激光与光电子学进展 先进成像

融合字典学习与视觉转换器的高分遥感影像场景分类方法

何晓军[1]，刘璇[1,2*]，魏宪[2]

[1]辽宁工程技术大学软件学院，辽宁 葫芦岛 125105；

[2]中国科学院福建物质结构研究所泉州装备制造研究中心，福建 泉州 362216

摘要 遥感影像场景分类方法多基于传统机器学习或卷积神经网络，此类方法的特征提取能力极为有限，尤其在处理类间相似度大、空间信息复杂、几何结构繁多的光学遥感影像时更容易出现特征信息丢失、分类精度受限等问题。基于此，提出一种融合字典学习与视觉转换器(ViT)的高分辨率遥感影像场景分类方法。该方法不仅能够挖掘图像内部的长距离依赖关系，而且可以利用字典学习抓取图像的深层非线性结构信息，从而达到提升分类准确度的目的。在PyTorch深度学习框架上，在RSSCN7、NWPU-RESISC45和Aerial Image Data Set(AID)3个公开的遥感影像数据集上对所提方法和模型进行了广泛实验，验证了所提方法的可行性，其分类正确率比原始视觉转换器模型分别高出1.763个百分点、1.321个百分点和3.704个百分点。与其他先进的场景分类方法相比，所提方法实现了更加优异的分类性能。

关键词 视觉转换器；字典学习；遥感场景分类；高分辨率遥感影像

中图分类号 TP753 **文献标志码** A **DOI**：10.3788/LOP222166

Classification Method of High-Resolution Remote Sensing Scene Image Based on Dictionary Learning and Vision Transformer

He Xiaojun[1], Liu Xuan[1,2*], Wei Xian[2]

[1]*College of Software, Liaoning Technical University, Huludao* 125105, *Liaoning, China*;

[2]*Quanzhou Institute of Equipment Manufacturing Haixi Institutes, Fujian Institute of Research on the Structure, Chinese Academy of Sciences, Quanzhou* 362216, *Fujian, China*

Abstract Classification methods of remote sensing scene images are mostly based on traditional machine learning or convolutional neural networks. The feature extraction capability of such methods is extremely limited, particularly for optical remote sensing images with large interclass similarity, complex spatial information, and various geometric structures, there are problems such as loss of feature information and low classification accuracy. To overcome these problems, we propose a high-resolution remote sensing scene image classification method that combines dictionary learning and Vision Transformer (ViT). This method can not only mine the long-distance dependencies inside the images but can also use dictionary learning to capture the deep nonlinear structural information of images to improve classification accuracy. Through extensive experiments performed on the RSSCN7, NWPU-RESISC45, and Aerial Image Data Set (AID) public remote sensing image datasets trained from scratch on the PyTorch deep learning framework, the effectiveness of the proposed method is verified; the results show that the classification accuracy of the proposed method for the mentioned datasets is 1.763 percentage points, 1.321 percentage points, and 3.704 percentage points higher than that of the original visual converter model, respectively. Moreover, the proposed method outperforms other advanced scene classification methods.

Key words Vision Transformer; dictionary learning; remote sensing image scene classification; high-resolution remote sensing image

收稿日期：2022-07-26；**修回日期**：2022-08-28；**录用日期**：2022-09-27；**网络首发日期**：2022-10-07

基金项目：国家自然科学基金(41801368)、辽宁省教育厅科学研究项目(LJKZ0350)、辽宁省教育厅重点项目(LJ2020ZD003)、福建省科技计划项目(2021T3003，2021T3068)、泉州市科技计划项目(2021C065L)

通信作者：*preciousisgfc@163.com

1 引 言

随着空间技术和传感器技术的飞速发展，遥感影像所包含的纹理细节信息能够被表达得更加清晰，空间细节信息也更加丰富和精细[1]，这使得其空间分辨率越来越高甚至可以达到亚米级[2-3]。然而，空间分辨率越来越高的遥感影像不可避免地包含了越来越多的冗余信息。如何在其中提取到有效特征信息，并进行适度表达以完成分类任务，是当前高分辨率遥感影像场景分类工作亟待解决的问题。

遥感影像场景分类是根据遥感影像的具体内容对其自动划分一个特定的语义标签的过程，这与传统图像分类类似。但高分辨率遥感图像的空间分辨率在几十米以下，其场景组成具有所含地物多样性较高、类内光谱方差差距大、空间分布复杂度较高、类间方差差距小等特点，使得分类工作难度更大。传统的遥感影像场景分类方法可分为有监督和无监督两大类[4]。有监督的分类方法大多是基于数理统计的，主要包括计算量较少但分类精度不高的最小距离分类法[5]、容易造成过度分类且计算速度较慢的混合距离分类法[6]、计算量较大且对训练样本分布要求较高的最大似然分类法[7]等。无监督的分类方法包括对 K 值的选取和异常点都十分敏感的 K-均值分类[8]、在 K-均值分类基础上增加对聚类结果进行合并和分裂两个操作但仍需指定大量参数的 ISODATA 分类方法[9]等。由此可见，这类方法人为因素对其影响较大。另外，传统非深度学习算法的泛化能力太低，分类精度也低于人工目视解译。基于此，人们将深度学习和机器学习引入遥感影像场景分类领域，如将支持向量机、随机森林和稀疏表示等机器学习算法应用到遥感影像分类中，虽然这类算法的分类精度均优于传统方法，但都难以从复杂的高分辨率遥感影像中学习到有效的特征信息，从而无法获得令人满意的分类效果。因此，探究一种分类准确率更高、效果更好的分类模型成为研究热点。

