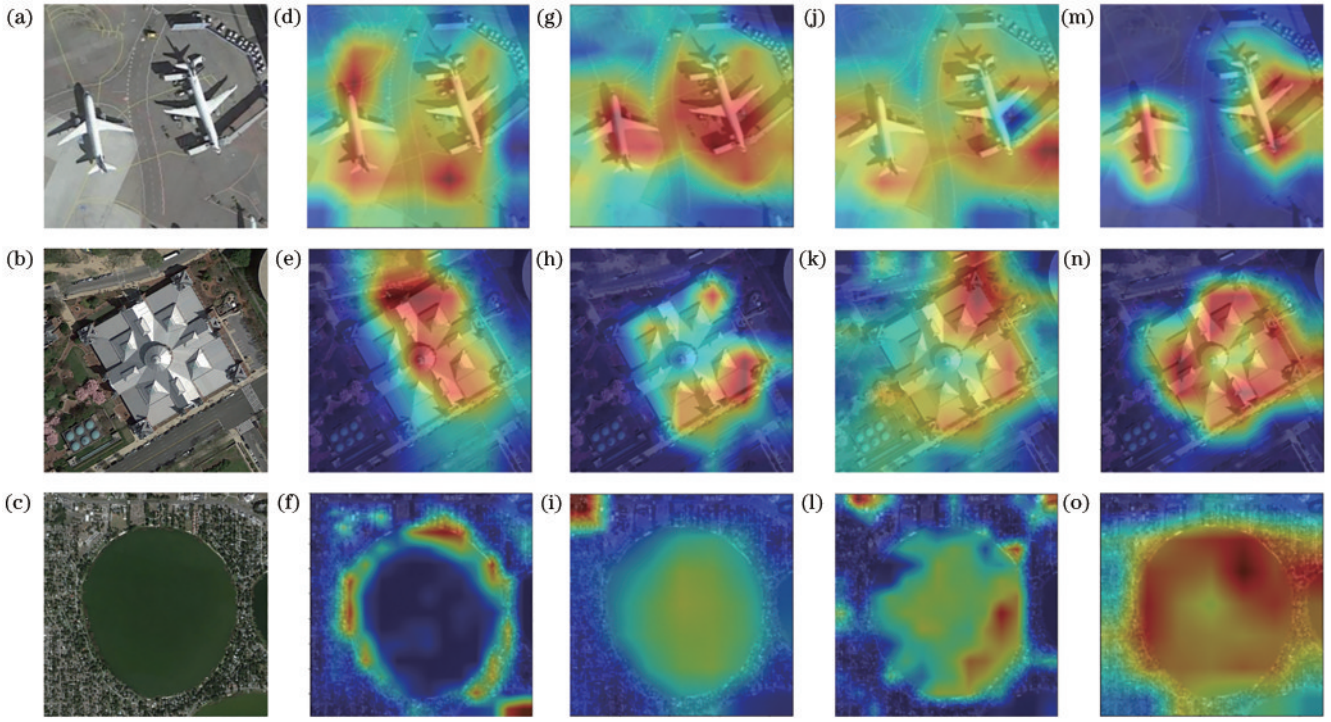

图 6　与其他相关模型的热力图对比。(a) 飞机；(b) 中心；(c) 池塘；(d)(e)(f) EfficientNet；(g)(h)(i) ResNet；(j)(k)(l) Swin-T；(m)(n)(o) 所提方法

Fig. 6　Heat map comparison with other related models. (a) Airplane; (b) center; (c) pond; (d) (e) (f) EfficientNet; (g) (h) (i) ResNet; (j) (k) (l) Swin-T; (m) (n) (o) proposed method

点；在 VGoogle 数据集 10% 的训练比例下提升 0.59 个百分点，20% 训练比例下减少 0.04 个百分点。相反地，当添加 Transformer 编码器结构后，分类准确率下降，在 AID 数据集不同训练比例下分别下降 7.38 个百分点和 2.52 个百分点；在 NWPU-RESISC45 数据集不同训练比例下，分别下降 5.75 个百分点和 2.67 个百分点；在 VGoogle 数据集不同训练比例下，分别下降 9.35 个百分点和 6.18 个百分点。

