**中图法分类号**:TP751　　**文献标识码**: A　　**文章编号**: 1006-8961(2024)08-2205-15

**论文引用格式**:Xue J, Huang H, Pu C Y, Yang Y M, Li Y and Liu Y X. 2024. Spatial-spectral model distillation network for hyperspectral scene classification. Journal of Image and Graphics, 29(08):2205-2219(薛洁, 黄鸿, 蒲春宇, 杨鄞铭, 李远, 刘英旭. 2024. 面向高光谱场景分类的空—谱模型蒸馏网络. 中国图象图形学报, 29(08):2205-2219)[DOI:10.11834/jig.230699]

# 面向高光谱场景分类的空—谱模型蒸馏网络

薛洁[1],黄鸿[1*],蒲春宇[2],杨鄞铭[1],李远[1],刘英旭[1]

1. 重庆大学光电技术与系统教育部重点实验室,重庆　210046; 2. 电磁空间安全全国重点实验室,成都　610036

**摘　要**:**目的**　现有场景分类方法主要面向高空间分辨率图像,但这些图像包含极为有限的光谱信息,且现有基于卷积神经网络(convolutional neural network,CNN)的方法由于卷积操作的局部性忽略了远程上下文信息的捕获。针对上述问题,提出了一种面向高光谱场景分类的空—谱模型蒸馏网络(spatial-spectral model distillation network for hyperspectral scene classification,SSMD)。**方法**　选择基于空—谱注意力的ViT方法(spatial-spectral vision Transformer,SSViT)探测不同类别的光谱信息,通过寻找光谱信息之间的差异性对地物进行精细分类。利用知识蒸馏将教师模型SSViT捕获的长距离依赖信息传递给学生模型VGG16(Visual Geometry Group 16)进行学习,二者协同合作,教师模型提取的光谱信息和全局信息与学生模型提取的局部信息融合,进一步提升学生分类性能并保持较低的时间代价。**结果**　实验在3个数据集上与10种分类方法(5种传统CNN分类方法和5种较新场景分类方法)进行了比较。综合考虑时间成本和分类精度,本文方法在不同数据集上取得了不同程度的领先。在OHID-SC(Orbita hyperspectral image scene classification dataset)、OHS-SC(Orbita hyperspectral scene classification dataset)和HSRS-SC(hyperspectral remote sensing dataset for scene classification)数据集上的精度,相比于性能第2的模型,分类精度分别提高了13.1%、2.9%和0.74%。同时在OHID-SC数据集中进行的对比实验表明提出的算法有效提高了高光谱场景分类精度。**结论**　提出的SSMD网络不仅有效利用高光谱数据目标光谱信息,并探索全局与局部间的特征关系,综合了传统模型和深度学习模型的优点,使分类结果更加准确。

**关键词**:高光谱场景分类;卷积神经网络(CNN);Transformer;空—谱联合自注意力机制;知识蒸馏(KD)

# Spatial-spectral model distillation network for hyperspectral scene classification

Xue Jie[1], Huang Hong[1*], Pu Chunyu[2], Yang Yinming[1], Li Yuan[1], Liu Yingxu[1]

*1. Key Laboratory of Optoelectronic Technology and Systems of the Education Ministry of China, Chongqing University, Chongqing　210046, China; 2. National Key Laboratory of Electromagnetic Space Security, Chengdu　610036, China*

**Abstract**: **Objective** In recent years, the development of remote sensing technology has enabled the acquisition of abundant remote sensing images and large datasets. Scene classification tasks, as key areas in remote sensing research, aim to distinguish and classify images with similar scene features by assigning fixed semantic labels to each scene image. Various scene classification methods have been proposed, including handcrafted feature- and deep learning-based methods. However, handcrafted feature-based methods have limitations in describing scene semantic information due to high require-

**收稿日期**:2023-09-28;**修回日期**:2023-12-22;**预印本日期**:2023-12-29

*通信作者:黄鸿　hhuang@cqu.edu.cn

**基金项目**:国家自然科学基金项目(42071302);重庆市留学人员回国创业创新支持计划(cx2019144)

**Supported by**: National Natural Science Foundation of China(42071302); Innovation Program for Chongqing Overseas Returnees (cx2019144)

ments for feature descriptors. Meanwhile, deep learning-based methods for scene classification of remote sensing images have shown powerful feature extraction capabilities and have been widely applied in scene classification. However, current scene classification methods mainly focus on remote sensing images with high spatial resolution, which are mostly three-channel images with limited spectral information. This limitation often leads to confusion and misclassification in visually similar categories such as geometric structures, textures, and colors. Therefore, integrating spectral information to improve the accuracy of scene classification has become an important research direction. However, existing methods have some shortcomings. For example, convolutional operations have translation invariance and are sensitive to local information, which causes difficulty in capturing remote contextual information. Meanwhile, although Transformer methods can extract long-range dependency information, they have limited capability in learning local information. Moreover, combining convolutional neural networks (CNNs) and Transformer methods incurs high computational complexity, which hinders the balance between inference efficiency and classification accuracy. This study proposes a high spectral scene classification method called the spatial-spectral model distillation (SSMD) network to address the aforementioned issues. **Method** In this study, we utilize spectral information to improve the accuracy of scene classification and overcome the limitations of existing methods. First, we propose a spatial-spectral joint self-attention mechanism based on ViT (SSViT) to fully exploit the spectral information of hyperspectral images. SSViT integrates spectral information into the Transformer architecture. By exploring the intrinsic relationships between pixels and between spectra, SSViT extracts richer features. In the spatial-spectral joint mechanism, SSViT leverages the spectral information of different categories to identify the differences between them, which enables fine-grained classification of land cover and improves the accuracy of scene classification. Second, we introduce the concept of knowledge distillation to further enhance the classification performance. In the framework of teacher-student models, SSViT is used as the teacher model, and a pretrained model, that is, Visual Geometry Group 16(VGG16), is used as the student model to capture contextual information of complex scenes. The teacher model extracts spectral information and global features among samples, while the student model focuses on capturing local features. The student model can learn and mimic the prior knowledge of the teacher model, which improves the discriminative ability of the student model. The joint training of the teacher-student models enables comprehensive extraction of land cover features, which improves the accuracy of scene classification. Specifically, the image is divided into 64 image patches in the spatial dimension, and 32 spectral bands in the spectral dimension. Each patch and band can be regarded as a token. Each patch and band are flattened into row vectors and mapped to a specific dimension through a linear layer. The learned vectors are concatenated with the embedded samples for the final prediction of image classification of the teacher model. A position vector is generated and directly concatenated with the token mentioned above as the input to the Transformer. The multi-head attention mechanism outputs encoded representations containing information from different subspaces to model global contextual information, which improves the representation capacity and learning effectiveness of the model. Finally, feature integration is performed through a multi-layer perceptron and a classification layer to achieve classification. The process of knowledge distillation consists of two stages. The first stage optimizes the teacher and student models by minimizing the loss function with distillation coefficients. In the second stage, the student model is further adjusted using the loss function, which leverages the supervision from the performance-excellent complex model to train the simple model. This adjustment aims for higher accuracy and better classification performance. The complex model is referred to as the teacher model, while the simpler model is referred to as the student model. The training mode of knowledge distillation provides the student model with more informative content, which allows it to directly learn the generalization ability of the teacher model. **Result** We compare our model with 10 models, including 5 traditional CNN classification methods and 5 latest scene classification methods on 3 public datasets, namely, OHID-SC (Orbita hyperspectral image scene classification dataset), OHS-SC (another Orbita hyperspectral scene classification dataset), and HSRS-SC (Hyperspectral remote sensing dataset for scene classification). The quantitative evaluation metrics include overall accuracy, standard deviation, and confusion matrix, and the confusion matrix on the three datasets is provided to clearly display the classification results of the proposed algorithm. Experimental results show that our model outperforms all other methods on OHID-SC, OHS-SC, and HSRS-SC datasets, and the classification accuracies on OHID-SC, OHS-SC, and HSRS-SC datasets are improved by 13.1%, 2.9%, and 0.74%, respectively, compared with the second-best model. Meanwhile, comparative experiments on

OHID-SC dataset show that the proposed algorithm can effectively improve the classification accuracy of hyperspectral scenes. **Conclusion** In this study, the proposed SSMD network not only effectively utilizes the target spectral information of hyperspectral data but also explores the feature relationship between global and local levels, synthesizes the advantages of traditional and deep learning models, and produces more accurate classification results.

**Key words:** hyperspectral scene classification; convolutional neural network (CNN); Transformer; spatial-spectral joint self-attention mechanism; knowledge distillation (KD)

## 0 引　言

随着航空和卫星传感器的发展，遥感图像包含的地物信息更加丰富完整，在环境监测、资源勘探、军事侦察和灾难评估等领域得到广泛应用(赵雪梅 等，2021；潘尔婷 等，2021)。遥感影像场景分类是遥感影像处理领域的一个基础问题，其主要目的是对遥感影像进行特征分析，挖掘遥感影像中包含的语义信息，将其映射到类别标签上(白坤 等，2022)。近年来，学者们提出了一系列的场景分类方法，主要包括基于手工特征的场景分类方法和基于深度学习的场景分类方法(邓培芳 等，2021)。

基于手工提取特征的方法有尺度不变特征变换(scale invariant feature transform, SIFT)(Yang 和 Newsam，2008)、局部二值模式(local binary pattern, LBP)(Ojala 等，2002)、梯度直方图(histogram of oriented gradient, HOG)(Dalal 和 Triggs，2005)和通用搜索树(generalized search tree, GIST)(Lim 等，2014)等，但其局部特性不能直接表示整幅图像。通过对手工提取特征编码使其描述整幅图像，如空间金字塔匹配(spatial pyramid matching, SPM)(Lazebnik 等，2006)、局部聚合描述子向量(vector of locally aggregated descriptors, VLAD)(Jégou 等，2012)和视觉词袋(bag of visual words, BOVW)(Zhou 等，2015)等。手工特征在纹理整齐和空间分布均匀的遥感图像上表现良好，但难以刻画出复杂遥感场景的语义信息(余甜微 等，2022)，导致模型泛化性能比较差，致使很难在遥感图像领域广泛应用。