近年来，由于计算力的提升以及卷积神经网络的出现，人工智能、深度学习领域得到了飞速发展，基于卷积神经网络的方法在许多应用上取得了成功，在图像处理领域尤为明显。正因卷积神经网络模型在图像处理领域的出色表现，其被越来越多地应用到高分辨率遥感影像场景分类中[10]，但是卷积神经网络采用分层的数据提取方式，从浅层逐步深入提取高级语义特征，高层数据特征表示高度依赖于底层特征。又由于核心操作即卷积核操作具有平移不变性，这种操作可以捕捉空间信息，但缺少对全局信息的理解，不能充分地利用全局信息，也无法建立特征之间的依赖关系[11]。因此单纯的卷积神经网络模型考虑不到像素级的语义分类问题，使得模型对高分辨率图像的局部信息提取不够敏感，造成了大量的信息冗余。同时，由于固定的卷积核提取到的图像特征通常固定不变，会出现梯度消失或者网格退化等现象。基于前述问题，Shelhamer等[12]在2015年提出了完全卷积网络（FCN），将全连接层改成卷积层后再增加一层反卷积层，从而实现了像素级的语义分类。2016年，Maggiori等[13]将FCN应用于遥感影像场景分类中。2018年，Zhu等[14]构建了一种通过结合卷积神经网络和稀疏矩阵来描述高分辨率遥感图像特征信息的自适应深度稀疏语义模型（ADSSM），该模型取得了较好的性能。

上述研究始终不能真正地解决卷积操作提取特征不充分及其特征信息因层而异造成的级联问题[15]。因此，受到自然语言处理领域相关研究的启发，有研究人员尝试将“转换器（Transformer）”模型[16]迁移到计算机视觉任务。相较于卷积神经网络，“转换器”模型的自注意力机制可以做到并行计算，深层特征的表达不再受浅层特征的影响，且可以较好地挖掘全局信息之间的依赖关系，可以根据不同任务选择不同的归纳偏置，能够有效降低噪声在特征提取中的影响，已经在诸多视觉任务中[17-20]取得了良好的效果。然而，从头训练的视觉转换器[21]与卷积神经网络相比，性能往往较差，原因有两点：首先，利用简单标记化的输入图像后无法对类似于边缘的相邻像素之间的局部结构进行建模，导致训练样本效率低；其次，因为视觉转换器的自注意力机制具有冗余性，在有限的计算资源和训练样本下得到的特征丰富度受到限制[22]。基于此，本文提出了一种将字典学习与视觉转换器融合以提升网络中注意力机制提取特征丰富度的算法，称之为融合字典学习与视觉转换器的高分遥感影像分类方法。

2 理论基础

2.1 视觉转换器

Dosovitskiy等[21]首次将自然语言处理领域的转换器（Transformer）模型应用于图像分类任务，并将其命名为视觉转换器（Vision Transformer），这种网络架构完全抛弃卷积操作而只采取自注意力机制操作，在大规模数据集分类测试中取得良好的效果。

研究人员为了在图像分类任务中使用转换器结构，把图像数据类比文字数据，将输入图片分为多个图像块；再将图像块展平为固定长度的向量；之后对向量进行线性投影变换后加入一个特殊的标志（对应最后的类别预测），得到输入序列；最后将序列传入转换器架构中。视觉转换器与自然语言处理领域的转换器模型的编码器部分中对应部分结构是完全一致的，因此需要深入理解转换器的编码器及其核心多头自注意力机制部分。视觉转换器结构由多个编码器组成，每层编码器的输入都是前一个编码器的输出，编码器由多头注意力层和前馈连接层两个子层构成，每个子层后面是跳跃连接层和层归一化。

其中多头自注意力模块采用的注意力机制是缩放

点积注意力，输入包含Query矩阵（$\boldsymbol{Q}$）、Key矩阵（$\boldsymbol{K}$）、Value矩阵（$\boldsymbol{V}$）。3个矩阵分别由矩阵$\boldsymbol{X}$产生，表达式为

$$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}=\text{embedding}(\boldsymbol{X})。\tag{1}$$

评分函数采用2015年Luong等[23]提出的点积注意力，具体计算公式为

$$\text{attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}。\tag{2}$$

相较于一般的注意力，这种缩放点积注意力在实际应用中会更快，更节省空间，其多头注意力的拼接的表达式为

$$\text{multihead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Concat}(h_{\text{head }1},\cdots,h_{\text{head }h})\boldsymbol{W}^{h}。\tag{3}$$

每个“头”的计算公式为

$$h_{\text{head }i}=\text{attention}(\boldsymbol{Q}\boldsymbol{W}_i^{\boldsymbol{Q}},\boldsymbol{K}\boldsymbol{W}_i^{\boldsymbol{K}},\boldsymbol{V}\boldsymbol{W}_i^{\boldsymbol{V}}),\tag{4}$$

式中：$\boldsymbol{Q}=\boldsymbol{I}\times\boldsymbol{W}_i^{\boldsymbol{Q}}$，$\boldsymbol{K}=\boldsymbol{I}\times\boldsymbol{W}_i^{\boldsymbol{K}}$，$\boldsymbol{V}=\boldsymbol{I}\times\boldsymbol{W}_i^{\boldsymbol{V}}$，$\boldsymbol{W}_i^{\boldsymbol{Q}},\boldsymbol{W}_i^{\boldsymbol{K}}\in\mathbb{R}^{D\times d_u}$，$\boldsymbol{W}_i^{\boldsymbol{V}}\in\mathbb{R}^{D\times d_e}$，$\boldsymbol{I}$为输入图像，$D$为所有参数矩阵的第一维度，$d_u$和$d_e$为参数矩阵的第二维度。

2.2 稀疏表示字典学习

任意一个信号都可以在一个过完备字典上被稀疏线性表出，因此一个信号被分解为有限信号的线性组合的形式，称为稀疏表示[24]。其形式化表达式为

$$\boldsymbol{Y}=\boldsymbol{D}\boldsymbol{X},\tag{5}$$

式中：$\boldsymbol{D}=[d_1,d_2,\cdots,d_K]\in\mathbb{R}^{m\times K}$，表示通过字典学习得到的具有$K$列原子的字典；$\boldsymbol{X}=[x_1,x_2,\cdots,x_k]\in\mathbb{R}^{K\times n}$，表示$\boldsymbol{Y}$中的样本稀疏编码是由$n$个$K$维列向量组成的。如图1所示，字典学习主要是将原始样本$\boldsymbol{Y}$给分解中字典矩阵$\boldsymbol{D}$和稀疏码矩阵$\boldsymbol{X}$的过程，这里的稀疏码$\boldsymbol{X}$自然是十分稀疏的，数据量比较少，字典$\boldsymbol{D}$存储了原始样本$\boldsymbol{Y}$中的特征。