相较于 MBConv 中的 SE 模块，CA 机制可以从通道和空间两个维度进行特征聚集，对特征区域的定位较为准确，因此模型检测精度有小幅度上升。而关于使用 Transformer 编码器使得模型检测精度降低的原因分析如下：首先 Transformer 对图像局部信息的捕获能力不如 CNN，Transformer 结构凭借动态的、全局的感受野虽然可以直接获取图像的全局信息，但却无法关注各个特征之间的局部相关性；其次嵌入的位置编码存在问题，遥感分类面临内类差异大、类间差异小的问题，精确定位有关类别的特征信息至关重要，然而位置编码在语义空间中并不具有这种差异性变化，它相当于一种人为设计的特殊索引，虽然可以通过网络训练得到不断更新，但是仍然不能很好地表示特征的位置信息；最后，模型的顶层梯度可能会消失，Transformer 结构实际上是由一些残差模块与层归一化组合而成的，因此最终的输出层与之前的 Transformer 层都没有直连的通路，梯度流会被层归一化模块阻断，故而顶层的参数很难被更新，导致梯度消失。

3.5　可视化实验

使用混淆矩阵来可视化最终的分类结果，如图 7～9 所示，分别表示模型对 3 个数据集的分类混淆矩阵，矩阵的行和列分别表示样本的预测值和真实值，矩阵坐标系数字为数据集中相应的类别名称，按英文首字母顺序进行排序，矩阵中元素 x_{ij} 表示将第 j 种类别预测为第 i 种类别的样本数占该类别样本总数的比例(p)。可以看出，所设计的模型分类性能较好，在 3 个数据集上均取得了较高的总体分类准确率。在 AID 数据集 50% 训练比例下，所设计的模型对 30 类场景的识别率均达 95% 以上；在 NWPU-RESISC45 数据集 20% 训练比例下，对 45 类场景中的 39 类的识别率达 95% 以上；在 VGoogle 数据集 20% 训练比例下，对 38 类场景中的 30 类的识别率达 95% 以上。如图 7 所示，在 AID 数据集 50% 训练比例下，所提方法在一定程度上解决了遥感图像具有的类内差异大而类间差异小的难题，针对高类间相似性的类别 22(resort)和类别 16(park)，高类内差异性的类别 6(church)，所提方法均能取得超过 95% 的分类准确率。

