文章编号　1004-924X(2023)17-2598-13

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

刘英旭[1]，蒲春宇[1]，许典坤[2]，杨沂川[1]，黄　鸿[1*]
(1. 重庆大学 光电技术及系统教育部重点实验室，重庆 400044；
2. 重庆大学 光电工程学院测控技术与仪器专业，重庆 400044)

摘要：针对目标场景复杂的空间布局和高光谱影像固有的空-谱信息冗余等挑战，提出了端到端的轻量化深度全局-局部知识蒸馏(Lightweight Deep Global-Local Knowledge Distillation，LDGLKD)网络。为探索空-谱特征的全局序列属性，教师模型视觉Transformer(Vision Transformer，ViT)被用来指导轻量化学生模型进行高光谱影像场景分类。LDGLKD选择预训练的VGG16作为学生模型来提取局部细节信息，将ViT和VGG16通过知识蒸馏协同训练后，教师模型将所学习到的远程上下文关系向小规模学生模型进行传递。LDGLKD可通过知识蒸馏结合上述两种模型的优点，在欧比特高光谱影像场景分类数据集OHID-SC及公开的高光谱遥感图像数据集HSRS-SC上的最佳分类精度分别达到91.62%和97.96%。实验结果表明：LDGLKD网络具有良好的分类性能。根据欧比特珠海一号卫星提供的遥感数据构建的OHID-SC可以反映详细的地表覆盖情况，并为高光谱场景分类任务提供数据支撑。

关　键　词：高光谱场景分类；特征提取；视觉Transformer；知识蒸馏；基准数据集
中图分类号：TP751　　**文献标识码**：A　　**doi**：10.37188/OPE.20233117.2598

Lightweight deep global-local knowledge distillation network for hyperspectral image scene classification

LIU Yingxu[1], PU Chunyu[1], XU Diankun[2], YANG Yichuan[1], HUANG Hong[1*]

(1. *Key Laboratory of Optoelectronic Technology and Systems of the Education Ministry of China, Chongqing University, Chongqing* 400044, *China*;
2. *Measurement and Control Technology and Instrument major, College of Optoelectronic Engineering, Chongqing University, Chongqing* 400044, *China*)
* *Corresponding author, E-mail: hhuang@cqu.edu.cn*

Abstract: To address the challenges of the complex spatial layouts of target scenes and inherent spatial-spectral information redundancy of HSIs, an end-to-end lightweight deep global - local knowledge distillation (LDGLKD) method is proposed herein. To explore the global sequence properties of spatial-spectral features, the vision transformer (ViT) is used as the teacher to guide the lightweight student model for HSI scene classification. In LDGLKD, pre-trained VGG16 is selected as the student model to extract lo-

收稿日期：2023-01-03；**修订日期**：2023-02-09.
基金项目：国家自然科学基金资助项目(No. 42071302)；重庆市留学人员回国创业创新支持计划资助项目(No. cx2019144)

cal detail information. After collaborative training of ViT and VGG16 through knowledge distillation, the teacher model transmits the learned long-range contextual information to the small-scale student model. By combining the advantages of the two models through knowledge distillation, the optimal classification accuracy of LDGLKD on the Orbita HSI scene classification dataset (OHID-SC) and hyperspectral remote sensing dataset for scene classification (HSRS) reached 91.62% and 97.96%, respectively. The experimental results revealed that the proposed LDGLKD method presented good classification performance. In addition, the OHID-SC based on the remote sensing data obtained by the Orbita Zhuhai-1 satellite could reflect the detailed information of land cover and provide data support for HSI scene classification.

Key words: hyperspectral scene classification; feature extraction; vision transformer; knowledge distillation; benchmark dataset

1 引　言

随着遥感技术的飞速发展，所获取的遥感影像数量呈爆炸式增长，对遥感影像的解译工作也随之成为研究热点。作为遥感影像解译领域的重要分支之一，场景分类问题旨在根据遥感影像中所包含的视觉信息来给每一幅影像分配预定义的场景标签。场景分类任务以庞大的遥感影像数据量和计算机高效的计算能力为基础，近年来广泛地应用在城乡规划、环境监测和自然灾害检测等领域[1]。

场景分类任务根据遥感数据的不同可划分为高分辨率影像场景分类和高光谱影像场景分类。由于高分辨率影像能够清晰地呈现地物细节信息和空间模式，现阶段场景分类任务主要面向高分辨率影像。但当前场景分类任务面临复杂背景干扰和地物结构多变等挑战，由RGB三通道构成的高分辨率遥感影像仅包含地物信息与空间结构，有限的光谱信息难以分辨纹理、形状和颜色等视觉感知类似的场景之间的细小差别，可能造成错分或混淆现象。因此，在保证场景空间分辨率的基础上引入丰富的光谱信息，则可以提升对视觉感知类似场景的判别能力。

近年来，人们对基于高光谱影像(Hyperspectral Image，HSI)的场景分类任务展开了积极研究。与高分辨率影像相比，HSI具有“图谱合一”的优势，有利于挖掘地物细微的光谱差异来实现不同场景的本征表达。虽然HSI内蕴丰富的空-谱信息，但其非线性强、高度冗余和数据量大等特点导致场景分类模型存在特征提取能力有限和计算复杂度高等问题。由于高光谱影像场景分类的相关研究仍处于探索阶段，现阶段面向场景分类的高光谱数据集相对匮乏。