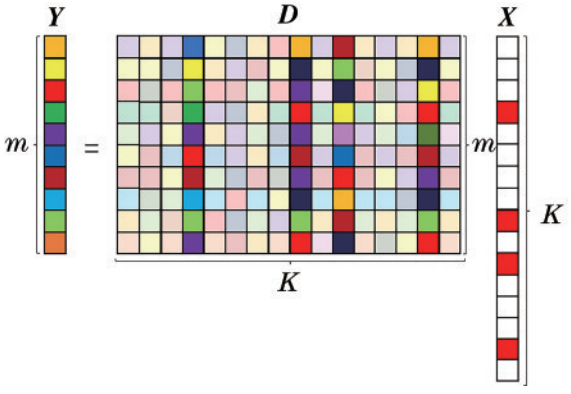

图1 字典学习示意图

Fig. 1 Diagram of dictionary learning

稀疏表示具有很好的图像表达能力，而字典学习作为一种十分有效且鲁棒性较好的稀疏表示算法已经成功应用于多种不同任务特征表示和选择方法。

一般情况下，给定数据$\boldsymbol{Y}=[y_1,y_2,\cdots,y_d]\in\mathbb{R}^{n\times d}$，字典学习的目标任务是寻找字典$\boldsymbol{D}=[d_1,d_2,\cdots,d_d]\in\mathbb{R}^{n\times d}$及其对应的稀疏矩阵$\boldsymbol{X}=[x_1,x_2,\cdots,x_d]\in\mathbb{R}^{d\times d}$，使得每个数据样本能够更好地被字典重构。若直接使用样本作为字典则会造成字典集过于庞大，会使数据冗余，产生噪声，导致重构学习效率低下[25]，因此可将字典学习表述为优化问题。

1）通过L_1-范数最小化获得最佳的稀疏编码和过完备字典：

$$\{\hat{\boldsymbol{D}},\hat{\boldsymbol{X}}\}=\arg\min_{\boldsymbol{D},\boldsymbol{X}}\left\{\|\boldsymbol{Y}-\boldsymbol{D}\boldsymbol{X}\|_F^2+\lambda\|\boldsymbol{X}\|_1\right\}。\tag{6}$$

因为L_0-范数最小化为非凸的NP-hard问题，而L_1-范数最小化是L_0-范数的最紧的凸松弛问题，且L_1-范数的解往往是稀疏性的最优解，所以可以通过L_1-范数去逼近最优解。

2）通过最小化重构误差获得最佳的稀疏编码和过完备字典：

$$\{\hat{\boldsymbol{D}},\hat{\boldsymbol{X}}\}=\arg\min_{\boldsymbol{D},\boldsymbol{X}}\left\{\|\boldsymbol{Y}-\boldsymbol{D}\boldsymbol{X}\|_2\right\}。\tag{7}$$

通过最小化重构误差，学习到的字典和稀疏编码的乘积与原始数据的差距最小，从而得到最佳的稀疏编码和过完备字典。

3 融合字典学习与视觉转换器的高分遥感影像分类模型

图2展示了所提基于视觉转换器架构的方法的整体流程。首先，类似Transformer处理流程，需要先对图像数据按照一定的大小进行分块；然后对分块图像信息进行线性变换后将结果映射到一维向量中，在将结果输入到编码器之前还需要嵌入分类信息及其空间位置信息，其中每个编码器由层归一化、多头注意力机制、多层感知机构成；最后输出的特征即为用于场景分类的特征。

值得注意的是，此处的多头注意力机制不是原始视觉转换器模型的自注意力机制，而是基于字典学习的注意力机制。在从高维输入信号到用于分类的低维特征的降维过程中，期望能够保留一些重要的几何特征，同时又能够从低维数据中恢复原始信号[26]，因此利用字典学习算法对高分遥感图像数据进行降维，获取遥感图像的空间和通道特征信息；随后使用视觉转换器架构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[27]。

3.1 基本网络架构

为了更好地读取到遥感场景图像中复杂的全局信息，采用视觉转换器架构，此架构只采用了Transformer中的编码层。此架构有5个关键子模块，分别是嵌入层（embedding）、注意力层、连接操作、归一化连接结构及多层感知机。其中，本文采用的注意力层是基于稀疏字典学习的注意力模块。