深度学习能挖掘数据潜在特征并且对数据表达更高效准确，广泛应用于遥感图像场景分类。现有工作表明，在大规模数据集(如 ImageNet)上使用预训练模型是一种很好的方法(徐科杰 等，2021)。Liang 等人(2016)利用各种卷积神经网络(convolutional neural network, CNN)预训练模型提取不同特征，并使用新的分类器用于高分辨率航空图像分类。Yuan 等人(2019)将预训练 VGG19Net 提取的局部特征重新排列，用于遥感图像场景分类。Lu 等人(2019)引入了一种特征聚合 CNN(feature aggregation CNN, FACNN)来进行场景分类。上述策略在小规模数据集上简单有效，但在训练样本不足的情况下，会限制其在遥感图像场景分类精度上的提升空间及其泛化能力(倪康 等，2022)。因而，人们提出了一些新的方法来解决遥感影像场景分类问题，如判别特征网络(discriminative feature network, DFN)(Yao 等，2016)、生成式对抗网络(generative adversarial networks, GAN)(Ma 等，2019)和少样本学习(few-shot learning, FSL)(Li 等，2021)等。上述方法可以获得很好的分类效果，但局部空间和短距离依赖信息会限制卷积的效果，很难获取视觉场景的全局上下文信息。

基于自注意力机制的 Transformer 可以从长序列数据中学习丰富的特征，不仅能够建立长期依赖关系且能并行训练(胡义 等，2023；Xie 等，2022)。余东行等人(2022)用 Transformer 替代传统 CNN 获取全局视野，并将基于 patch 的框架用基于图像分类的框架代替来解决训练和测试效率低下问题。张艺超等人(2023)利用 Transformer 自注意力机制提取上下文信息，通过层级融合防止分层传播过程中有效信息的损失。辛紫麒等人(2023)提出了一种光谱—空间联合 Transformer 模型，通过光谱和空间支路引入多阶特征交互层，将浅层边缘信息和深层语义信息融合从而实现分类。Li 等人(2020) 设计了双分支双注意力机制(double branch dual attention, DBDA)网络来优化光谱特征提取能力。上述方法虽然能捕获全局信息和表征光谱序列信息，但是联合模型的计算复杂度大，不能兼顾推理效率和分类性能。

当前大部分场景分类方法面向高分辨率遥感影像，包含极为有限的光谱信息，在几何结构、纹理和颜色等视觉属性相似的类别中容易产生混淆和错分

现象，而高光谱图像拥有“图谱合一”的表达方式，通过挖掘目标光谱差异将空间特征和光谱特征两类表达方式结合在一起，可以提升对地观测技术在相似场景上的判别能力，实现场景精细分类。Xu等人(2021)以ResNet18(residual network 18)为主干网络，采用空—谱注意力机制，重点学习关键波段和位置信息。Liu等人(2021)利用多尺度卷积模块检测高光谱图像局部区域中像素之间空间和光谱维数的细微变化，用密集交叉注意机制捕获最具代表性的特征。注意力机制关注高光谱图像的空—谱信息，卷积模块关注局部区域，两者忽略了全局上下文信息。因此，如何将局部信息和全局信息充分结合并有效利用高光谱影像的空—谱信息对影像进行精确场景分类是目前研究的重点和难点(徐科杰 等，2021)。

为了实现高光谱场景分类，本文提出一种面向高光谱场景分类的空—谱模型蒸馏网络(spatial-spectral model distillation network for hyperspectral scene classification，SSMD)，该算法由教师模型和学生模型联合组成。教师模型是基于空—谱联合自注意力机制的ViT(spital-spectral vision Transformer，SSViT)，不仅可以充分学习高光谱图像特有的光谱信息，还能获取长距离依赖信息；学生模型是基于VGG16(Visual Geometry Group 16)的预训练模型，用于提取局部细节信息。二者协同优化，学生模型利用教师模型提取到的鉴别信息与自身提取到的特征信息进行判别学习分类。在OHID-SC(Orbita hyperspectral image scene classification dataset)、OHS-SC(Orbita hyperspectral scene classification dataset)和HSRS-SC(hyperspectral remote sensing dataset for scene classification)数据集上的分类结果验证了本文算法在高光谱场景分类任务上的有效性。

## 1 本文方法

### 1.1 面向高光谱场景分类的空—谱模型蒸馏网络

高光谱遥感场景空间布局复杂且蕴涵丰富的光谱信息，如何有效利用目标光谱信息以及探索远距离依赖关系是提高分类性能的有效方法。因此本文通过在Transformer中引入空—谱联合自注意力，有效利用光谱信息，并通过知识蒸馏(knowledge distillation，KD)探索高光谱图像全局与局部之间的特征关系进行场景分类。提出的SSMD方法总体架构如图1所示。图1中，SSMD算法包含教师模型和学生模型两个部分。考虑到其光谱学习能力和全局建模能力，引入SSViT作为SSMD算法的教师模型，探究高光谱场景图像块之间的关系，并采用VGG16作为学生模型学习局部结构信息。在模型训练阶段，通过两阶段训练模式优化教师和学生模型，将长距离依赖知识从SSViT传递到VGG16，且两个阶段的损失系数不相同。

### 1.2 教师模型

SSViT可以捕获序列长距离依赖关系，同时空—谱联合自注意力机制能探索目标空间信息和光谱信息，可以发挥高光谱数据“图谱合一”的优势，所以本文选择用SSViT作为教师模型，其中空—谱联合自注意力机制框架如图2所示。对于输入为$H\times W\times C$的每一幅高光谱图像($H$、$W$和$C$分别为图像的长、宽和通道数)，在空间维和光谱维分别将其划分为$N_1$个$P\times P\times C$大小的图像块和$N_2$个$B\times B\times 1$的光谱波段，其中$N_1=(H\times W)/P^2$，$N_1$和$P$分别为空间维度划分的图像块个数和大小；$N_2=(H\times W)/B^2$，$N_2$和$B$分别为光谱维度划分的波段数和长度。每一个图像块和光谱波段分别可以看做一个$token$，记为$token_{spa}$和$token_{spe}$。

将每个图像块和波段展平成行向量，经过线性映射层映射到维度$D$，此时空间维序列$\boldsymbol{S}_{spa}$尺寸为$N_1\times(P\times P\times C)$，光谱维序列$\boldsymbol{S}_{spe}$尺寸为$N_2\times(B\times B\times 1)$。接着，利用一个可学习的向量$\boldsymbol{L}_{emb}$与已经完成嵌入的样本进行级联，该向量用于教师模型对影像分类的预测。最后，位置向量$\boldsymbol{P}_{emb}$与上述$token$拼接作为SSViT的编码输入部分，输入部分由空间维$\boldsymbol{I}_{spa}$和光谱维$\boldsymbol{I}_{spe}$组成，分别表示为

$$\boldsymbol{I}_{spa}=[\boldsymbol{L}_{emb};S_{spa}^{1}D;S_{spa}^{2}D;\cdots;S_{spa}^{N_1}D]+\boldsymbol{P}_{emb} \tag{1}$$

$$\boldsymbol{I}_{spe}=[\boldsymbol{L}_{emb};S_{spe}^{1}D;S_{spe}^{2}D;\cdots;S_{spe}^{N_2}D]+\boldsymbol{P}_{emb} \tag{2}$$

编码部分作为SSViT最重要的部分，由$L$层相同结构堆叠而成，每层均由多头自注意力(multi-head self-attention，MSA)和多层感知机(multilayer perceptron，MLP)交叉组成，每一个模块都应用残差连接且后面紧跟一个层归一化(layer norm，LN)模块，从而实现高光谱场景图像内的信息流动。SSViT模型中第$l$层详细结构如图3所示。