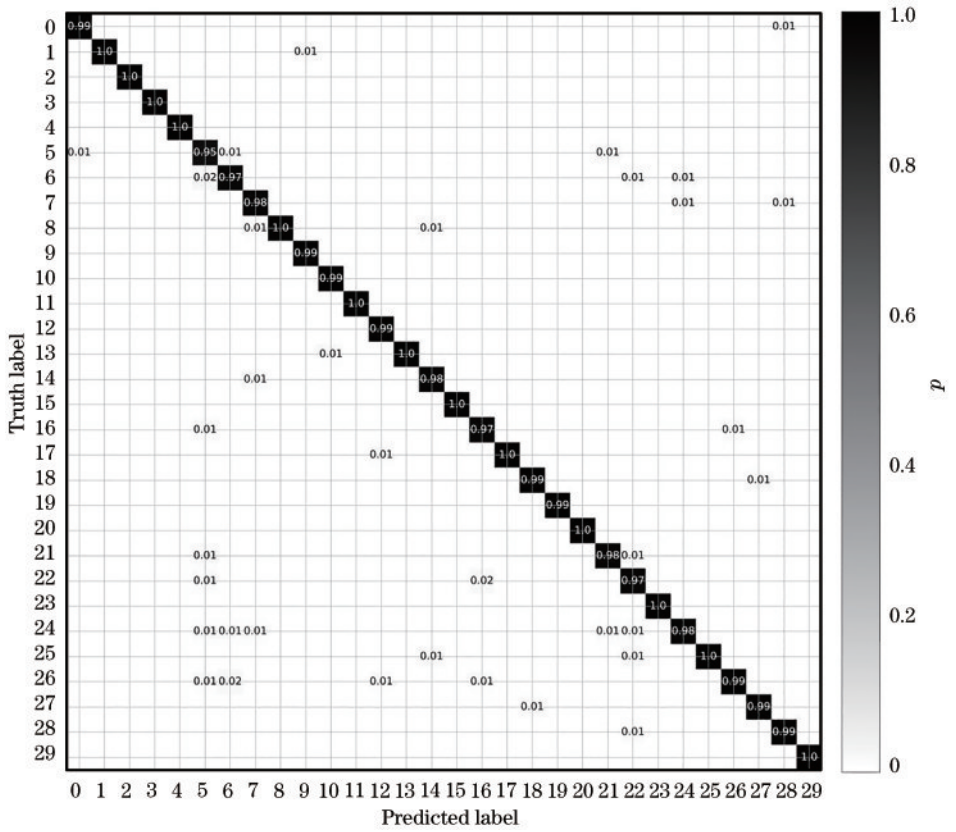

图 7　50% 训练比例下 AID 数据集的混淆矩阵

Fig. 7　Confusion matrix of the AID dataset at 50% training scale

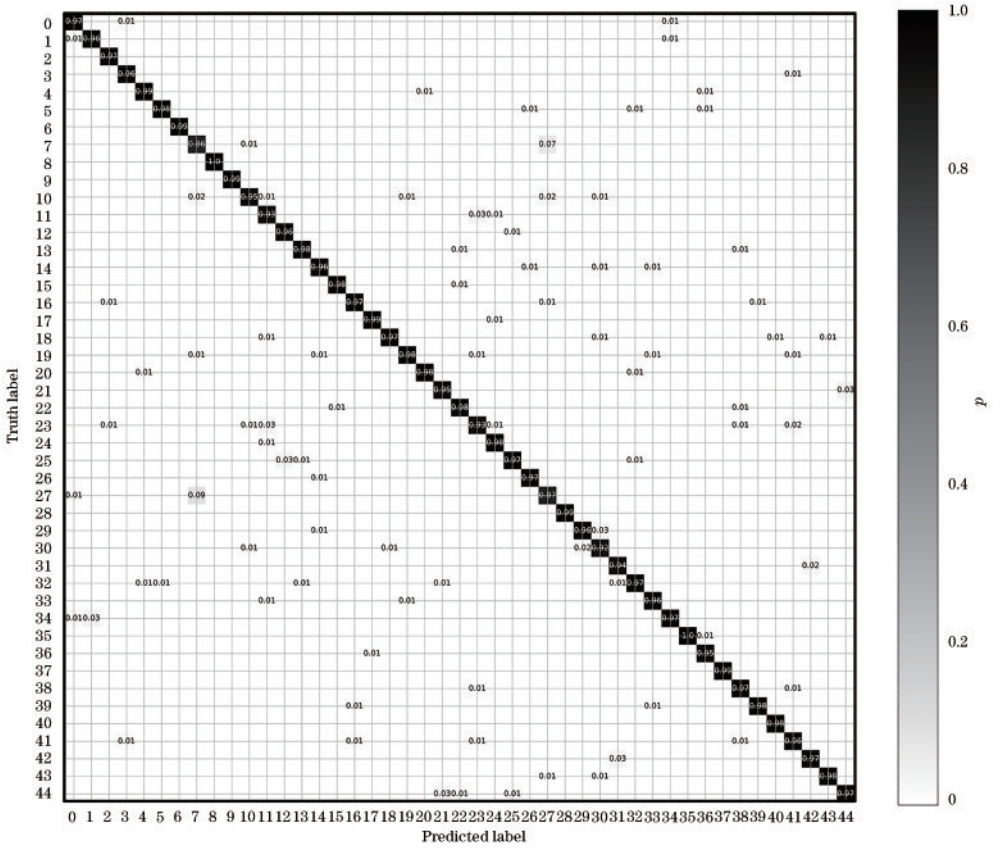

图 8　20% 训练比例下 NWPU-RESISC45 数据集的混淆矩阵

Fig. 8　Confusion matrix of the NWPU-RESISC45 dataset at 20% training scale

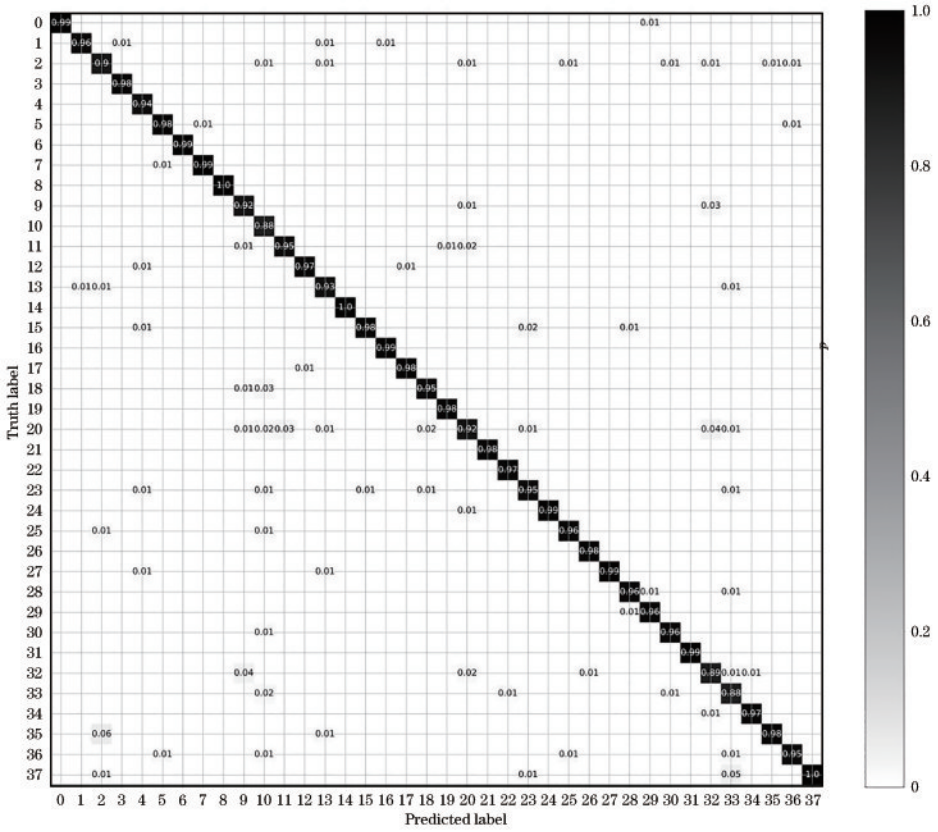

图 9　20% 训练比例下 VGoogle 数据集的混淆矩阵

Fig. 9　Confusion matrix of the VGoogle dataset at 20% training scale

4　结　　论

为解决遥感图像难分类问题，提出一种结合 CNN 与 Transformer 优点的图像分类方法。在 CNN 提取图像特征信息时向通道中嵌入空间位置注意力机制，提取图像中感兴趣的特征区域，接着利用 Transformer 编码器赋予感兴趣区域较大的权重，以重点关注显著性区域和显著性特征，从而提升模型的分类准确率。从在 3 个数据集上不同训练比例下的实验结果可知，相较于其他基于深度学习的遥感图像分类方法，所提方法有效地解决了遥感图像分类任务所面临的类内差异大而类间差异小这一难题，在保持分类准确率的同时降低模型复杂度。在进一步的工作中，尝试将完整的 Transformer 架构（编码器加解码器）应用于遥感图像分类任务。另外，针对遥感图像的空间尺度分布复杂的问题，多尺度遥感图像分类网络也将是未来的工作重点之一。

参　考　文　献

[1] Longbotham N, Chapel C, LBleiler, et al. Very high-resolution multiangle urban classification analysis[J]. IEEE Transactions on Geoscience and Remote Sensing, 2012, 50(4): 1155-1170.

[2] Tayyebi A, Pijanowski B C, Tayyebi A H. An urban growth boundary model using neural networks, GIS, and radial parameterisation: an application to Tehran, Iran[J]. Landscape and Urban Planning, 2011, 100(1/2): 35-44.

[3] Huang X, When D W, Li J Y, et al. Multi-level monitoring of subtle urban changes for the megacities of China using high-resolution multi-view satellite imagery [J]. Remote Sensing of Environment, 2017, 196: 56-75.