1）嵌入层

根据自然语言处理领域采用的Transformer概念

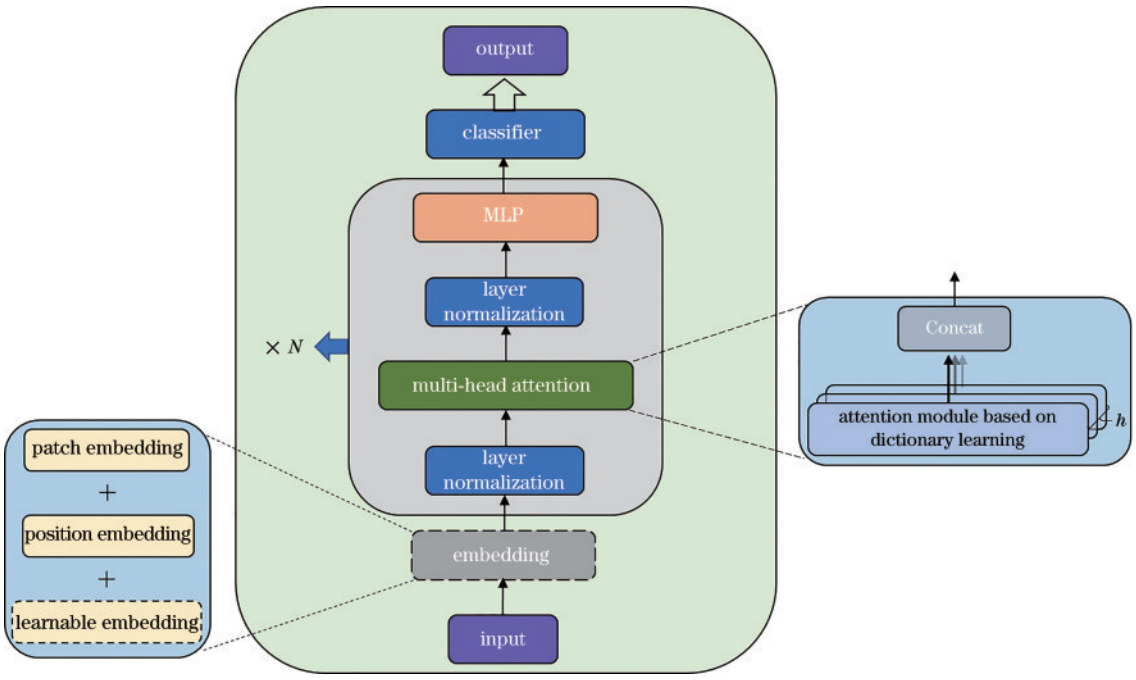

图 2　所提方法流程

Fig. 2　Flowchart of the proposed method

和 Dosovitskiy 等[21]的建议，一张图像可以被重新映射为一连串扁平化的二维图像块 $\boldsymbol{x}_p \in \mathbb{R}^{H\times(P^2\times C)}$。在 Transformer 中输入矩阵形状为 (N, D)，N 代表序列长度，D 代表序列中每个向量的维度。因此在得到二维图像块之后，还需要对每个维度为 $P^2\times C$ 的图像块进行一个线性变换（全连接层），将维度压缩为 D。此外，在分类任务中还需要加上一个可学习的嵌入向量，作为类别信息和位置编码，因此，适当的表达形式为

$$z_0 = \left[x_{\text{class}}; x_p^1 E; x_p^2 E; x_p^3 E; \cdots; x_p^N E\right] + E_{\text{pos}}, \quad (8)$$

式中：E 代表线性变换层；$P^2\times C$ 为输入维度；D 是输出维度。可训练变量 E_{pos} 用于表示添加序列的位置信息。当位置接近时，它们往往有相似的编码，同一行/列的补丁也有相似的位置编码。

位置编码一般包括正余弦位置编码、学习位置向量和相对位置表达等方式。本架构所采用的是可学习位置向量。因为 Transformer 中没有时间步的概念，所以将偏置向量当作位置向量，从每一个图像块的位置学得一个独立的向量。在代码中用 nn. Parameter 实现可学习的过程。

2）连接操作

连接操作（Concat）为多头注意力层所特有的操作，目的是增强网络架构捕捉不同结构特征信息的能力。每个注意力模块的输出结果为一个“头（head）”，因此多头注意力中需要连接操作。若有 h 个头，则可表示为

$$h(\text{attention}) = \text{Concat}(h_{\text{head }1}, h_{\text{head }2}, \cdots, h_{\text{head }h})。 \quad (9)$$

3）层归一化

随后对整个多头注意力模块计算得到的结构进行层归一化操作（layer normalization）。区别于批量归一化操作（batch normalization），层归一化操作对某一层的所有神经元进行归一化，不受 batchsize 大小的影响，二者区别如图 3 所示。

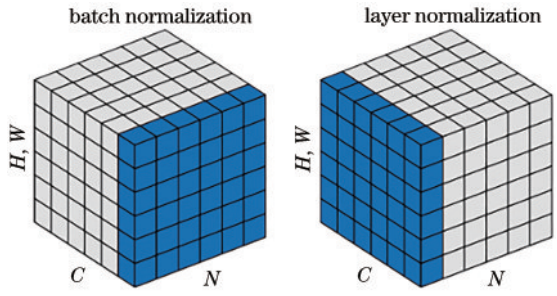

图 3　批量归一化和层归一化

Fig. 3　Batch normalization and layer normalization

假设某层有 M 个神经元，则该层的输入为 $\{z_1^l, z_2^l, \cdots, z_M^l\}$，其均值为 $\mu = \frac{1}{M}\sum_{m=1}^{M} z_m^l$，方差为 $\sigma^2 = \frac{1}{M}\sum_{m=1}^{M}\left(z_m^l - \mu\right)^2$。归一化公式可表示为

$$\widehat{z_m^l} = \frac{z_m^l - \mu}{\sqrt{\sigma^2 + \epsilon}} * \gamma + \beta, \quad (10)$$

式中：γ 和 β 分别代表缩放和平移的参数，作用是使归一化过程不影响网络的表示能力。

4）多层感知机

多层感知机主要由输入层、隐藏层和输出层构成。这是一种连接方式比较简单的前馈神经网络结构，多层感知机的作用主要是模拟复杂非线性函数功能。隐

藏层中可以采用不同的激活函数，使得模型具有非线性功能。输入层不承担函数处理功能，只有隐藏层和输出层会对数据进行加工处理。

在多层感知机中，相邻层之间的神经元通常通过全连接的方式进行连接，因此也被称为全连接层。每个隐藏层的神经元数量是可以变化的，有更多的神经元意味着会有更好的拟合能力，但同时也会更容易造成模型过拟合，多层感知机可以应用于几乎所有任务的多功能学习方法，包括分类、回归甚至是无监督学习，结构如图 4 所示。正则化方法能很好地解决过拟合问题，L_1-正则化可以使特征矩阵一部分数据系数缩小到 0，从而间接实现特征选择，这种情况对应字典学习的第一种求解思路；L_2-正则化可以使所有特征系数都缩小，但不会缩小至 0，这种操作会使优化求解过程稳定快速，这种情况与字典学习的第二种求解思路目标一致。因此所提方法可以很好地规避网络过拟合的状况。

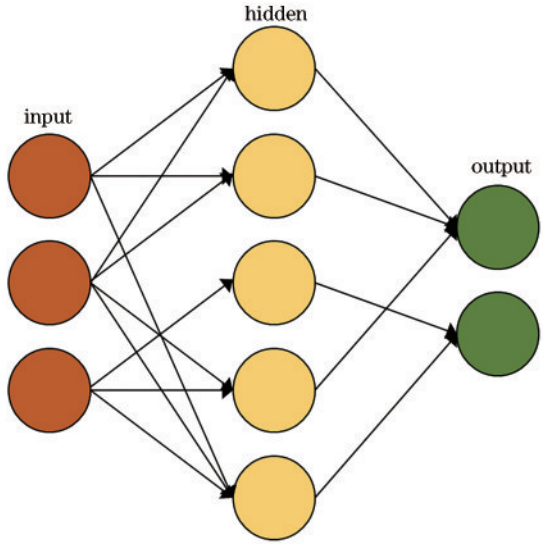

图 4　多层感知机示意图

Fig. 4　Schematic of multilayer perceptron

3.2　基于稀疏字典学习的注意力模块

不管在自然中还是在工业领域，通过诸如生物、人工等传感器（人眼、人耳、摄像头等）能够捕获到大规模的高维图像数据信息。但是，这些数据存在着较大的冗余信息，对这些数据进行直接处理可能会出现计算困难、计算周期长等问题。