多头自注意力输出的编码信息可以用于建模全

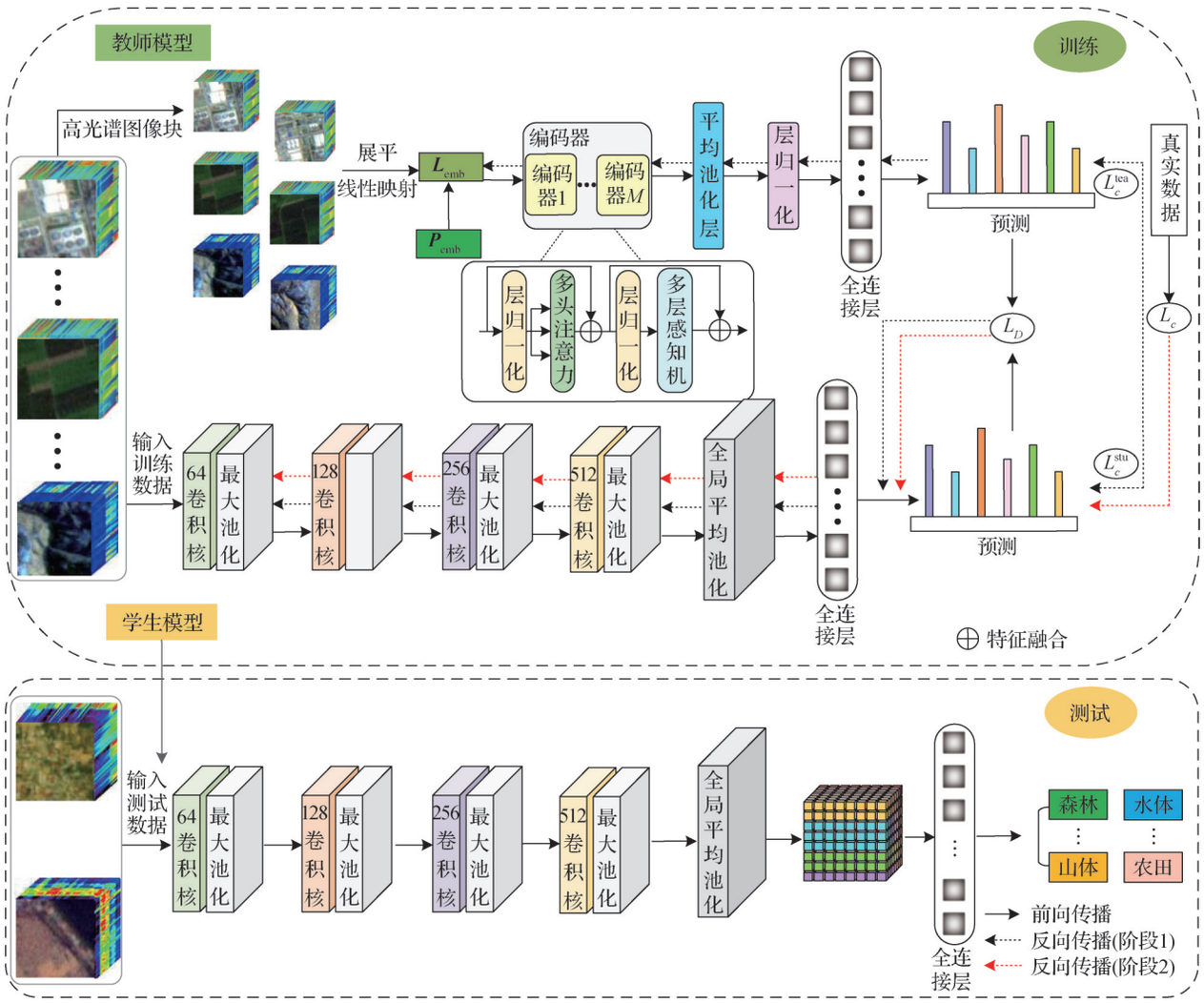



图 1　SSMD 算法总体框图

Fig. 1　The framework of the SSMD algorithm

局上下文特征，提高模型表示能力和学习效果。与自注意力区别在于其将注意力分为多个头（子空间），最后把每个头的输出拼接在一起。以光谱维为例，将输入 $\boldsymbol{I}_{\mathrm{spe}}$ 传入全连接层并将输出结果重塑为矩阵，分别作为查询向量 $\boldsymbol{Q}$、键向量 $\boldsymbol{K}$ 和值向量 $\boldsymbol{V}$。即

$$\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}^{Q} \tag{3}$$

$$\boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}^{K} \tag{4}$$

$$\boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}^{V} \tag{5}$$

式中，$\boldsymbol{X}$ 为输入矩阵，$\boldsymbol{W}^{Q}$，$\boldsymbol{W}^{K}$，$\boldsymbol{W}^{V}$ 是 3 个可训练的参数矩阵。

多头自注意力定义多组不同的参数矩阵（$\boldsymbol{W}_0^{Q}$，$\boldsymbol{W}_0^{K}$，$\boldsymbol{W}_0^{V}$，$\cdots$，$\boldsymbol{W}_i^{Q}$，$\boldsymbol{W}_i^{K}$，$\boldsymbol{W}_i^{V}$）与输入矩阵进行乘积，每组分别计算生成不同的 $\boldsymbol{Q}^i$，$\boldsymbol{K}^i$，$\boldsymbol{V}^i$，进行不同参数的学习，$head(i)$ 表示多头注意力机制的第 $i$ 个头，此过程可以表示为

$$head_i = f_{\mathrm{Attention}}(\boldsymbol{Q}\boldsymbol{W}_i^{Q}, \boldsymbol{K}\boldsymbol{W}_i^{K}, \boldsymbol{V}\boldsymbol{W}_i^{V}) \tag{6}$$

其次，每组的 $\boldsymbol{Q}$ 和 $\boldsymbol{K}^{\mathrm{T}}$（$\boldsymbol{K}$ 的转置）相乘进行相似度计算，并对向量做归一化得到权重 $f_{\mathrm{Attention}}$，其计算式为

$$f_{\mathrm{Attention}}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = f_{\mathrm{softmax}}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{7}$$

式中，$1/\left(d_k\right)^{0.5}$ 为比例因子。

权重矩阵与对应的 $\boldsymbol{V}$ 相乘，加权求和之后将输出结果拼接（$Concat$）到一起，乘以矩阵 $\boldsymbol{W}^{O}$，$f_{\mathrm{MultiHead}}$ 为多头注意力，此过程表示为

$$f_{\mathrm{MultiHead}}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = Concat(head_0, \cdots, head_i)\boldsymbol{W}^{O} \tag{8}$$

多层感知机由两层全连接层和激活函数组成，能解决非线性问题，还有较好的泛化能力。所用的

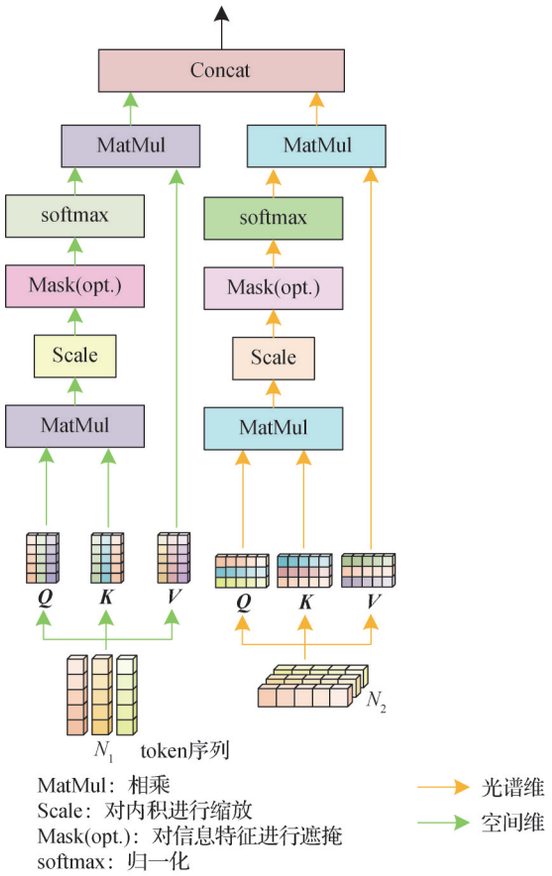



图2　空—谱联合自注意力机制框架

Fig. 2　The framework of spatial-spectral joint self-attention mechanism

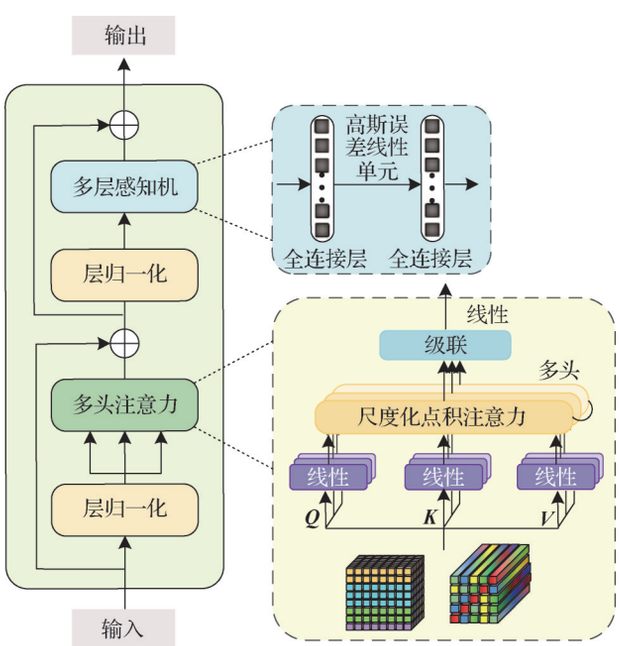



图3　SSViT模型第$l$层详细结构

Fig. 3　The detailed structure of the layer $l$ in the SSViT

激活函数为高斯误差线性单元(Gaussian error linear unit,GELU),表达式为

$$f_{\mathrm{GELU}}(x)=x\Phi(x) \tag{9}$$

$$\Phi(x)=\int_{-\infty}^{x}\frac{\mathrm{e}^{-t^2/2}}{\sqrt{2\pi}}\mathrm{d}t=2\left[1+\mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right] \tag{10}$$

式中,$\int_{-\infty}^{x}\frac{\mathrm{e}^{-t^2/2}}{\sqrt{2\pi}}\mathrm{d}t$是标准正态分布的累积分布函数,而同时,标准正态分布的累积分布函数可以用$erf$来表示:$erf(x)=\frac{2}{\sqrt{\pi}}\int_{0}^{x}\mathrm{e}^{-t^2}\mathrm{d}t$。

GELU的最终定义为

$$f_{\mathrm{GELU}}(x)=x\Phi(x)=x\cdot\frac{1}{2}\left[1+erf\left(\frac{x}{\sqrt{2}}\right)\right] \tag{11}$$

最终分类将最后一层的第1个元素$\boldsymbol{I}_L^0$作为预测语义类别的表示。因此,教师模型的最终输出$\boldsymbol{Z}_{\mathrm{tea}}$可以表示为

$$\boldsymbol{Z}_{\mathrm{tea}}=LN(\boldsymbol{I}_L^0) \tag{12}$$

式中,$LN()$表示层归一化。

### 1.3　学生模型

CNN由于权重共享机制对特征的全局位置不敏感,但因为卷积特征图的局部敏感性,可以充分提取局部特征信息,从而有效地与教师模型提取的远距离依赖信息形成互补。本文引入了经典的VGG16分类网络作为学生模型,如图4所示。

学生模型VGG16有较深的网络结构、较小的卷积核和池化采样域,能够在获得更多图像特征的同时避免过多的计算量,从图4可知,VGG16由6部分组成,前5部分分别由2—3层卷积层和1层池化层组成,最后一部分通过3层全连接层输出并采用softmax激活函数对分类目标进行预测。学生模型的最终输出$\boldsymbol{Z}_{\mathrm{stu}}$可以表示为

$$\boldsymbol{Z}_{\mathrm{stu}}=f_{\mathrm{softmax}}(z_i)=\frac{\exp(z_i)}{\sum_j\exp(z_j)} \tag{13}$$

式中,$z$是向量,$z_i$和$z_j$是其中的一个元素。

### 1.4　蒸馏损失系数

知识蒸馏通过性能出色的复杂模型所提取的监督信息训练简单模型,学生模型使用地面真实数据对硬标签进行训练,而且对教师模型产生的具备更多潜在信息的软标签进行优化,以此达到更高的精度和更好的分类性能。在SSMD算法中,知识蒸馏作用于SSViT和VGG16之间,将SSViT中的远程依赖知识和光谱信息传递到VGG16中,SSMD算法的