[4] Wang Y B, Zhang L Q, Tong X H, et al. A three-layered graph-based learning approach for remote- sensing image retrieval[J]. IEEE Transactions on Geoscience and Remote Sensing, 2016, 54(10): 6020-6034.

[5] Mishra N B, Crews K A. Mapping vegetation morphology types in a dry savanna ecosystem: integrating hierarchical object-based image analysis with Random Forest[J]. International Journal of Remote Sensing, 2014, 35(3): 1175-1198.

[6] 朱国龙，汪云甲，乔浩然，等. 高分辨率遥感图像最近邻模糊分类器的研究及实现[J]. 测绘科学，2010，35(6)：96-98.
Zhu G L, Wang Y J, Qiao H R, et al. Study and implement of the nearest neighbour fuzzy classifier for high-resolution remote sensing images[J]. Science of Surveying and Mapping, 2010, 35(6): 96-98.

[7] 赵蔷，解争龙，李红，等. 基于 PCA-K-means 的卫星遥感图像的颜色特征提取技术[J]. 微电子学与计算机，2012，29(10)：64-68.
Zhao Q, Xie Z L, Li H, et al. Color-feature extraction of remote sensing images based on principal components analysis and K-means[J]. Microelectronics & Computer, 2012, 29(10): 64-68.

[8] 冯逍，肖鹏峰，李琦，等. 三维Gabor滤波器与支持向量机的高光谱遥感图像分类[J]. 光谱学与光谱分析，2014，34(8)：2218-2224.
Feng X，Xiao P F，Li Q，et al. The hyperspectral image classification based on a 3D Gabor filter and support vector machines[J]. Spectroscopy and Spectral Analysis，2014，34(8)：2218-2224.

[9] Simonyan Kand Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04)[2022-10-08]. https://arxiv.org/abs/1409.1556.