高维数据因数据本身内部特征的限制，往往会产生维度上的冗余，且高维空间中的流形在局部具有欧氏空间的性质[28]，因此只需要比较少维度的信息就能对数据进行唯一标识。为了能够高效地提取到高维图像数据的有用信息，降维方法被应用到各个领域。而字典学习的本质也是一种降维方法，因此本文用流形优化和字典学习的方法。流形优化是一种从高维数据中探测出低维流形结构的方法，流形是一个局部具有欧几里得空间性质的空间。所提基于字典学习的注意力机制模块采用了结合流形优化和字典学习的降维方法，使得在对高维图像数据进行降维操作时，可以尽可能地保留数据的结构特征，方法处理流程如图 5 所示。首先将待处理的高维数据 $\boldsymbol{X}=[\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N]\in\mathbb{R}^{N\times S}$ 约束在 Stiefel 流形上，其流形空间记为 $\mathcal{M}$，降维后的数据为 $\boldsymbol{x}=[\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n]\in\mathbb{R}^{n\times s}$。随后求过完备字典以及其对应的稀疏表示，字典学习的目标函数为

$$\min_{\boldsymbol{D},\boldsymbol{\Phi}}\sum_{i=1}^{s}\frac{1}{2}\left|\boldsymbol{x}_i-\boldsymbol{D}\boldsymbol{\phi}_i\right|_2^2+g\left(\boldsymbol{\phi}_i\right),\ \text{s.t.}\ \alpha h(\boldsymbol{D}),\tag{11}$$

式中：$h(\boldsymbol{D})=\sum_{i\neq j}\left\|\boldsymbol{d}_i^{\mathrm{T}}\boldsymbol{d}_j\right\|_F^2$ 表示字典的正则化项。此外，字典集是 s 与 $K-1$ 维单位球体的乘积流形，$\boldsymbol{D}\in\mathcal{S}(n,k)=\{\boldsymbol{D}\in\mathbb{R}^{n\times k}:\mathrm{diag}(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D})=\boldsymbol{I}_k\}$，即 $\mathcal{S}(n,k)$ 约束所有 $\boldsymbol{d}_i\in\mathbb{R}^k$ 都有单位范数。$g\left(\boldsymbol{\phi}_i\right)$ 通常利用范数约束来控制稀疏编码 $\boldsymbol{\phi}_i$ 的稀疏性，L_1-范数和 L_2-范数正则化分别称岭回归和 Lasso 回归。弹性网络很好地结合了岭回归的稳定性和 Lasso 回归的稀疏性。因此，可以得到 $g\left(\boldsymbol{\phi}_i\right)=\lambda\left\|\boldsymbol{\phi}_i\right\|_1+\frac{\beta}{2}\left\|\boldsymbol{\phi}_i\right\|_2^2$，稀疏编码的闭式解的形式为

$$\boldsymbol{\phi}_i^*=\left(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D}+\lambda\boldsymbol{I}\right)^{-1}\left(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{x}_i\right)。\tag{12}$$

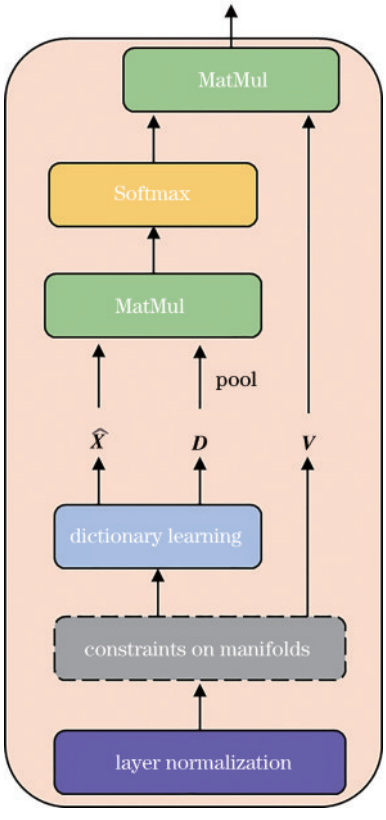

图 5　注意力模块方法的流程

Fig. 5　Flowchart of attention module method

通过初始化不同的字典 $\boldsymbol{D}$，得到不同的稀疏编码矩阵 $\boldsymbol{\Phi}_1$ 和 $\boldsymbol{\Phi}_2$，之后分别对 $\boldsymbol{\Phi}_1$ 和 $\boldsymbol{\Phi}_2$ 进行层归一化和平均池化操作，得到矩阵 $\boldsymbol{Q}$ 和矩阵 $\boldsymbol{K}$，计算公式分别为

$$\boldsymbol{Q}=\text{layernorm}(\boldsymbol{\Phi}_1), \tag{13}$$

$$\boldsymbol{K}=\text{AvgPool2d}(\boldsymbol{\Phi}_2)。 \tag{14}$$

随后对矩阵 $\boldsymbol{Q}$ 和矩阵 $\boldsymbol{K}$ 进行矩阵相乘操作，最后进行 Softmax 操作，得到所需特征信息：

$$\boldsymbol{D}_{\text{Attn}}=\text{Softmax}(\boldsymbol{QK}^{\text{T}})。 \tag{15}$$

所提方法在注意力模块中应用字典学习来重建和发现关键特征区域，引入字典和稀疏编码的转换后可以通过重建矩阵将空间注意力和通道注意力转移到模型上，和原始的自注意力机制相比，去除了较多冗余信息，得到更多有利于分类的特征信息，操作的具体流程如图 6 所示。