特征提取是实现高光谱场景分类的关键步骤。根据所提取特征的表示能力不同，面向场景分类任务的方法可分为传统分类方法和基于深度学习的方法两类。传统场景分类方法主要是通过目标地物的颜色、纹理、光谱等信息手工制作浅层特征，如尺度不变特征变换(Scale-Invariant Feature Transform，SIFT)特征[2]，定向梯度(Histogram of Oriented Gradients，HOGs)特征[3]和局部二值(Local Binary Pattern，LBP)特征[4]。将手工特征与特征编码方法相结合可以提取遥感影像高阶的统计信息，代表方法有视觉词袋(Bag-of-Visual-Words，BoVW)[5]，隐含狄利克雷分布(Latent Dirichlet Allocation，LDA)[6]和概率潜在语义分析(Orobabilistic Latent Semantic Analysis，PLSA)[7]。上述传统方法虽然易于实现，但是高性能特征描述符的构造严重依赖领域内相关经验且费时耗力。此外，浅层特征只能对场景信息的特定方面进行有限表征，导致上述方法的分类性能难以满足复杂的场景分类任务。

深度学习作为机器学习的一种新范式，能通过层级结构自主实现对输入数据从浅层到深层的抽象表示，在遥感场景分类任务中已得到广泛应用[8]。当前，基于深度学习的分类方法大致可以分为三类：基于全训练的方法、基于预训练模型的方法和基于微调的方法。周勤等[9]基于深度

置信网络(Deep Belief Network,DBN)提出迭代特征学习算法,该方法根据查询重构权值来实现判别特征的选择。为充分利用航拍场景不同层级的特征,Bi等[10]构建了一个多实例密集连接卷积神经网络。上述两种模型属于第一类方法,此类方法根据目标数据进行建模,因此不受固定网络架构的限制。由于遥感影像标记样本与模型可训练参数量之间的不平衡性,从零开始训练高性能模型需要高昂的计算成本。

考虑到图像的基本元素相同,人们采用在大规模数据集(如ImageNet)上预训练的模型(如GoogLeNet,VGGNet,ResNet)作为特征提取器,并对输出的特征映射进行编码,以提高场景分类性能。程塬等[11]利用现有卷积神经网络提取到的深度特征作为视觉词进行场景分类。为提高预训练模型的分类性能,李二珠等[12]利用特征编码方法将多层特征进行有效融合。然而,自然图像与遥感图像在空间布局、成像角度和背景信息等方面存在巨大差异,预训练模型直接学习到的特征难以准确描述目标场景。

为提高特征鉴别能力,许多方法利用遥感数据对预训练模型进行微调。与基于预训练模型的方法相比,利用遥感数据对预训练模型进行微调被认为是一种提高特征识别能力的有效策略。Chaib等[13]通过判别相关分析算法(Discriminant Correlation Analysis,DCA)对预训练VGGNet的输出特征再次进行挖掘。双流特征聚合深度神经网络(Two-Stream Feature Aggregation Deep Neural Network,TFADNN)[14]通过鉴别特征流和一般特征流来提升场景的表征能力。房杰等[15]采用圆形卷积模块(Circular Convolutional Module,CCM)对来自空间域的鉴别特征和频率域的位置特征进行有效融合。虽然上述方法在一定程度上提高了场景的分类精度,但针对遥感图像复杂的空间布局,关键的地物远程依赖性却被忽略。

Transformer[16]具有学习远程上下文关系的能力,这在遥感领域得到了充分验证。Bazi等[17]采用Transformer对远程信息进行提取,并引入多种数据增强方法来提升模型判别能力。遥感Transformer(Remote Sensing Transformer,TRS)[18]集成了卷积神经网络和Transformer两种架构的优势,并在公开数据集上取得了具有竞争力的分类结果。然而,Transformer在对空-谱信息高度冗余的HSI进行分析处理时,具有较高的计算复杂度和时间成本,这极大地限制了基于Transformer的方法在高光谱影像场景分类任务上的应用。知识蒸馏(Knowledge Distillation,KD)在保持良好性能的同时可降低模型的复杂度,因此它是实现模型压缩的较优策略。KD可以将教师网络学到的知识有效转移到学生网络,即将原始训练集与提炼出来的知识进行融合,以实现对学生网络的训练。石程等[19]设计了尺度蒸馏网络,该框架利用多尺度教师模型探索鉴别信息。Hu等[20]通过变分估计以概率的方式进行KD,实现交互信息的端到端优化。不过,基于KD的方法目前面临以下问题:(1)通常以卷积神经网络为教师模型,不能有效探索数据内蕴的长距离上下文信息;(2)教师和学生模型被视为两个独立的环节进行训练,难以实现协同优化;(3)为获取较低的计算复杂度,学生模型架构过于简单,教师模型的性能远优于学生模型。

针对上述问题,本文提出了轻量化深度全局-局部知识蒸馏(Lightweight Deep Global-Local Knowledge Distillation,LDGLKD)网络,以提升高光谱影像场景分类性能,降低模型的计算复杂度。由于HSI的高维特性强且空-谱信息冗余,部署ViT作为教师模型,通过多头注意力机制实现远程上下文关系的学习。预训练的学生模型VGG16用来提取数据中的局部细节信息。在LDGLKD中,改进的知识蒸馏架构通过有效控制学习率来实现学生模型与教师模型的协同优化,促使轻量化的学生模型能更有效地学习来自教师模型的鉴别信息。结合两种模型优点的学生模型单独进行测试,在保证分类精度的同时,降低了模型的计算复杂度。