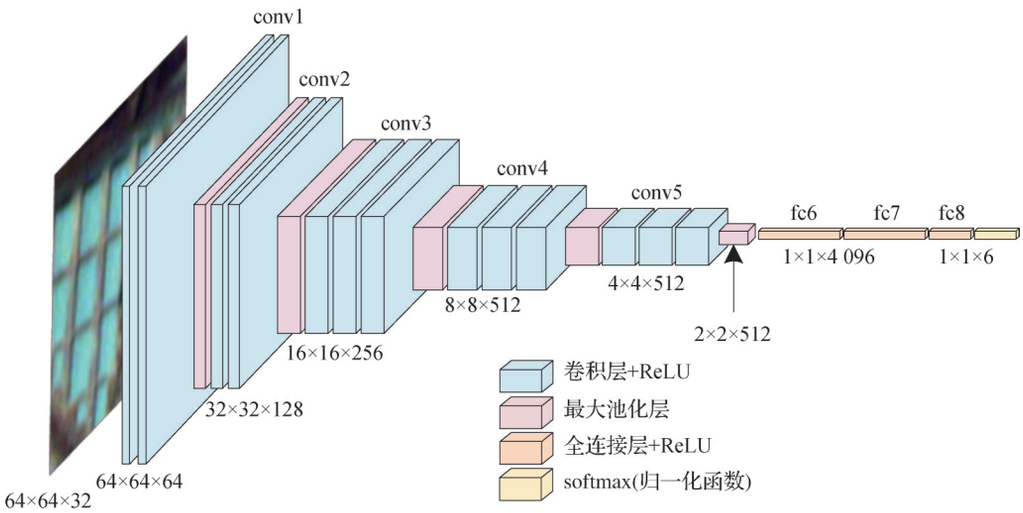



图4　VGG16网络框架

Fig. 4　The framework of the VGG16

优化策略为两阶段训练且采用不同的蒸馏损失系数$\alpha$。

第1阶段是通过蒸馏系数最小化损失函数来优化教师和学生模型。损失函数可定义为

$$L_1 = \alpha L_D + (1-\alpha)L_c^{\text{stu}} + L_c^{\text{tea}} \tag{14}$$

$$L_D = D_{\text{KL}}(p \| q) = \sum p(x)\log\frac{p(x)}{q(x)} \tag{15}$$

式中，$\alpha$为蒸馏损失系数，$L_D$为蒸馏损失函数，$L_c^{\text{tea}}$是教师模型与地面真值的交叉熵函数，$L_c^{\text{stu}}$是学生模型与地面真值的交叉熵函数。$D_{\text{KL}}$为KL(Kullback-Leibler)散度，用来度量两个概率分布相似度的指标，$p(x)$和$q(x)$分别是教师模型和学生模型预测结果的概率分布。

第2阶段通过损失函数调整学生模型，计算为

$$L_2 = \alpha L_D + (1-\alpha)L_c^{\text{stu}} \tag{16}$$

使用高温softmax函数代替传统的softmax（$T=1$）生成类别概率，$T$值越高产生的概率分布越软。温度为$T$时，教师模型和学生模型在第$i$类上的输出可表示为

$$\boldsymbol{Q}_{\text{tea}}^{\text{T}}(i) = \frac{\exp\left(\frac{Z_{\text{tea}}^i}{T}\right)}{\sum_n \exp\left(\frac{Z_{\text{tea}}^n}{T}\right)} \tag{17}$$

$$\boldsymbol{Q}_{\text{stu}}^{\text{S}}(i) = \frac{\exp\left(\frac{Z_{\text{stu}}^i}{T}\right)}{\sum_n \exp\left(\frac{Z_{\text{stu}}^n}{T}\right)} \tag{18}$$

式中，$\boldsymbol{Q}_{\text{tea}}^{\text{T}}(i)$、$\boldsymbol{Q}_{\text{stu}}^{\text{S}}(i)$分别代表教师模型和学生模型在第$i$类上的输出，$n$为总标签数量，$Z_{\text{tea}}^i$和$Z_{\text{stu}}^i$分别表示教师模型和学生模型的输出产生的第$i$个概率值。为了增加训练样本的多样性，在每幅图像上随机进行了几种形式的变换，具体设置如表1所示。

**表1　数据增强设置**

**Table 1　Settings of data augmentation**

| 方式 | 参数 | 补充说明 |
|---|---|---|
| 仿射变换 | 90°，180°，360° | 每幅图像执行0～3个增广函数 |
| 翻转 | 0.5 | 每幅图像有50%的概率进行镜像翻转或上下翻转 |
| 通道变换 | (10,100) | 对图像的某几个通道进行变换 |

本文提出的SSMD运算过程步骤如下：

1)训练阶段。

阶段(1)：以小的知识蒸馏系数$\alpha$进行训练；

输入：高光谱图像训练集$\boldsymbol{S}_{\text{in}}^{\text{train}}$，教师模型SSViT，学生模型VGG16，蒸馏损失系数$\alpha$，温度$T$，教师模型函数集合$z_{\text{tea}}(\cdot)$，学生模型函数集合$z_{\text{stu}}(\cdot)$；

输出：优化后的教师模型和学生模型；

①$\boldsymbol{Z}_{\text{tea}} \leftarrow z_{\text{tea}}(\boldsymbol{S}_{\text{in}}^{\text{train}})$；

②$\boldsymbol{Z}_{\text{stu}} \leftarrow z_{\text{stu}}(\boldsymbol{S}_{\text{in}}^{\text{train}})$；

③$L_1 \leftarrow \alpha L_D + (1-\alpha)L_c^{\text{stu}} + L_c^{\text{tea}}$；

④通过蒸馏系数最小化损失函数$L_1$优化教师模型和学生模型；

阶段(2)：以大的知识蒸馏系数$\alpha$进行训练；

输入：高光谱图像数据集$\boldsymbol{S}_{\text{in}}^{\text{train}}$，教师模型SSViT，学生模型VGG16，蒸馏损失系数$\alpha$，温度$T$，教师模型

函数集合$z_{tea}(\cdot)$,学生模型函数集合$z_{stu}(\cdot)$;

输出:优化后的学生模型;

①$\boldsymbol{Z}_{tea} \leftarrow z_{tea}(\boldsymbol{S}_{in}^{train})$;

②$\boldsymbol{Z}_{stu} \leftarrow z_{stu}(\boldsymbol{S}_{in}^{train})$;

③$L_2 \leftarrow \alpha L_D + (1-\alpha)L_c^{stu}$;

④通过损失函数$L_2$优化学生模型,教师模型停止训练;

2)测试阶段。实现场景分类的阶段。

输入:高光谱图像测试集$\boldsymbol{S}_{in}^{test}$,优化后的学生模型VGG16,学生模型函数集合$z_{tea}(\cdot)$;

输出:预测标签$O_{label}$;

1)$\boldsymbol{Z}_{stu} \leftarrow z_{stu}(\boldsymbol{S}_{in}^{train})$;

2)$O_{label} \leftarrow f_{softmax}(\boldsymbol{Z}_{stu})$;

3)测试集每一幅图像生成语义标签$O_{label}$。

## 2 实验验证

### 2.1 数据集

为评估提出方法的有效性,选取3个数据规模不一致的高光谱场景分类数据集OHID-SC(刘英旭等,2023)、OHS-SC(Wang等,2022)和HSRS-SC(徐科杰 等,2021)进行实验。

OHID-SC数据集来源于重庆大学,图像采集自珠海1号OHS系列卫星,共包含2 280幅高光谱影像,尺寸为64×64像素,空间分辨率为10 m,光谱分辨率为2.5 nm,幅宽为150 km,波段范围400~1 000 nm,分为6个典型场景:城镇、农田、森林、山体、水体和未利用区,每个类别分别有353、439、284、428、398和378幅影像,如图5所示。

OHS-SC数据集来源于中南大学,图像采集自OHS高光谱卫星,包含9 050幅高光谱图像,尺寸为64×64像素,空间分辨率为10 m,光谱分辨率为2.5 nm,幅宽150 km,波段范围400~1 000 nm,波段数64,总共标记为9类,包括农田、荒地、森林、工业、住宅、河流、道路、农村和海洋,如图6所示。

HSRS-SC数据集来源于重庆大学,图像来自轻便机载光谱成像仪,包含1 385幅高光谱图像,尺寸为256×256像素,空间分辨率为1 m,波长范围(可见光到近红外,380~1 050 nm),光谱分辨率为3.5 nm,传感器光谱带宽2.4 nm,共48个波段,被标记为5类,包括农田、建筑、城市建筑、未利用区和水体,能够反映详实的地物空间信息和丰富的光谱信息,如图7所示。