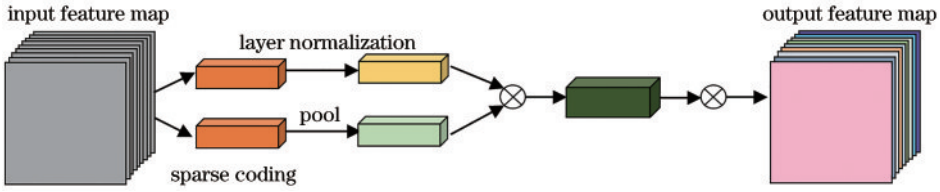

图 6　基于字典学习的注意力模块

Fig. 6　Attention module based on dictionary learning

4　实验结果及分析

4.1　数据集

为了在不同环境下测试所提算法的有效性，实验分别使用了 3 种不同的公开遥感场景图像数据集 RSSCN7、NWPU-RESISC45 和 Aerial Image Data Set（AID）。所选择的 3 个数据集都具有较高分类难度的遥感影像场景，3 个数据集的规模大小、类别信息以及图像大小等均不相同。关于数据集的简易信息，如表 1 所示。

表 1　数据集简介

Table 1　Introduction of datasets

Dataset	Number of scene classes	Number of total images	Image size	Spatial resolution /m	Year
RSSCN7	7	2800	400×400		2015
NWPU-RESISC45	45	31500	256×256	~30–0.2	2016
AID	30	10000	600×600	~8–0.5	2017

RSSCN7 数据集[29]图像来自于谷歌地球，共计包含 2800 幅遥感影像，均匀分布在 7 个场景类中，每个类包含 400 张基于 4 种不同尺度采样的像素大小为 400×400 的样本。这些图像的采样环境分别处于不同的季节以及天气条件下，并且采用了不同的比例进行最终选择，因此场景图像的多样性较复杂，致使分类工作具有较大的挑战性。各类遥感影像场景如图 7 所示。

图 7　RSSCN7 数据集

Fig. 7　RSSCN7 dataset

NWPU-RESISC45 数据集[30]是由西北工业大学（NWPU）创建，用于遥感影像场景分类的公开数据集，图像同样来自于谷歌地球。该数据集有 31500 张图像，覆盖了 45 个不同的场景类别，每个场景类别包含 700 张从光照、视角、背景、遮挡以及空间分辨率不同角度采集的场景图像，图像大小为 256×256，空间分辨率为 0.2～30 m。此数据集在场景类的数量和图像总数上都是三个所选数据集中最大的，具有规模大、图像丰富、类内多样性丰富、类间相似度高等特征，场景分类难度较大。各类遥感影像场景如图 8 所示。

Aerial Image Data Set（AID）[31]也是一个从谷歌地球图像中收集到的样本组成的数据集，这是一个大尺度的航空影像数据集。数据集包含 30 个场景类型，共计 10000 张图像。不同类别的图像数量相差较大，从 220 张到 420 张不等。每个图像的尺寸固定为 600×600，空间分辨率为 0.5～8 m。此外，每个类别的图像都是在不同的成像条件、不同时间和季节采集的，导致类内差异很大。各类遥感影像场景如图 9 所示。

图 8　NWPU-RESISC45 数据集

Fig. 8　NWPU-RESISC45 dataset

图 9　AID 数据集

Fig. 9　AID dataset

4.2　实验设置

本研究的所有实验都基于 PyTorch 深度学习框架，使用 NVIDIA TITAN Xp 图形处理器，运行频率达 1.6 GHz，显存为 12G，CPU 为 Intel(R) Xeon(R) Silver 4210 CPU@2.20 GHz。使用分类正确率(accuracy)、召回率(recall)、精准率(precision)、F1 值(F1)作为遥感影像场景分类的评价指标。具体实验配置如表 2 所示。

表 2　实验环境

Table 2　Laboratory environment

Laboratory environment	Environment configuration
Language	Python3.8.6
Tool	PyCharm11.0.11
Framework	PyTorch1.9.1
CUDA	10.2

4.3　分类正确率对比

为了验证所提算法的有效性，首先使用 RSSCN7 数据集，将图片按 7∶3 划分为训练集和测试集。先使用训练集对所提算法进行训练，再在测试集测试分类算法的分类效果。由于基于自注意力机制的视觉转换器(ViT)模型在大型数据集上对设备要求较高，因此所提模型和原始基于自注意力机制的视觉转换器模型均是未采用迁移学习、未经过预训练、从头开始训练的网络模型。

为了验证所提算法的有效性，实验设置与文献[32]相同，比较了原始 ViT 网络、Transformer in Transformer(TNT)[17]以及现有经典的卷积网络 AlexNet、VGG、ResNet。分类结果如表 3 所示，由表 3 可知：原始 ViT 网络的分类正确率比经典卷积网络和 TNT 高；所提算法在 RSSCN7 数据集上的表现均优于其他网络架构，总体分类正确率最高，达 91.406%。

已经在 RSSCN7 数据集上证明了所提算法的优越性，为进一步证明所提算法的有效性，在 NWPU-RESISC45 数据集上再进行实验。NWPU-RESISC45 数据集的场景类别数上升到 45，并且图片总数达 31500 张，增加了分类难度，并且与文献[30]一致，都采用 2∶8 的比例划分数据集。文献[30]中的网络模型

表 3 不同网络在 RSSCN7 数据集上的分类正确率

Table 3 Accuracy of different networks on RSSCN7 dataset

Network	Accuracy /%
AlexNet	82.230
VGG	80.833
ResNet50	89.048
TNT	84.833
ViT	89.643
Proposed network	91.406

均是在 ImageNet 上预训练后经过微调的 AlexNet、VGG、GoogLeNet 等，而所提算法是未采用迁移学习的网络模型。实验还采用了从头开始训练的原始 ViT 网络进行消融实验，并且采用 TNT 网络进行对比实验，分类结果如表 4 所示。由表 4 可知：原始 ViT 网络性能优于经过微调的 AlexNet、GoogLeNet、TNT，分类正确率仅仅比经过微调后的 VGG 网络差 0.105 个百分点；所提算法在 NWPU-RESISC45 数据集上的表现均优于其他算法，总体分类正确率最高，达 91.576%。

表 4 不同网络在 NWPU-RESISC45 数据集上的分类正确率

Table 4 Accuracy of different networks on NWPU-RESISC45 dataset

Network	Accuracy /%
Fine-tuned AlexNet	85.160
Fine-tuned VGGNet-16	90.360
Fine-tuned GoogLeNet	86.020
TNT	85.031
ViT	90.255
Proposed network	91.576