2 原　理

高光谱影像具有复杂的空间布局和丰富的空-谱信息,探索远距离上下文关系可以进一步

提高分类性能。LDGLKD方法通过改进的知识蒸馏充分探索高光谱影像的远距离上下文信息和局部结构特征，其总体结构如图1所示。LDGLKD网络主要包括教师模型与学生模型两个部分。其中，教师模型ViT旨在提取输入样本块间的长距离上下文联系。结构相对简单的学生模型VGG16用来学习影像的局部细节信息。针对训练阶段和测试阶段损失函数的特定构造，两种模型在训练阶段可进行协同优化，以保证ViT将所学习到的判别信息有效转移到VGG16中。值得注意的是，仅有完成了知识蒸馏的学生模型进入测试阶段，旨在提高分类性能的同时减少模型的计算成本。

2.1　教师模型

ViT在影像分类任务上具有出色的性能，它可以通过多头注意力机制来提取序列之间的长距离依赖关系。因此，在LDGLKD网络中，ViT充当教师模型来向学生模型传递样本间的远程联系。

以本文所构造的32通道高光谱数据集为例，将每一幅高光谱影像的尺寸设为$H\times W\times 32$，其中H和W为影像的长和宽。将影像裁剪成N个尺寸为$P\times P\times 32$的样本，其中$N=H\times W/P^2$为样本总数，P为样本的边长，将所有样本拉伸从而得到一个尺寸为$N\times(P^2\times 32)$的序列$\boldsymbol{S}$。然后，采用一个尺寸为$(P^2\times 32)\times D$的嵌入

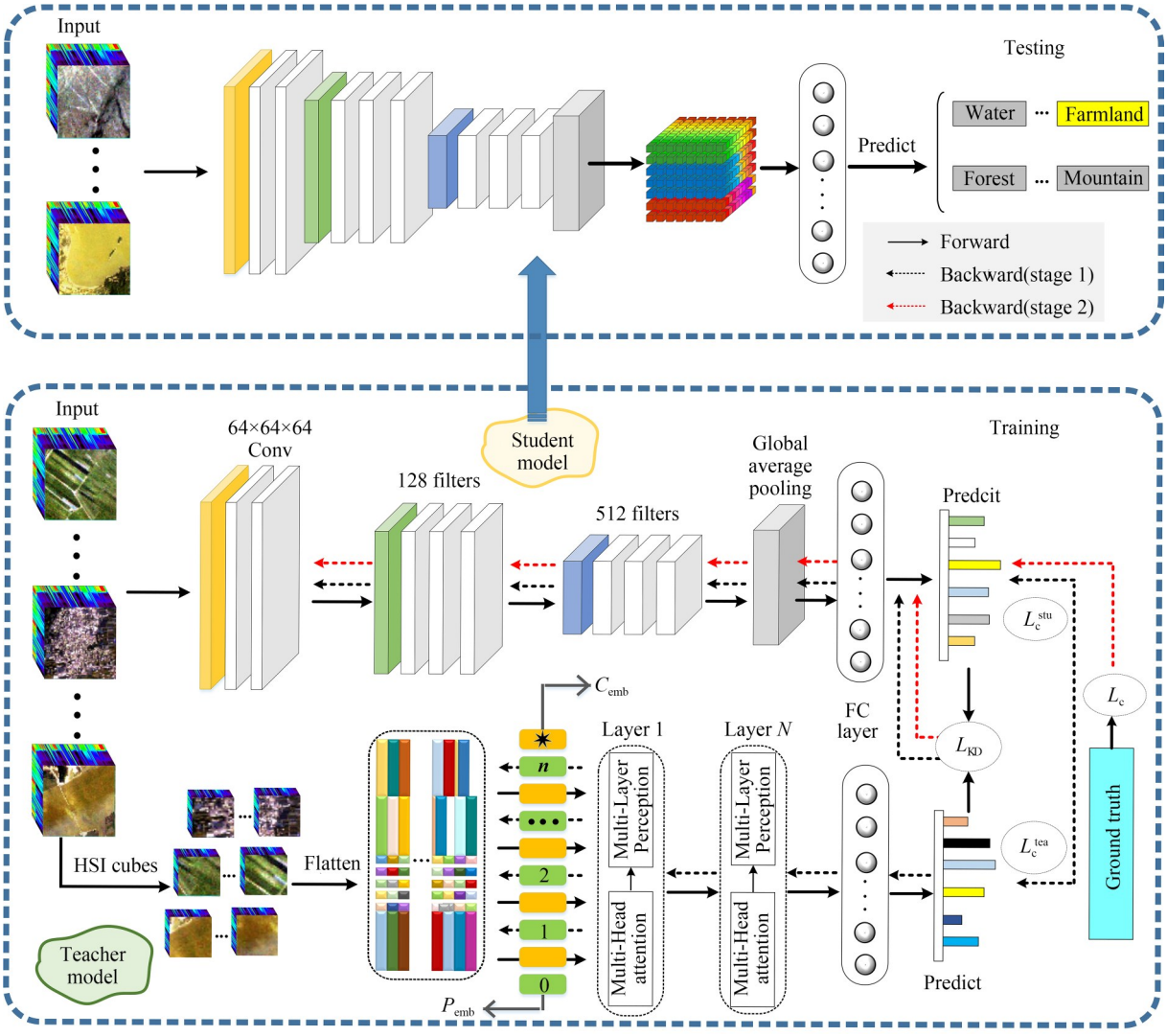


图1　轻量化深度全局-局部知识蒸馏网络

Fig. 1　Framework of proposed Lightweight Deep Globle-local Knowledge Distillation(LDGLKD) network