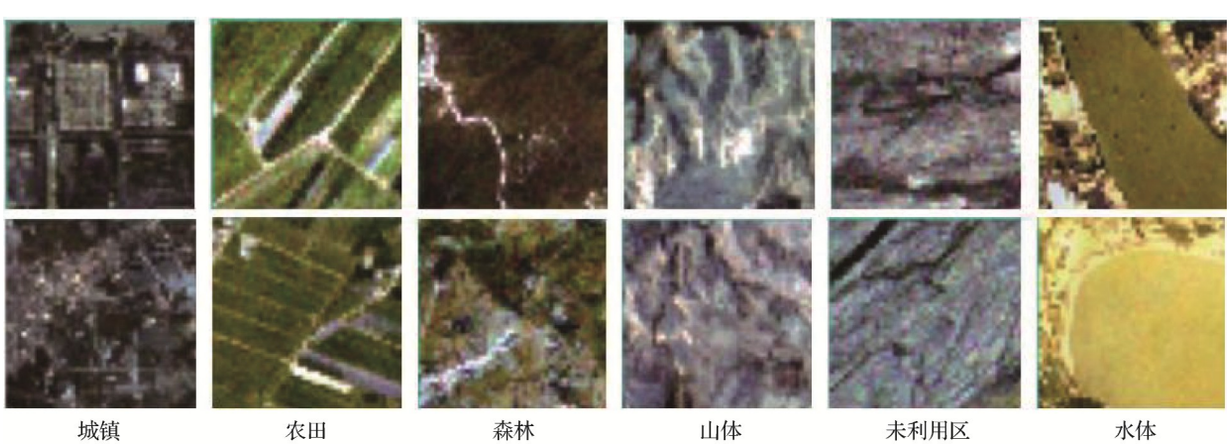
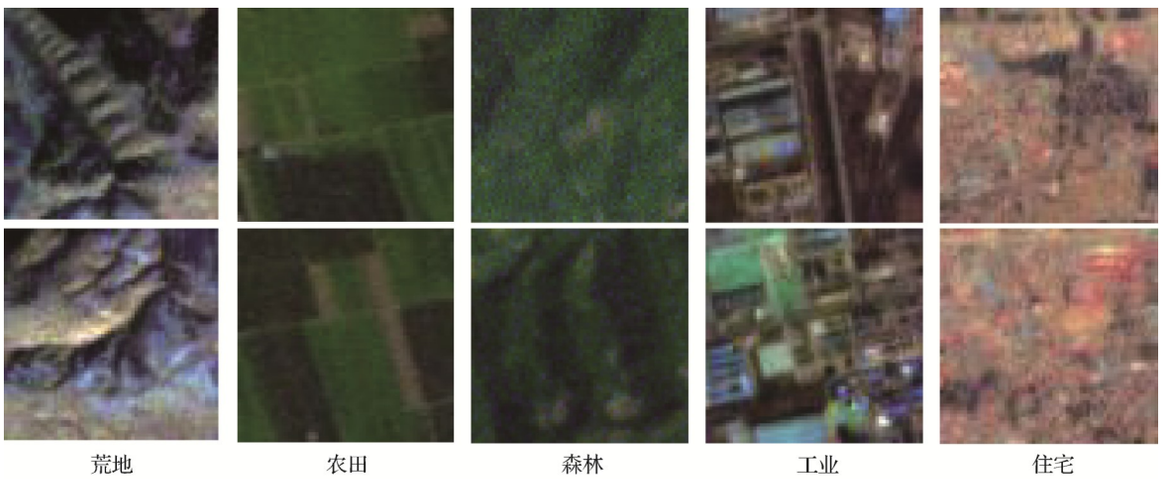



图5 OHID-SC数据集(6类)

Fig. 5 OHID-SC dataset(six classes)

### 2.2 实验设置

实验平台利用256 GB内存的PANYAO 7048GR服务器,实验中使用的GPU是NVIDIA Titan RTX,深度学习框架为PyTorch,共迭代60个epoch,批大小设置为12,所有数据集使用原图尺寸进行训练。此外,选择随机梯度下降法(stochastic gradient descent,SGD)对SSViT模型进行优化直至收敛,动量大小设置为0.9,权重衰减为0.000 1。对于学生网络(VGG16),所采用的优化器为Adam,权重衰减也设置为0.000 1,二者学习率均设置为0.000 1。在验证实验中,性能评估指标采用的是总体分类精度(overall accuracy,OA)和混淆矩阵。所有的实验



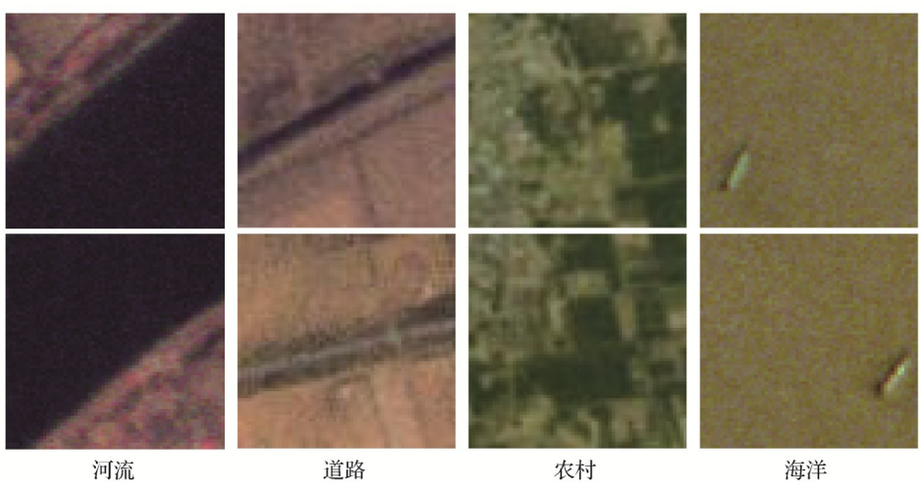




图 6　OHS-SC 数据集(9 类)

Fig. 6　OHS-SC dataset(nine classes)

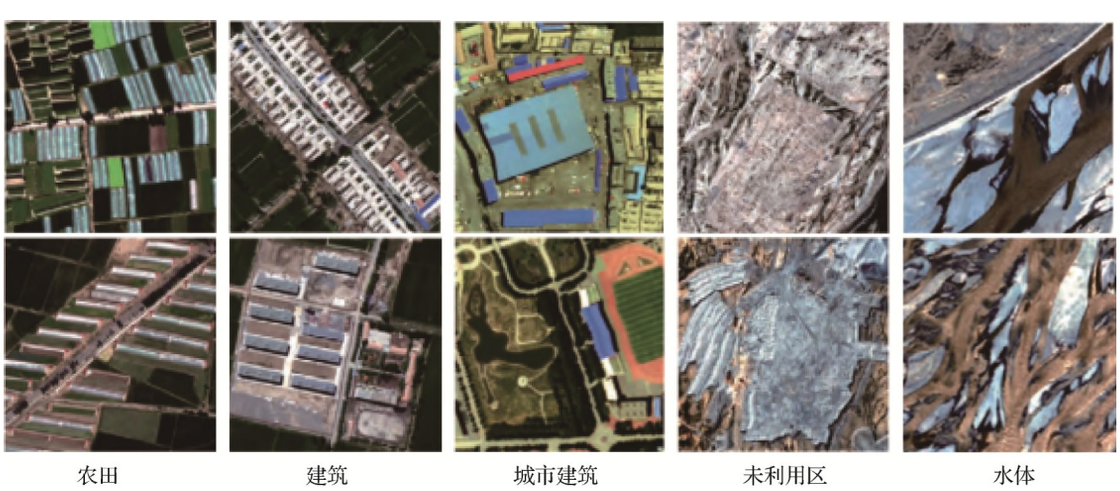



图 7　HSRS-SC 数据集(5 类)

Fig. 7　HSRS-SC dataset(five classes)

均重复 5 次，以保证实验结果的有效性和可靠性，且每次实验的训练集和测试集按比例随机划分，在 3 个高光谱场景分类数据集中均选用 30% 的样本进行训练。

### 2.3　参数设置

对于所提出的 SSMD 算法，蒸馏损失系数 $\alpha$ 和温

度$T$两个参数决定了最终的分类性能。在OHID-SC数据集(训练样本占比：30%)上进行参数敏感性实验，分析其对SSMD算法的影响并选取最优值。

$\alpha$可以平衡不同损失函数之间的重要性，为了验证不同的$\alpha$对分类结果的影响，取值区间设置为$\{0.1, 0.15, \cdots, 0.5\}$。图8为OHID-SC数据集不同蒸馏损失系数$\alpha$对分类精度的影响。根据图8可知，分类精度先逐步提高，达到峰值后开始下降。由于教师模型存在一定的偏置，教师在训练完成后训练集正确率会高于测试集，为了减缓这种趋势，首先通过设置较大的值快速拟合教师网络，再逐步调低$\alpha$的值，来确保网络的分类正确率。实验结果说明$\alpha$在0.3时可以很好地平衡不同损失函数的权重，后续实验中参数$\alpha$均设置为0.3。

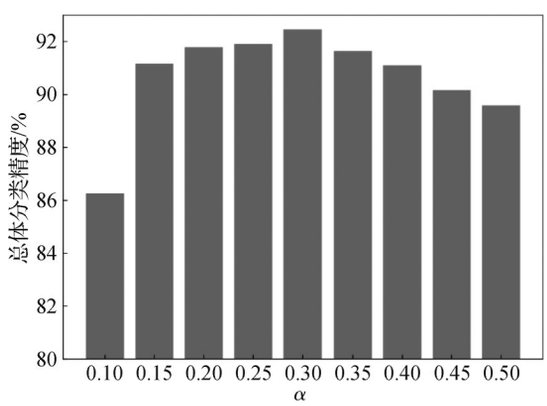



图8　OHID-SC数据集不同蒸馏损失系数$\alpha$对分类精度的影响

Fig. 8　OAs with respect to different values of KD loss coefficient on the OHID-SC dataset

在参数优化的过程中，$T$值的大小影响学生模型对负标签的关注程度。为了研究算法在不同$T$下的分类性能，$T$的变化区间设置为$\{5, 10, 15, \cdots, 55\}$。根据图9所示，$T$等于30时精度达到最高，并在之后精度出现小幅度的上升又下降的情况，但始终未曾超过$T$在30时的分类精度。因为$T$值较小时，负标签会变得更小，学生模型对于负标签的关注也会减少；$T$值较大时软标签分布包含更加丰富的信息，而学生模型无法捕捉其全部的信息。

### 2.4　消融实验

为了评估本文算法的分类效果，在OHID-SC、OHS-SC和HSRS-SC这3个数据集上对ViT、VGG和基于空—谱联合的自注意力机制蒸馏模型(SSMD)展开消融实验，训练样本均占比30%，其分类精度如图10所示。

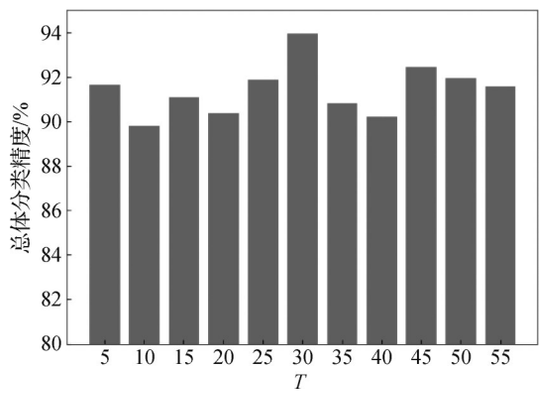



图9　OHID-SC数据集上不同温度$T$对分类精度的影响

Fig. 9　OAs with respect to different values of temperature on the OHID-SC dataset