为了进一步验证所提算法的有效性，实验又选取了 AID 数据集。相较于 NWPU-RESISC45 数据集，AID 数据集中的空间分辨率精度更高，而且图片尺寸也更大，但单类场景图像样本较少。本实验设置与文献[31]一致，并且选取其中 CaffeNet、VGG-VD-16、ResNet152、GoogLeNet 网络模型以及原始 ViT 网络和 TNT 网络作为对比，实验结果如表 5 所示。在三种不同的数据集上都验证了所提算法的优越性。

表 5 不同网络在 AID 数据集的分类正确率

Table 5 Accuracy of different networks on AID dataset

Network	Accuracy /%
CaffeNet	86.860
VGG-VD-16	86.590
ResNet152	89.130
GoogLeNet	83.440
TNT	80.450
ViT	85.514
Proposed network	89.218

4.4 消融实验

所提方法是基于视觉转换器架构的，原始的视觉转换器架构是基于自注意力机制的。将自注意力机制替换成基于字典学习的注意力机制，为了验证所提算法的有效性，就是否采用基于字典学习的注意力机制在三种不同的公开数据集 RSSCN7、NWPU-RESISC45 和 AID 上进行消融实验，采用 kappa 系数、F1 系数、召回率(recall)及精准度(precision)4 个分类指标对结果进行分析，具体数据如表 6 所示。通过表 6 可知，在未经过预训练的情况下，所提模型在 3 个数据集上的表现均优于原始 ViT 模型。其中，相较于 RSSCN7、NWPU-RESISC45 数据集，两个模型在 AID 数据集上的表现均有降低，原因在于 AID 数据集的单类样本量较少，遥感影像场景种类较多，在进行分类任务时具有一定的难度。

表 6 两种方法在三个数据集上的参数指标

Table 6 Parameter indicators of two methods on three datasets

Parameter	RSSCN7		NWPU-RESISC45		AID	
	ViT	Proposed method	ViT	Proposed method	ViT	Proposed method
kappa	0.900	0.916	0.934	0.947	0.883	0.909
F1	86.222	90.890	88.927	90.207	84.202	87.768
recall	85.986	91.142	88.984	90.286	84.147	87.662
precision	86.417	91.002	89.039	90.317	84.558	88.004

4.5 模型参数量对比

在遥感影像场景分类应用中，分类网络的参数量也是需要考虑的重要内容，在算法参数量测试中，以训练 RSSCN7 数据集时网络架构中总的参数量为例，各个分类框架包含的参数量如表 7 所示。从表 7 可知，当用基于字典学习的注意力机制替换原有的自注意力机制时，模型所需要的参数量有了明显降低，这有效减少了显存的消耗并且大大降低了计算量，节省训练成本。实验结果证实了所提融合字典学习和视觉转换器的高分遥感影像分类模型在高分遥感影像场景分类方面具有一定的创新性、实用性、高效性。

4.6 鲁棒性测试

为了测试所提方法对高斯噪声干扰的鲁棒性，在 RSSCN7 测试集中加入了不同扰动波幅的高斯噪声，利

表 7 不同分类框架的参数量

Table 7 Parameters of different classification frameworks

Network	Number of parameter /10^6
AlexNet	6
VGG	13.3
ResNet50	2.55
TNT	2.25
ViT	2.6
Proposed method	1.84

用这种方式来破坏图像以进行鲁棒性评估，实验结果如图 10 所示。由图 10 可以看出，在高斯噪声扰动波幅从 10 μV 到 50 μV 变化的过程中，所提模型对受到高斯噪声扰动的图像的分类正确率的变化幅度一直比 ViT 小，其鲁棒性比 ViT 更好，这是因为所提模型能够抓取图像的深层非线性结构信息，所以不易受到噪声干扰。

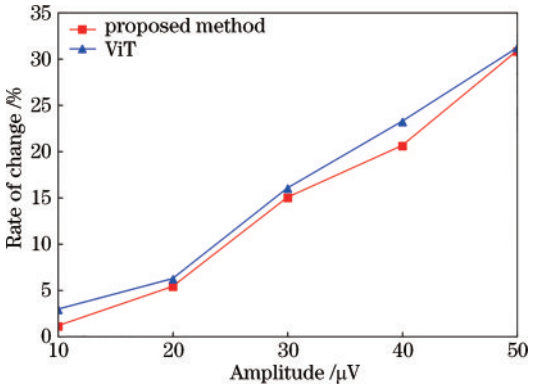

图 10 高斯噪声图像的分类正确度变化率

Fig. 10 Rate of change of classification accuracy on Gaussian noise images

5 结 论

为了解决卷积神经网络处理分类任务时提取特征不充分、卷积操作特有的级联问题以及原始 ViT 中的注意力机制为抓取到输入数据的全局信息造成数据冗余和忽略其中非线性空间关系，从而无法在从头训练的任务中获得具有代表性和鉴别力的特征表达的问题，提出了一个融合字典学习和视觉转换器架构的网络模型。该模型不仅具有 ViT 的全局信息抓取能力，解决了卷积操作的级联问题，并且具有字典学习的非线性特征挖掘能力，解决了特征提取不充分的问题。结合实验结果可知，与自注意力机制相比，基于字典学习的注意力机制可以有效提高网络框架的特征提取能力，从而提高模型的分类识别正确率、kappa 系数、F1 系数、召回率、精准度及鲁棒性。因此，所提算法在遥感影像场景分类领域具有一定的创新性、高效性、可行性和实用性。

参 考 文 献

[1] 孙伟伟，杨刚，陈超，等. 中国地球观测遥感卫星发展现状及文献分析[J]. 遥感学报，2020，24(5)：479-510.
Sun W W，Yang G，Chen C，et al. Development status and literature analysis of China's earth observation remote sensing satellites[J]. Journal of Remote Sensing，2020，24(5)：479-510.

[2] 赵济. 面向高分辨率遥感影像分类的条件随机场模型研究[D]. 武汉：武汉大学，2017.
Zhao J. Conditional random fields for high resolution remote sensing image classification[D]. Wuhan：Wuhan University，2017.