矩阵 $\boldsymbol{M}$ 将序列 $\boldsymbol{S}$ 线性映射到 D 维空间中。其次，利用一个可学习的向量 $\boldsymbol{C}_{\mathrm{emb}}$ 与已经完成嵌入的样本进行级联，该向量用作教师模型最后对影像分类的预测。考虑到记录每个样本所在位置的重要性，将一个维度为 $(N+1)\times D$ 的位置向量 $\boldsymbol{P}_{\mathrm{emb}}$ 与嵌入后的样本执行元素级加法。因此，教师模型的输入 $\boldsymbol{R}_0$ 可以表示为：

$$\boldsymbol{R}_0=[\boldsymbol{C}_{\mathrm{emb}},\boldsymbol{S}^1\boldsymbol{M},\boldsymbol{S}^2\boldsymbol{M},\ldots,\boldsymbol{S}^N\boldsymbol{M}]+\boldsymbol{P}_{\mathrm{emb}}. \tag{1}$$

Transformer 编码器作为 ViT 模型中最关键的部分，由 L 层相同结构堆叠而成，即由多头注意力机制（Multihead Self-attention，MSA）和多层感知器（Multilayer Perception，MLP）模块交替构成，这里将 L 设置为 12。在每一个模块之前均进行层标准化（Layer Norm，LN），并采用残差连接来传递影像中的信息。输入 $\boldsymbol{R}_0$ 被编码器处理后，MSA 会强调输入影像的长距离依赖信息。如图 2 所示，对于第 l 层，中间变量 $\boldsymbol{R}'_l$ 和输出 $\boldsymbol{R}_l$ 可以表示为：

$$\boldsymbol{R}'_l=\mathrm{MSA}(\mathrm{LN}(\boldsymbol{R}_{l-1}))+\boldsymbol{R}_{l-1}\quad(l=1,\ldots,L), \tag{2}$$

$$\boldsymbol{R}_l=\mathrm{MLP}(\mathrm{LN}(\boldsymbol{R}'_l))+\boldsymbol{R}'_l\ (l=1,\ldots,L). \tag{3}$$

在影像分类的过程中，将最后一层的第一个元素 $\boldsymbol{R}_L^0$ 进行层标准化处理后，教师模型的输出 $\boldsymbol{F}_{\mathrm{tea}}$ 可最终表示为：

$$\boldsymbol{F}_{\mathrm{tea}}=\mathrm{LN}(\boldsymbol{R}_L^0). \tag{4}$$

作为编码器中最重要的模块，MSA 由线性映射、自注意力机制、级联操作以及最后的线性映射四层组成。不同于传统的注意力机制，这里所采用的自注意力机制采取了一系列的查询向量 $\boldsymbol{Q}$、键向量 $\boldsymbol{K}$ 和值向量 $\boldsymbol{V}$ 来计算输出向量，这 3 个可学习的向量维度分别为 $d_{\mathrm{q}},d_{\mathrm{k}},d_{\mathrm{v}}$。对于每一个在输入序列 $\boldsymbol{R}$ 中的元素，$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 可以通过将 $\boldsymbol{R}$ 与 3 个可学习的矩阵 $\boldsymbol{U}_Q,\boldsymbol{U}_K,\boldsymbol{U}_V$ 相乘得到，$\boldsymbol{U}_Q,\boldsymbol{U}_K,\boldsymbol{U}_V$ 的维度均为 $D\times d_{\mathrm{k}}$，其计算公式为：

$$[\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}]=\boldsymbol{R}[\boldsymbol{U}_Q,\boldsymbol{U}_K,\boldsymbol{U}_V]. \tag{5}$$

为确定输入样本间的远距离联系，将每一个查询向量与所有键向量进行点积运算来生成注意力权重。注意力权重 $\boldsymbol{W}^{\mathrm{att}}$ 的计算公式如下：

$$\boldsymbol{W}^{\mathrm{att}}=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right)\boldsymbol{V}, \tag{6}$$