[10] Szegedy C，Liu W，Jia Y Q，et al. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)，June 7-12，2015，Boston，MA，USA. New York：IEEE Press，2015.

[11] He K M，Zhang X Y，Ren S Q，et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)，June 27-30，2016，Las Vegas，NV，USA. New York：IEEE Press，2016：770-778.

[12] Howard A G，Zhu M L，Chen B，et al. MobileNets：efficient convolutional neural networks for mobile vision applications[EB/OL]. (2017-04-17)[2022-10-08]. https://arxiv.org/abs/1704.04861.

[13] Tan M X，Le Q V. EfficientNet：rethinking model scaling for convolutional neural networks[EB/OL]. (2019-05-28)[2022-10-08]. https://arxiv.org/abs/1905.11946.

[14] Krizhevsky A，Sutskever I，Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM，2017，60(6)：84-90.

[15] 车思韬，郭荣佐，李卓阳，等. 注意力机制结合残差收缩网络对遥感图像分类[J]. 计算机应用研究，2022，39(8)：2532-2537.
Che S T，Guo R Z，Li Z Y，et al. The attention mechanism was combined with a residual shrinkage network to classify remote sensing images[J]. Application Research of Computers，2022，39(8)：2532-2537.

[16] Vaswani A，Shazeer N，Parmar N，et al. Attention is all You need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems，December 4-9，2017，Long Beach，California，USA. New York：ACM，2017：6000-6010.

[17] Dosovitskiy A，Beyer L，Kolesnikov A，et al. An image is worth 16×16 words：transformers for image recognition at scale[EB/OL]. (2020-10-22)[2022-10-08]. https://arxiv.org/abs/2010.11929.

[18] 陈辉，张甜，陈润斌. 基于轻量级卷积Transformer的图像分类方法及在遥感图像分类中的应用[J/OL]. 电子与信息学报：1-9[2022-12-22]. http://kns.cnki.net/kcms/detail/11.4494.TN.20220705.1638.014.html.
Chen H，Zhang T，Chen R B. The image classification method based on a lightweight convolution Transformer and its application in remote sensing image classification[J/OL]. Journal of Electronics and Information Technology：1-9[2022-12-22]. http://kns.cnki.net/kcms/detail/11.4494.TN.20220705.1638.014.html.

[19] Zhao X D，Zhang M M，Tao R，et al. Fractional Fourier image transformer for multimodal remote sensing data classification[J]. IEEE Transactions on Neural Networks and Learning Systems，2022：1-13.

[20] Hou Q B，Zhou D Q，Feng J S. Coordinate attention for efficient mobile network design[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)，June 20-25，2021，Nashville，TN，USA. New York：IEEE Press，2021：13708-13717.

[21] Hendricks D，Gimpel K. Gaussian error linear units (GELUs)[EB/OL]. (2016-06-27)[2022-10-08]. https://arxiv.org/abs/1606.08415.

[22] Shaw P，Uszkoreit J，Vaswani A. Self-attention with relative position representations[EB/OL]. (2018-03-06)[2022-10-08]. https://arxiv.org/abs/1803.02155.

[23] 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.

[24] Cheng G，Li Z P，Yao X W，et al. Remote sensing image scene classification using a bag of convolutional features[J]. IEEE Geoscience and Remote Sensing Letters，2017，14(10)：1735-1739.

[25] Hou D Y，Miao Z L，Xing H Q，et al. Two novel benchmark datasets from ArcGIS and Bing world imagery for remote sensing image retrieval[J]. International Journal of Remote Sensing，2021，42(1)：240-258.

[26] 徐从安，吕亚飞，张筱晗，等. 基于双重注意力机制的遥感图像场景分类特征表示方法[J]. 电子与信息学报，2021，43(3)：683-691.
Xu C A，Lü Y F，Zhang X H，et al. A discriminative feature representation method based on a dual attention mechanism for remote sensing image scene classification[J]. Journal of Electronics & Information Technology，2021，43(3)：683-691.

[27] Yuan Y，Fang J，Lu X Q，et al. Remote sensing image scene classification using rearranged local features[J]. IEEE Transactions on Geoscience and Remote Sensing，2019，57(3)：1779-1792.

[28] Xiong W，Lü Y F，Cui Y Q，et al. A discriminative feature- learning approach for remote- sensing image retrieval[J]. Remote Sensing，2019，11(3)：281.

[29] Selvaraju R，Cogswell M，Das A，et al. Grad-CAM：visual explanations from deep networks via gradient-based localization[C]//2017 IEEE International Conference on Computer Vision (ICCV)，October 22-29，2017，Venice，Italy. New York：IEEE Press，2017：618-626.