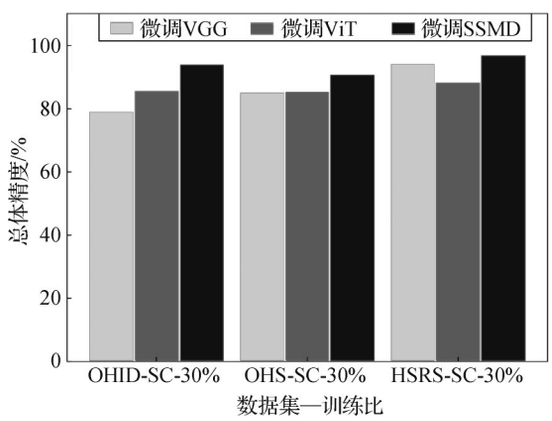



图10　OHID-SC、OHS-SC和HSRS-SC数据集上不同算法消融实验

Fig. 10　Ablation study of different methods on the OHID-SC, OHS-SC and HSRS-SC datasets

从图10可知，微调后的ViT在OHID-SC数据集和OHS-SC数据集上的分类性能比微调VGG优越，因为ViT可以通过自注意力提取图像上下文特征的优秀能力，但由于ViT在较小的数据集上性能受限，所以在HSRS-SC数据集上的分类表现稍显不足。SSMD算法捕获光谱信息和长距离依赖信息并通过知识蒸馏与局部信息相融合，增强特征表征能力，因此SSMD算法在3个数据集上的分类精度都有明显提升。

### 2.5　混淆矩阵实验

为了进一步直观评估SSMD算法的混淆程度和分类误差，在3个数据集上生成了9个混淆矩阵，图11—图13分别是OHID-SC数据集、OHS-SC数据集和HSRS-SC数据集的实验结果，均为30%的训练

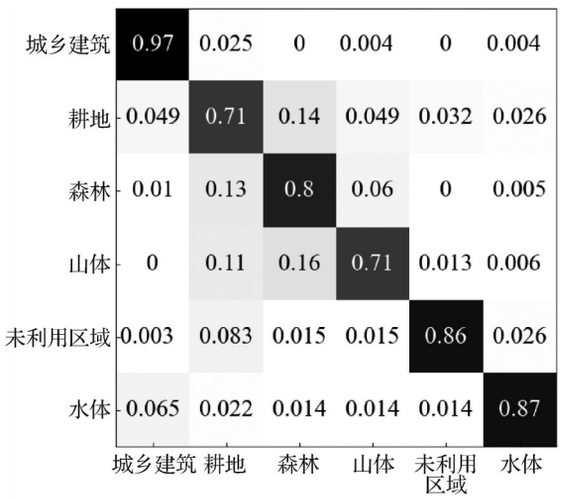
样本。

OHID-SC 数据集上得到的 VGG、ViT 和 SSMD 混淆矩阵如图 11 所示。由图中可知，SSMD 算法在城镇、农田、森林、未利用土地和水体等类别上分类性能均高于 VGG 和 ViT，其中城镇、农田、未利用土地和水体均达到 96% 以上，且在山体等相似场景分类上，也有精度为 85% 的较好效果。原因是 SSMD 算法中引入了光谱信息，缓解了数据集类间相似性高产生的影响，进而减少了混淆程度。

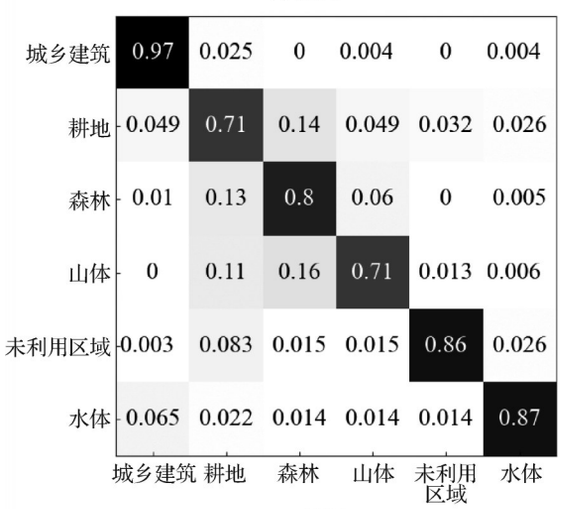
OHS-SC 数据集上得到的混淆矩阵如图 12 所示，在混淆矩阵中，荒地、海洋和森林的分类精度高达 100%，农田和农村的分类精度也在 97% 以上。工业区—住宅、住宅—工业区、道路—荒地、道路—农田几类混淆度高的样本经对数据集查看分析可知，工业区和住宅的空间几何结构相似，道路和荒地以及农田的地物结构和空间模式相似，在预测时产生错分，但与 VGG 和 ViT 单个算法相比，SSMD 在分类精度方面也有相对提升，这也说明本文算法在场景极其相似的情况下还略有不足，但在某些场景中的分类精度优异，这也证实了所提出的算法在一定程度上可以汲取 SSViT 的光谱信息和全局信息探索地物在空—谱维度上的分布，以此较好地完成多分类任务。

HSRS-SC 数据集上得到如图 13 所示的混淆矩阵，SSMD 与 VGG 和 ViT 相比取得了优异的分类效果，在未利用区域、农田和水体上的分类精度几乎达到 100%。SSMD 通过知识蒸馏来挖掘图像中潜在的远距离依赖信息，结合 VGG 出色的学习能力，在保证其优秀分类能力的基础上增强特征判别能力，从而进一步提高场景分类精度，对于 HSRS-SC 数据集发生在建筑和农田的混淆，取得了具有竞争力的分类结果。

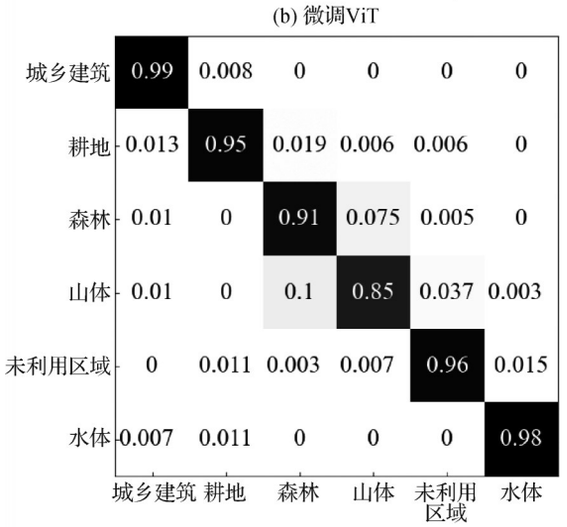
为了验证本文算法的有效性，选取了 GoogLeNet、ResNet18、MobileNet_v3_large、ShuffleNet_v2_x1_0 和 VGG16 5 种预训练分类网络以及 SKAL、ARCNet 和 MF2CNet 这 3 种较新的场景分类方法，在 OHID-SC、OHS-SC 和 HSRS-SC 这 3 个数据集上对上述 10 种方法基于 30% 的训练样本进行对比，对比结果如表 2 所示。由表 2 可知，所提出的 SSMD 在 OHID-SC、OHS-SC 和 HSRS-SC 数据集上均获得最佳分类精度。预训练分类网络在 HSRS-SC 数据集上的分类效果都较优，原因是 HSRS-SC 数据集较为简单，上述方法可以致力于局部关键特征感知，使网络关注在更显著的区域上。SKAL、ARCNet



(a) 微调VGG



(b) 微调ViT



(c) SSMD(本文)

图 11　OHID-SC 数据集混淆矩阵

Fig. 11　The confusion matrix of OHID-SC dataset

((a) fine-tuned VGG; (b) fine-tuned ViT; (c) SSMD(ours))

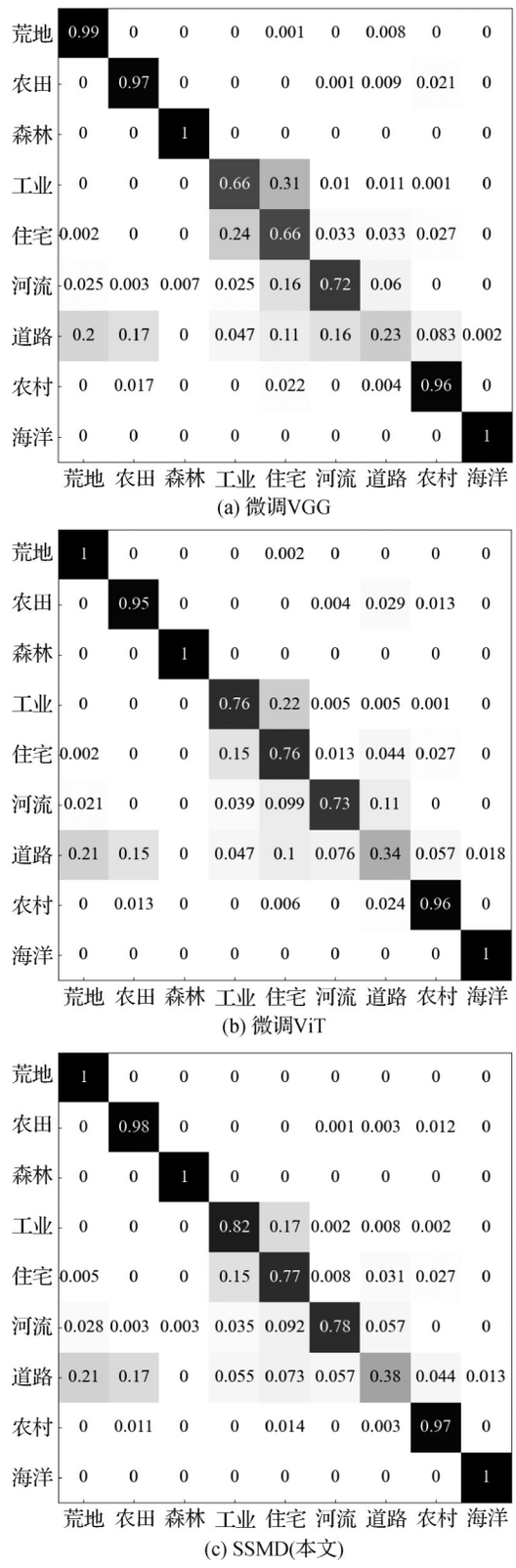



图12 OHS-SC数据集混淆矩阵

Fig. 12 The confusion matrix of OHS-SC dataset

((a) fine-tuned VGG; (b) fine-tuned ViT; (c) SSMD(ours))