其中：$1/\sqrt{d_{\mathrm{k}}}$ 为比例因子，Softmax 是激活函数。

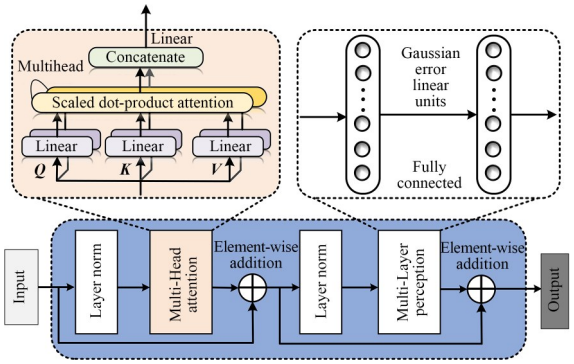


图 2　ViT 模型中第 l 层的详细结构

Fig. 2　Detail structure of layer l in ViT model

为了提高自注意力机制的性能并减少计算复杂度，$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 向量通过不同的线性映射重复计算 t 次，然后并行执行式（6）。随后，将生成的 t 个权重矩阵级联，其中每一个权重矩阵均称为一个“头”。MSA 模块的结果可以表示为：

$$\mathrm{MSA}(\boldsymbol{R})=[\boldsymbol{W}_1^{\mathrm{att}},\boldsymbol{W}_2^{\mathrm{att}},\ldots,\boldsymbol{W}_t^{\mathrm{att}}]\boldsymbol{W}, \tag{7}$$

其中：$\boldsymbol{W}$ 为维度是 $t\times d_{\mathrm{v}}\times D$ 的可学习矩阵，$[\cdot,\cdot]$ 表示级联操作。需要说明的是，$d_{\mathrm{k}}=d_{\mathrm{v}}=D/t$ 且 $t=12,D=768$。

当 MSA 模块的输出生成之后，使用 MLP 模块来产生 Transformer 编码器中的输出，MLP 模块由两层全连接层组成，中间由高斯误差线性单元（Gaussian Error Linear Units，GELU）作为激活函数将其连接。GELU 的计算公式如下：

$$\mathrm{GELU}(x)=x\varphi(x)=x\cdot\frac{1}{2}\left[1+\mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right], \tag{8}$$

$$\mathrm{erf}(x)=(2/\pi)\int_0^x \mathrm{e}^{-\eta^2}\,\mathrm{d}\eta, \tag{9}$$

其中 $\varphi(x)$ 代表标准高斯分布。在 Transformer 编码器中 LN 的计算公式如下：

$$\mu^f=\frac{1}{B}\sum_{i=1}^{B}a_i^f, \tag{10}$$

$$\sigma^f=\sqrt{\frac{1}{B}\sum_{i=1}^{B}(a_i^f-\mu^f)^2}, \tag{11}$$

式中：μ^f 代表第 f 层的均值，σ^f 代表第 f 层的标准差，B 代表一层中隐藏单元的个数，a_i^f 代表第 f 层中第 i 个隐藏单元的输入总和，每层中的所有隐藏单元采用相同的归一化项 μ 和 σ。归一化隐藏

单元 $\hat{a}^f$ 的计算公式如下：

$$\hat{a}^f=\frac{a^f-\mu^f}{\sqrt{(\sigma^f)^2+\varepsilon}}, \tag{12}$$

式中 ε 是为了计算稳定性而加上的一个常数。

2.2　学生模型

卷积神经网络（Convolutional Neural Networks，CNNs）根据其层级结构，可以逐步地从数据中提取到深层特征。近年来，各种各样的CNNs相继被提出，VGG16作为一种典型模型，具有特征提取能力强、结构相对简单、参数量较多等特点。学生模型利用一个全局平均池化层（Global Average Pooling，GAP）和一个全连接层来替代 VGG16的三个全连接层，以进一步简化VGG16的网络架构并减少可训练的参数量。因此，本文选取 VGG16 作为知识蒸馏中轻量化的学生模型。

全局平均池化层可以将一个大小为 $Y\times Z\times C$ 的特征图 $\boldsymbol{U}$ 转化为一个 C 维的特征向量，具体公式如下：

$$\mathrm{GAP}(\boldsymbol{U})=\frac{1}{Y\times Z}\sum_{y=1}^{Y}\sum_{z=1}^{Z}u_{y,z,c}$$
$$(c=1,2,\ldots,C), \tag{13}$$

式中 Y,Z,C 分别代表特征图的长、宽和通道数。

2.3　知识蒸馏

知识蒸馏是一种可以将知识从复杂神经网络（教师模型）转移到简单神经网络（学生模型）的特征压缩融合技术，它在保持高精度的同时还降低了计算复杂度。学生模型通过学习教师模型输出的“软标签”来获得更加丰富的信息，这其中不仅包括了正标签信息，也包括了负标签中的海量信息。同时，学生模型还对来自于地面真值的硬标签进行学习，以此来增强学生模型的稳定性。

在LDGLKD方法中，通过共同优化教师模型ViT和学生模型VGG16，将教师模型所提取的远距离上下文联系转移到学生模型。本文所提出的知识蒸馏策略共分为两个步骤。

在第一阶段，教师模型和学生模型通过最小化损失函数 L_{fir} 进行共同优化。然后，教师模型的学习率逐渐减小到0，学生模型单独进入到第二阶段，并以损失函数 L_{sec} 进行优化。其中，损失函数 L_{fir} 和 L_{sec} 分别为：

$$L_{\mathrm{fir}}=\alpha L_{\mathrm{KD}}+(1-\alpha)(L_{\mathrm{C}}^{\mathrm{tea}}+L_{\mathrm{C}}^{\mathrm{stu}}), \tag{14}$$

$$L_{\mathrm{sec}}=2\alpha L_{\mathrm{KD}}+(1-2\alpha)L_{\mathrm{C}}^{\mathrm{stu}}, \tag{15}$$