[3] Zhao J，Zhong Y F，Shu H，et al. High-resolution image classification integrating spectral-spatial-location cues by conditional random fields[J]. IEEE Transactions on Image Processing，2016，25(9)：4033-4045.

[4] Li M，Zang S Y，Zhang B，et al. A review of remote sensing image classification techniques：the role of spatio-contextual information[J]. European Journal of Remote Sensing，2014，47(1)：389-411.

[5] Patil M B，Desai C G，Umrikar B N. Image classification tool for land use/land cover analysis：a comparative study of maximum likelihood and minimum distance method[J]. International Journal of Geology，Earth and Environmental Sciences，2012，2(3)：189-196.

[6] 陈斯娅. 基于距离的遥感图像分类方法研究[D]. 长春：东北师范大学，2017.
Chen S Y. Research of remote sensing image classification methods based on distance[D]. Changchun：Northeast Normal University，2017.

[7] Peng J T，Li L Q，Tang Y Y. Maximum likelihood estimation-based joint sparse representation for the classification of hyperspectral remote sensing images[J]. IEEE Transactions on Neural Networks and Learning Systems，2019，30(6)：1790-1802.

[8] Rollet R，Benie G B，Li W，et al. Image classification algorithm based on the RBF neural network and K-means[J]. International Journal of Remote Sensing，1998，19(15)：3003-3009.

[9] Abbas A W，Minallh N，Ahmad N，et al. K-means and ISODATA clustering algorithms for landcover classification using remote sensing[J]. Sindh University Research Journal-SURJ (Science Series)，2016，48(2)：315-318.

[10] 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.

[11] 周飞燕，金林鹏，董军. 卷积神经网络研究综述[J]. 计算机学报，2017，40(6)：1229-1251.
Zhou F Y，Jin L P，Dong J. Review of convolutional neural network[J]. Chinese Journal of Computers，2017，40(6)：1229-1251.

[12] Shelhamer E，Long J，Darrell T. Fully convolutional networks for semantic segmentation[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition

(CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 640-651.

[13] Maggiori E, Tarabalka Y, Charpiat G, et al. Fully convolutional neural networks for remote sensing image classification[C]//2016 IEEE International Geoscience and Remote Sensing Symposium, July 10-15, 2016, Beijing, China. New York: IEEE Press, 2016: 5071-5074.

[14] Zhu Q Q, Zhong Y F, Zhang L P, et al. Adaptive deep sparse semantic modeling framework for high spatial resolution image scene classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2018, 56(10): 6180-6195.

[15] 王友伟, 郭颖, 邵香迎. 基于改进级联算法的遥感图像目标检测[J]. 光学学报, 2022, 42(24): 2428004.
Wang Y W, Guo Y, Shao X Y, et al. Remote sensing images object detection based on improved cascade R-CNN[J]. Acta Optica Sinica, 2022, 42(24): 2428004.

[16] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[EB/OL]. (2017-06-12)[2022-03-05]. https://arxiv.org/abs/1706.03762.

[17] Han K, Xiao A, Wu E H, et al. Transformer in transformer[EB/OL]. (2021-02-27)[2022-03-05]. https://arxiv.org/abs/2103.00112.

[18] Xu K J, Deng P F, Huang H. Vision Transformer: an excellent teacher for Guiding small networks in remote sensing image scene classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 5618715.

[19] Hao S Y, Wu B, Zhao K, et al. Two-stream swin transformer with differentiable Sobel operator for remote sensing image classification[J]. Remote Sensing, 2022, 14(6): 1507.

[20] Lü P Y, Wu W J, Zhong Y F, et al. SCViT: a spatial-channel feature preserving vision transformer for remote sensing image scene classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 4409512.

[21] 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)[2022-03-05]. https://arxiv.org/abs/2010.11929v2.

[22] 刘嘉敏, 郑超, 张丽梅, 等. 基于图像重构特征融合的高光谱图像分类方法[J]. 中国激光, 2021, 48(9): 0910001.
Liu J M, Zheng C, Zhang L M, et al. Hyperspectral image classification method based on image reconstruction feature fusion[J]. Chinese Journal of Lasers, 2021, 48(9): 0910001.

[23] Luong M T, Pham H, Manning C D. Effective approaches to attention-based neural machine translation [EB/OL]. (2015-08-17)[2022-03-06]. https://arxiv.org/abs/1508.04025.

[24] Tošić I, Frossard P. Dictionary learning[J]. IEEE Signal Processing Magazine, 2011, 28(2): 27-38.

[25] Vu T H, Monga V. Fast low-rank shared dictionary learning for image classification[J]. IEEE Transactions on Image Processing, 2017, 26(11): 5160-5175.

[26] 郑思龙, 李元祥, 魏宪, 等. 基于字典学习的非线性降维方法[J]. 自动化学报, 2016, 42(7): 1065-1076.
Zheng S L, Li Y X, Wei X, et al. Nonlinear dimensionality reduction based on dictionary learning[J]. Acta Automatica Sinica, 2016, 42(7): 1065-1076.

[27] 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.
Wang J N, Gao Y, Shi J, 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.

[28] 刘嘉敏, 杨松, 黄鸿. 基于局部重构 Fisher 分析的高光谱遥感影像分类[J]. 中国激光, 2020, 47(7): 0710001.
Liu J M, Yang S, Huang H. Hyperspectral remote sensing image classification based on local reconstruction Fisher analysis[J]. Chinese Journal of Lasers, 2020, 47(7): 0710001.

[29] Zou Q, Ni L H, Zhang T, et al. Deep learning based feature selection for remote sensing scene classification [J]. IEEE Geoscience and Remote Sensing Letters, 2015, 12(11): 2321-2325.

[30] Cheng G, Han J W, Lu X Q. Remote sensing image scene classification: benchmark and state of the art[J]. Proceedings of the IEEE, 2017, 105(10): 1865-1883.

[31] 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.

[32] 李彦甫, 范习健, 杨绪兵, 等. 基于自注意力卷积网络的遥感图像分类[J]. 北京林业大学学报, 2021, 43(10): 81-88.
Li Y F, Fan X J, Yang X B, et al. Remote sensing image classification framework based on self-attention convolutional neural network[J]. Journal of Beijing Forestry University, 2021, 43(10): 81-88.