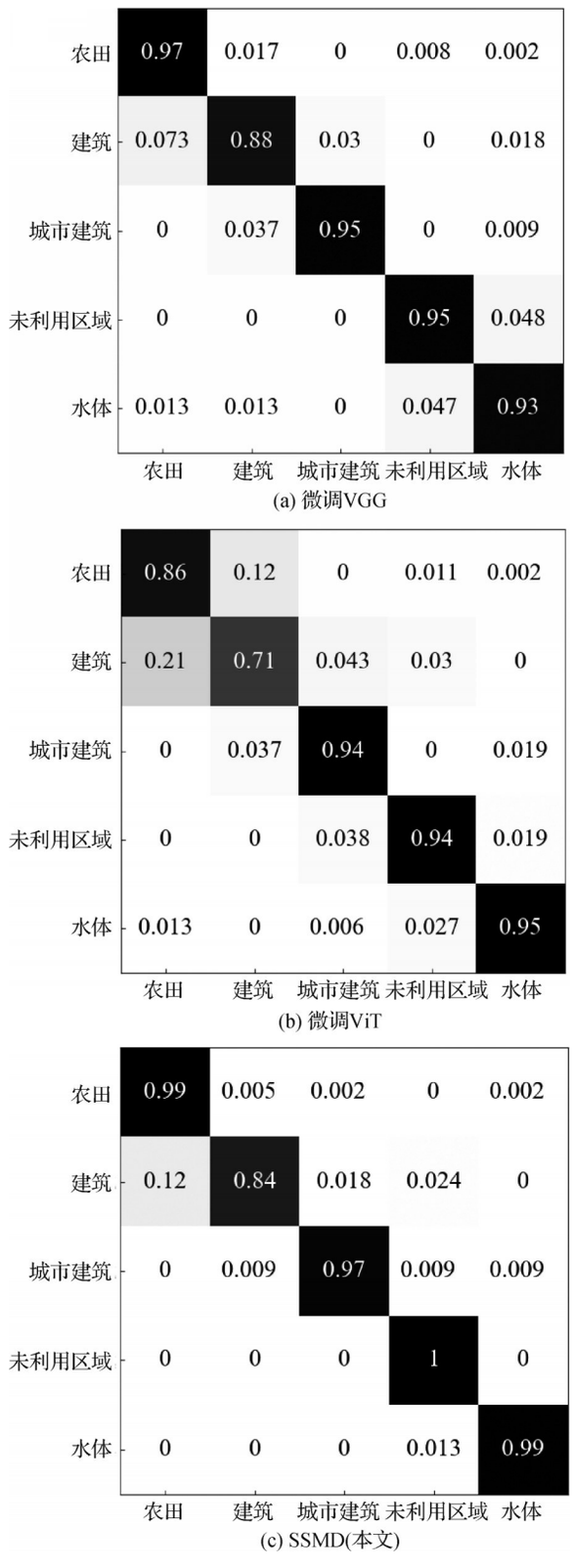



图13 HSRS-SC数据集混淆矩阵

Fig. 13 The Confusion matrix of HSRS-SC dataset

((a) fine-tuned VGG; (b) fine-tuned ViT; (c) SSMD(ours))

和MF2CNet这3种场景分类方法主要针对高分辨率遥感影像场景分类，不能充分学习高光谱图像的空—谱信息。前期一些传统、优秀的网络模型由于数据集的噪声、不均匀性导致过拟合，以及降维操作

会损失一些重要信息致使分类精度不高。基于CNN和Transformer端到端的处理方式可以直接在图像的空间—光谱维度上进行学习，捕获局部空间信息和光谱信息，从而有效地提高分类精度，为高光谱图像分析提供了新的有效工具。

**表2　各类方法在3个数据集上的分类精度对比（平均值±标准差）**

**Table 2　Overall accuracies of various methods on three datasets (mean ± std)**

/%

| 分类网络 | OHID-SC | OHS-SC | HSRC-SC |
|---|---|---|---|
| GoogLeNet | 47.28±2.91 | 77.77±1.11 | 93.98±0.47 |
| ResNet18 | 66.09±1.25 | 81.22±0.94 | 96.27±0.29 |
| ShuffleNet | 42.88±3.09 | 67.40±5.72 | 95.94±0.13 |
| MobileNet | 45.02±3.66 | 76.84±1.3 | 95.36±0.19 |
| VGG16 | 79.02±1.70 | 85.11 | 94.27±0.01 |
| SKAL-G | 55.78±1.01 | 79.18±0.70 | 89.46±1.48 |
| SKAL-V | 70.18±0.6 | 81.73±0.18 | 92.52±0.60 |
| SKAL-R | 62.09±2.35 | 77.77±1.13 | 86.74±1.51 |
| ARCNet-R | 78.41±0.15 | 87.80±0.05 | 94.51±0.62 |
| MF2CNet | 80.94±1.66 | 84.89±0.15 | 95.64±0.13 |
| SSMD（本文） | **94.12±0.53** | **90.70±0.20** | **97.01±0.40** |

注：加粗字体表示各列最优结果。

SSMD算法在Transformer中引入丰富的光谱信息，通过知识蒸馏将教师模型提取到的长距离依赖信息迁移到学生模型上，与学生模型获取的局部信息形成互补，使得学生模型更具判别性，提高图像分类的准确率，通过在上述3个不同尺寸规模数据集上的实验也说明本文算法有比较好的泛化性。

为了对SSMD运行时间进行分析，不同方法在OHID-SC、OHS-SC和HSRC-SC数据集上的运行时间如表3所示，所有方法均在同一台计算机上运行。表3中SKAL、ARCNet和MF2CNet是3种场景分类方法，但主要针对高分辨率遥感影像数据，因而运行时间较短，MF2CNet、ShuffleNet和MobileNet在GPU上采用并行计算，因为有足够大的显存在推理速度上未有提升。ResNet在batchsize设置较大时占用显存大，因此推理时间加长。VGG运算量大导致运行时间增加。综合考虑分类精度和计算时间，本文SSMD不仅具有较高分类精度且有较好的计算效率。

**表3　各类方法在3个数据集上的运行时间对比**

**Table 3　Running time of various methods on three datasets**

/s

| 方法 | OHID-SC | | OHS-SC | | HSRS-SC | |
|---|---|---|---|---|---|---|
| | 训练 | 测试 | 训练 | 测试 | 训练 | 测试 |
| GoogLeNet | 158 | 159 | 731 | 757 | 790 | 815 |
| ResNet18 | 69 | 75 | 331 | 396 | 575 | 754 |
| ShuffleNet | 158 | 161 | 641 | 631 | 612 | 726 |
| MobileNet | 165 | 163 | 718 | 768 | 653 | 793 |
| VGG16 | 58 | 59 | 256 | 257 | 593 | 743 |
| SKAL-G | 82 | 128 | 253 | 306 | 291 | 324 |
| SKAL-V | 53 | 57 | 235 | 272 | 256 | 294 |
| SKAL-R | 43 | 55 | 206 | 293 | 259 | 308 |
| ARCNet-R | 37 | 56 | 133 | 195 | 243 | 277 |
| MF2CNet | 165 | 214 | 245 | 283 | 285 | 305 |
| SSMD（本文） | 135 | 113 | 883 | 752 | 528 | 601 |

## 3　结　论

传统CNN分类算法通过获取图像空间信息和局部特征表达进行分类，且当前的场景分类方法主要针对于高光谱遥感影像，不仅忽略了光谱信息和全局依赖信息，而且高光谱场景分类研究较为匮乏。鉴于此，提出一种基于空—谱联合ViT和知识蒸馏的高光谱场景分类方法，空—谱联合自注意力机制通过挖掘像素与像素之间序列关系和光谱波段与波段之间的关系，提取空间特征和光谱特征，知识蒸馏在空—谱特征基础上有效地将上下文信息和局部信息结合在一起，在高光谱场景分类数据集上获得满意的效果。实验中，在OHID-SC、OHS-SC和HSRS-SC数据集上利用SSMD算法得到的最高分类精度分别为94.12%、90.70%以及97.01%。本文知识蒸馏只有一个教师模型，学生模型得到的知识来源较为单一。若能将多个教师的预测分布进行聚合，根据所提供知识的重要程度分配不同的权重，让学生网络从多个模型中进行学习，以此达到近似或比模型融合更佳的效果，从而解决相似场景错分问题，提高场景分类性能。因此下一步工作将注重多教师知识蒸馏算法的研究。

## 参考文献(References)

Bai K, Mu X D, Chen X B, Zhu Y Q and You X A. 2022. Unsupervised remote sensing image scene classification based on semi-supervised learning. Acta Geodaetica et Cartographica Sinica, 51(5): 691-702 (白坤, 慕晓冬, 陈雪冰, 朱永清, 尤轩昂. 2022. 融合半监督学习的无监督遥感影像场景分类. 测绘学报, 51(5): 691-702) [DOI: 10.11947/j.AGCS.20210270]

Dalal N and Triggs B. 2005. Histograms of oriented gradients for human detection//Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, USA: IEEE: 886-893 [DOI: 10.1109/CVPR.2005.177]

Deng P F, Xu K J and Huang H. 2021. CNN-GCN-based dual-stream network for scene classification of remote sensing images. National Remote Sensing Bulletin, 25(11): 2270-2282 (邓培芳, 徐科杰, 黄鸿. 2021. 基于CNN-GCN双流网络的高分辨率遥感影像场景分类. 遥感学报, 25(11): 2270-2282) [DOI: 10.11834/jrs.20210587]