式中：L_{KD} 为知识蒸馏损失函数，$L_{\mathrm{C}}^{\mathrm{tea}}$ 为教师模型与地面真值的交叉熵损失函数，$L_{\mathrm{C}}^{\mathrm{stu}}$ 为学生模型与地面真值的交叉熵损失函数，α 代表知识蒸馏系数，用来平衡知识蒸馏损失函数 L_{KD} 与交叉熵损失函数 L_{C} 之间的重要程度。其中，L_{KD} 是让学生模型VGG16去模仿教师模型ViT概率分布的关键，其定义如下：

$$L_{\mathrm{KD}}=T^2\times\mathrm{KLdiv}(\boldsymbol{Q}_{\mathrm{stu}}^{\mathrm{T}},\boldsymbol{Q}_{\mathrm{tea}}^{\mathrm{T}}), \tag{16}$$

式中：T 代表知识蒸馏的温度，确定了模型输出预测标签的平滑程度；KLdiv表示Kullback-Leibler（KL）散度，可以测量两个分布之间的距离，其计算公式如下：

$$\mathrm{KLdiv}(p\|q)=\sum p(x)\log\frac{p(x)}{q(x)}, \tag{17}$$

式中 $p(x)$ 和 $q(x)$ 是两个分离随机变量的概率分布。

在式（16）中，与传统的Softmax（$T=1$）计算概率分布不同，$\boldsymbol{Q}_{\mathrm{stu}}^{\mathrm{T}}$ 和 $\boldsymbol{Q}_{\mathrm{tea}}^{\mathrm{T}}$ 是通过高温Softmax函数计算出的学生模型和教师模型输出结果的各类概率分布。蒸馏温度 T 越高，则计算得到的各类概率分布越平缓。$\boldsymbol{Q}_{\mathrm{stu}}^{\mathrm{T}}$ 和 $\boldsymbol{Q}_{\mathrm{tea}}^{\mathrm{T}}$ 中第 j 类的计算公式如下：

$$\boldsymbol{Q}_{\mathrm{stu}}^{\mathrm{T}}(j)=\frac{\exp(F_{\mathrm{stu}}^{j}/T)}{\sum_{b}\exp(F_{\mathrm{stu}}^{b}/T)}, \tag{18}$$

$$\boldsymbol{Q}_{\mathrm{tea}}^{\mathrm{T}}(j)=\frac{\exp(F_{\mathrm{tea}}^{j}/T)}{\sum_{b}\exp(F_{\mathrm{tea}}^{b}/T)}, \tag{19}$$

式中：F_{stu}^{j} 和 F_{tea}^{j} 分别是学生模型和教师模型所产生的logits输出 F_{stu}，F_{tea} 中的第 j 类，b 为 F_{stu} 和 F_{tea} 中输出的类别总数。

在LDGLKD网络中，L_{KD} 是将教师模型中的长距离上下文联系向学生模型转移的桥梁，L_{C} 提供了硬标签信息并与 L_{KD} 形成互补。因此，LDGLKD网络在模型收敛后，学生模型可以同时提取输入数据中长距离上下文联系和局部细节信息。表1是LDGLKD网络的运算过程。

表 1　LDGLKD 算法步骤

Tab. 1　Steps of LDGLKD algorithm

算法：轻量化深度全局－局部知识蒸馏

训练过程：

阶段 1：以小的知识蒸馏系数 α 进行训练

输入：训练集图像 $\boldsymbol{I}_{\text{in}}$，温度 T，知识蒸馏系数 α，教师模型 ViT，学生模型 VGG16，教师模型函数集合 $f_{\text{tea}}()$，学生模型函数集合 $f_{\text{stu}}()$；

输出：经过优化的教师模型与学生模型；

$F_{\text{tea}} \leftarrow f_{\text{tea}}(\boldsymbol{I}_{\text{in}})$；

$F_{\text{stu}} \leftarrow f_{\text{stu}}(\boldsymbol{I}_{\text{in}})$；

$L_{\text{KD}} \leftarrow T^2 \times \text{KLdiv}(\boldsymbol{Q}_{\text{stu}}^{\text{T}}, \boldsymbol{Q}_{\text{tea}}^{\text{T}})$，$\boldsymbol{Q}_{\text{stu}}^{\text{T}}$ 和 $\boldsymbol{Q}_{\text{tea}}^{\text{T}}$ 由公式(18)和(19)计算；

$L_{\text{fir}} \leftarrow \alpha L_{\text{KD}} + (1-\alpha)(L_{\text{C}}^{\text{tea}} + L_{\text{C}}^{\text{stu}})$；

以 L_{fir} 为损失函数对教师模型以及学生模型进行共同优化；

阶段 2：以大的知识蒸馏系数 α 进行训练

输入：训练集图像 $\boldsymbol{I}_{\text{in}}$，温度 T，知识蒸馏系数 α，教师模型 ViT，学生模型 VGG16，教师模型函数集合 $f_{\text{tea}}()$，学生模型函数集合 $f_{\text{stu}}()$；

输出：经过优化的学生模型；

$F_{\text{tea}} \leftarrow f_{\text{tea}}(\boldsymbol{I}_{\text{in}})$；

$F_{\text{stu}} \leftarrow f_{\text{stu}}(\boldsymbol{I}_{\text{in}})$；

$L_{\text{KD}} \leftarrow T^2 \times \text{KLdiv}(\boldsymbol{Q}_{\text{stu}}^{\text{T}}, \boldsymbol{Q}_{\text{tea}}^{\text{T}})$；$\boldsymbol{Q}_{\text{stu}}^{\text{T}}$ 和 $\boldsymbol{Q}_{\text{tea}}^{\text{T}}$ 由公式(18)和(19)计算；

$L_{\text{sec}} \leftarrow 2\alpha L_{\text{KD}} + (1-2\alpha) L_{\text{C}}^{\text{stu}}$；

以 I_{sec} 为损失函数对学生模型进行优化，并停止对教师模型的训练；

测试阶段：

场景分类

输入：测试集图像，$I_{\text{in}}^{\text{test}}$，学生模型 VGG16，学生模型函数集合，$f_{\text{stu}}()$；

输出：预测标签 Y_{label}；

$F_{\text{stu}} \leftarrow f_{\text{stu}}(\boldsymbol{I}_{\text{in}}^{\text{test}})$；

$Y_{\text{label}} \leftarrow \text{softmax}(F_{\text{stu}})$；

生成每一张测试集图像的语义标签 Y_{label}。

3　实验结果和分析

3.1　实验数据

为验证 LDGLKD 网络的分类性能，本文采用所构建的欧比特高光谱遥感影像场景分类数据集(Orbita Hyperspectral Image Scene Classification Dataset，OHID-SC)和公开的高光谱遥感图像数据集(Hyperspectral Remote Sensing Dataset for Scene Classification，HSRS-SC)进行验证。

3.1.1　OHID-SC 数据集

为实现高光谱数据在场景分类任务上的应用，刘康[21]和徐科杰[22]等分别构建了天宫一号高光谱遥感场景分类数据集(TianGong-1 Hyperspectral Remote Sensing Scene Classification Dataset，TG1HRSSC)和 HSRS-SC。然而，上述数据集的场景覆盖相对单一且规模较小，难以有效反映各种典型的地表覆盖情况。因此，本文根据欧比特珠海一号遥感数据服务平台所提供的 HSI 数据，构建了欧比特高光谱遥感影像场景分类数据集。首先，利用 ENVI 平台对原始 32 张蕴含着不同波段信息的遥感影像进行波段融合。其次，在该遥感影像上进行感兴趣区域(Regions of Interest，ROI)裁剪，该步骤是构建本数据集最重要的环节之一。最后，在高光谱遥感影像中剪裁出 64×64 的地物级别场景影像，并利用谷歌地球高清卫星地图对所裁剪区域进行人工目视标注，同时根据中华人民共和国土地利用现状分类国家标准为其分配场景语义标签，最终形成 OHID-SC 数据集。数据集的总体构建流程如图 3 所示。

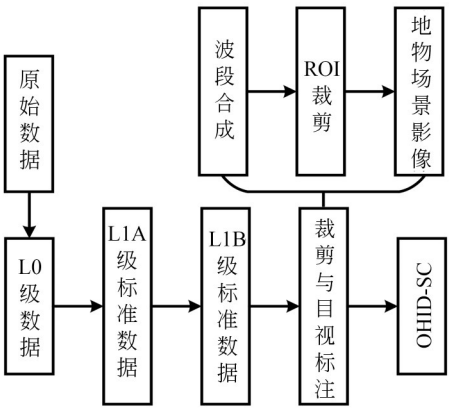


图 3　OHID-SC 数据集的总体构建流程

Fig. 3　Overall build process for OHID-SC dataset

珠海一号 OHS 系列卫星运行轨道覆盖我国境内大多数地区，OHID-SC 数据集选取其中最具有代表性的 6 类场景(城乡建筑、耕地、森林、山体、未利用土地、水体)作为语义类别，共计 2 280

张影像。该数据集影像的空间尺寸均为64×64，波段数均为32，每个语义类别的影像数量从284到439幅不等。图4提供了每个语义类别的示例影像和样本数量。

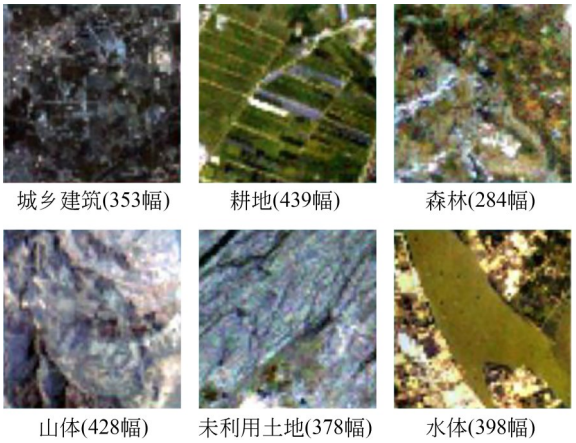


图4　OHID-SC数据集各类别的示例图像

Fig. 4　Examples of scene in constructed OHID-SC dataset

3.1.2　HSRS-SC数据集

该数据集源自黑河生态水文遥感试验航空数据。图像尺寸为256 × 256像素，波段数为48，共计1 385幅高光谱场景图像，该数据集被划分为农田、建筑、城市建筑、未利用土地和水体5个语义类别。HSRS-SC数据集各类别示例图像和数量如图5所示。