Hu Y, Huang B C and Li F. 2023. Image classification based on Transformer adaptive feature vector fusion. Journal of Optoelectronics·Laser, 34(6): 602-609 (胡义, 黄勃淳, 李凡. 2023. 基于Transformer自适应特征向量融合的图像分类. 光电子·激光, 34(6): 602-609) [DOI: 10.16136/j.joel.2023.06.0303]

Jégou H, Perronnin F, Douze M, Sánchez J, Pérez P and Schmid C. 2012. Aggregating local image descriptors into compact codes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(9): 1704-1716 [DOI: 10.1109/TPAMI.2011.235]

Lazebnik S, Schmid C and Ponce J. 2006. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories//Proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE: 2169-2178 [DOI: 10.1109/CVPR.2006.68]

Li L J, Han J W, Yao X W, Cheng G and Guo L. 2021. DLA-MatchNet for few-shot remote sensing image scene classification. IEEE Transactions on Geoscience and Remote Sensing, 59(9): 7844-7853 [DOI: 10.1109/TGRS.2020.3033336]

Li R, Zheng S Y, Duan C X, Yang Y and Wang X Q. 2020. Classification of hyperspectral image based on double-branch dual-attention mechanism network. Remote Sensing, 12(3): #582 [DOI: 10.3390/rs12030582]

Liang Y L, Monteiro S T and Saber E S. 2016. Transfer learning for high resolution aerial image classification//Proceedings of 2016 IEEE Applied Imagery Pattern Recognition Workshop (AIPR). Washington, USA: IEEE: 1-8 [DOI: 10.1109/AIPR.2016.8010600]

Lim C H, Risnumawan A and Chan C S. 2014. A scene image is nonmutually exclusive—a fuzzy qualitative scene understanding. IEEE Transactions on Fuzzy Systems, 22(6): 1541-1556 [DOI: 10.1109/TFUZZ.2014.2298233]

Liu R M, Ning X, Cai W W and Li G J. 2021. Multiscale dense cross-attention mechanism with covariance pooling for hyperspectral image scene classification. Mobile Information Systems, 2021: #9962057 [DOI: 10.1155/2021/9962057]

Liu Y X, Pu C Y, Xu D K, Yang Y C and Huang H. 2023. Lightweight deep global-local knowledge distillation network for hyperspectral image scene classification. Optics and Precision Engineering, 31(17): 2598-2610 (刘英旭, 蒲春宇, 许典坤, 杨沂川, 黄鸿. 2023. 面向高光谱影像场景分类的轻量化深度全局-局部知识蒸馏网络. 光学精密工程, 31(17): 2598-2610) [DOI: 10.37188/OPE.20233117.2598]

Lu X Q, Sun H and Zheng X T. 2019. A feature aggregation convolutional neural network for remote sensing scene classification. IEEE Transactions on Geoscience and Remote Sensing, 57(10): 7894-7906 [DOI: 10.1109/TGRS.2019.2917161]

Ma D A, Tang P and Zhao L J. 2019. SiftingGAN: generating and sifting labeled samples to improve the remote sensing image scene classification baseline *in vitro*. IEEE Geoscience and Remote Sensing Letters, 16(7): 1046-1050 [DOI: 10.1109/LGRS.2018.2890413]

Ni K, Zhao Y Q and Chen Z. 2022. Multi-scale convolutional neural network driven by sparse second-order attention mechanism for remote sensing scene classification. Acta Photonica Sinica, 51(6): #0610004 (倪康, 赵雨晴, 陈志. 2022. 稀疏二阶注意力机制驱动的多尺度卷积遥感图像场景分类网络. 光子学报, 51(6): #0610004 [DOI: 10.3788/gzxb20225106.0610004]

Ojala T, Pietikainen M and Maenpaa T. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7): 971-987 [DOI: 10.1109/TPAMI.2002.1017623]

Pan E T, Ma Y, Huang J, Fan F, Li H and Ma J Y. 2021. Hyperspectral image classification evaluated across different datasets. Journal of Image and Graphics, 26(8): 1969-1977 (潘尔婷, 马泳, 黄珺, 樊凡, 李皞, 马佳义. 2021. 跨数据集评估的高光谱图像分类. 中国图象图形学报, 26(8): 1969-1977) [DOI: 10.11834/jig.210123]

Wang Y Z, Xiao R, Qi J and Tao C. 2022. Cross-sensor remote-sensing images scene understanding based on transfer learning between heterogeneous networks. IEEE Geoscience and Remote Sensing Letters, 19: #8021705 [DOI: 10.1109/LGRS.2021.3116601]

Xie Y X, Yan J, Kang L, Guo Y M, Zhang J H and Luan X D. 2022. FCT: fusing CNN and Transformer for scene classification. International Journal of Multimedia Information Retrieval, 11(4): 611-618 [DOI: 10.1007/s13735-022-00252-7]

Xin Z Q, Li Z W, Wang L Q, Xu M M, Hu Y B and Liang J. 2023. Hyperspectral image classification of Yellow River Delta wetlands based on a spectral-spatial unified Transformer model. Marine Sciences, 47(5): 90-101 (辛紫麒, 李忠伟, 王雷全, 许明明, 胡亚斌, 梁建. 2023. 基于光谱—空间联合Transformer模型的黄河三角

洲湿地高光谱影像分类. 海洋科学，47(5)：90-101)［DOI：10.11759/hykx202204290012］

Xu K J，Deng P F and Huang H. 2021. HSRS-SC：a hyperspectral image dataset for remote sensing scene classification. Journal of Image and Graphics，26(8)：1809-1822（徐科杰，邓培芳，黄鸿. 2021. HSRS-SC：面向遥感场景分类的高光谱图像数据集. 中国图象图形学报，26(8)：1809-1822）［DOI：10.11834/jig.200835］

Xu K J，Huang H and Deng P F. 2021. Attention-based deep feature learning network for scene classification of hyperspectral images// Proceedings of the 55th Asilomar Conference on Signals，Systems，and Computers. Pacific Grove，USA：IEEE：1690-1693［DOI：10.1109/IEEECONF53345.2021.9723419］

Xu K J，Huang H，Deng P F and Li Y. 2022. Deep feature aggregation framework driven by graph convolutional network for scene classification in remote sensing. IEEE Transactions on Neural Networks and Learning Systems，33(10)：5751-5765［DOI：10.1109/TNNLS.2021.3071369］

Yang Y and Newsam S. 2008. Comparing SIFT descriptors and Gabor texture features for classification of remote sensed imagery//Proceedings of the 15th IEEE International Conference on Image Processing. San Diego，USA：IEEE：1852-1855［DOI：10.1109/ICIP.2008.4712139］

Yao X W，Han J W，Cheng G，Qian X M and Guo L. 2016. Semantic annotation of high-resolution satellite images via weakly supervised learning. IEEE Transactions on Geoscience and Remote Sensing，54(6)：3660-3671［DOI：10.1109/TGRS.2016.2523563］

Yu D H，Xu Q，Zhao C，Guo H T，Lu J，Lin Y Z and Liu X Y. 2023. Attention-guided feature fusion and joint learning for remote sensing image scene classification. Acta Geodaetica et Cartographica Sinica，52(4)：624-637（余东行，徐青，赵传，郭海涛，卢俊，林雨准，刘相云. 2023. 注意力引导特征融合与联合学习的遥感影像场景分类. 测绘学报，52(4)：624-637）［DOI：10.11947/j.AGCS.2023.20210659］

Yu H Y，Xu Z，Zheng K，Hong D F，Yang H and Song M P. 2022. MSTNet：a multilevel spectral-spatial Transformer network for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing，60：#5532513［DOI：10.1109/TGRS.2022.3186400］

Yu T W，Zheng E R，Shen J G and Wang K. 2022. Optical remote sensing image scene classification based on multi-level cross-layer bilinear fusion. Acta Photonica Sinica，51(2)：#0210007（余甜微，郑恩让，沈钧戈，王凯. 2022. 基于多级别跨层双线性融合的光学遥感图像场景分类. 光子学报，51(2)：#0210007）［DOI：10.3788/gzxb20225102.0210007］

Yuan Y，Fang J，Lu X Q and Feng Y C. 2019. Remote sensing image scene classification using rearranged local features. IEEE Transactions on Geoscience and Remote Sensing，57(3)：1779-1792［DOI：10.1109/TGRS.2018.2869101］

Zhang Y C，Zheng X T，Lu X Q. 2023. Hyperspectral image classification method based on hierarchical Transformer network. Acta Geodaetica et Cartographica Sinica，52(7)：1139-1147（张艺超，郑向涛，卢孝强. 2023. 基于层级Transformer的高光谱图像分类方法. 测绘学报，52(7)：1139-1147）［DOI：10.11947/j.AGCS.2023.20220540］

Zhao X M，Wu J and Chen R X. 2021. RMFS-CNN：new deep learning framework for remote sensing image classification. Journal of Image and Graphics，26(2)：297-304（赵雪梅，吴军，陈睿星. 2021. RMFS-CNN：遥感图像分类深度学习新框架. 中国图象图形学报，26(2)：297-304）［DOI：10.11834/jig.200397］

Zhou W X，Shao Z F，Diao C Y and Cheng Q M. 2015. High-resolution remote-sensing imagery retrieval using sparse features by auto-encoder. Remote Sensing Letters，6(10)：775-783［DOI：10.1080/2150704X.2015.1074756］

## 作者简介

薛洁，女，硕士研究生，主要研究方向为遥感图像处理和深度学习。E-mail：xuejie@cqu.edu.cn

黄鸿，通信作者，男，教授，博士生导师，主要研究方向为流形学习、模式识别和遥感影像智能化处理。E-mail：hhuang@cqu.edu.cn

蒲春宇，男，博士研究生，主要研究方向为模式识别和遥感影像图像处理。E-mail：puchunyu@stu.cqu.edu.cn

杨鄞铭，男，硕士研究生，主要研究方向为遥感影像分割和目标检测。E-mail：yangyinming@cqu.edu.cn

李远，男，博士研究生，主要研究方向为医学图像处理、机器学习和模式识别。E-mail：yuan_li@cqu.edu.cn

刘英旭，男，硕士研究生，主要研究方向为遥感图像处理和机器学习。E-mail：liuyingxu@stu.cqu.edu.cn