3.2　实验设置

3.2.1　实验环境

运行本实验的电脑硬件配置如下：操作系统为Windows 10专业版，内存为36 G，处理器为Intel(R) Core(TM) i7-7800X CPU，显卡型号为NVIDIA TITAN Xp。运行本实验的电脑软件环境配置如下：Python版本为python3.7.3，Pytorch版本为1.7.0。

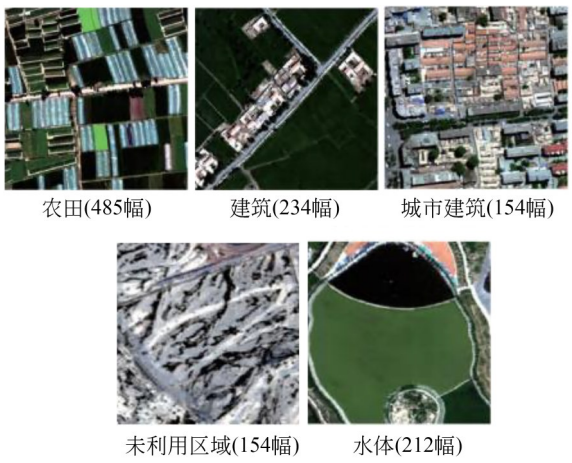


图5　HSRS-SC数据集各类别示例图像及数量

Fig. 5　Examples and numbers of scene in HSRS-SC dataset

3.2.2　参数设置

本实验所采用的ViT和VGG16模型均经过预训练。由于GPU显存的限制，batch size设置为12。教师模型(ViT)输入切割样本的空间尺寸为16×16。对于教师模型，采用SGD优化器对ViT进行微调，相关的动量设置为0.9，权重衰减设置为0.000 1。学生模型VGG16选用Adam作为优化器，权重衰减参数设置为0.000 1。教师模型中每层学习率为0.001，学生模型中的全连接层学习率设置为0.001，其余每层的学习率为0.000 1。在模型训练阶段，每经过12个轮次，学习率就会衰减为自身的0.1倍。

3.2.3　数据处理和评价指标

本实验采取两种数据增强的方法，分别为随机翻转和随机旋转。在所构建的OHID-SC数据集和公开数据集HSRS-SC中，标记样本的20%和80%被随机分配给训练集与测试集，以充分验证所提出LDGLKD的分类性能。为确保实验结果的公平性，所有实验均重复5次。此外，总体准确度(Overall Accuracy，OA)、标准差(Standard Deviation，STD)和混淆矩阵用来对分类结果进行有效评估。

3.2.4　参数敏感性分析

为验证蒸馏温度T不同取值对LDGLKD网络分类结果的影响，本文根据先验知识采用网格搜索方法从{1，5，15，25，35，45，55}中找到T的最佳取值。由图6可知，LDGLKD网络在OHID-SC数据集的OA随着T=1到T=5时迅速增加，并在T=5到T=55阶段保持稳定。这是因为当T=1时，学生模型使用硬标签进行训练，无法有效获取教师模型中的长距离依赖关系；而当T大于1时，教师模型利用软标签对学生模型进行指导，学生模型可以同时提取到正标签信息和负标签信息；此外，LDGLKD网络在T=35时取得最佳分类精度。

在LDGLKD网络的两个训练阶段中，知识蒸馏系数α用来平衡总损失函数中知识蒸馏损失

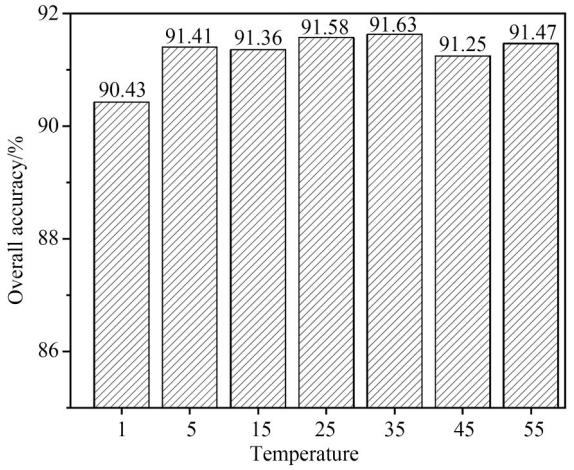


图 6　不同温度的总体分类精度

Fig. 6　OAs with respect to different temperatures

和交叉熵损失的重要程度。为获取最佳知识蒸馏系数 α，α 分别选取{0.1,0.15,0.2,0.25,0.3,0.35,0.4}进行实验。由图 7 可知，当 α 比较小（α=0.1）和比较大时（α=0.4），所得到的 OAs 较低，当 α 在 0.1～0.4 时，OAs 较高且保持稳定。当 α=0.3 时，LDGLKD 网络取得最高分类精度，这是因为合适的 α 可以有效平衡从教师模型处获得的软标签和从地面真值处获得的硬标签的重要性，并促进鉴别知识从教师模型向学生模型转移。具体来说，较小的 α 难以使学生模型充分学习来自于教师模型的远程上下文信息；当 α 太大时，学生模型又无法有效利用地面真值信息。综上所述，LDGLKD 在 OHID-SC 和 HSRSC-SC 数据集均选取 T=35 和 α=0.3 进行实验。

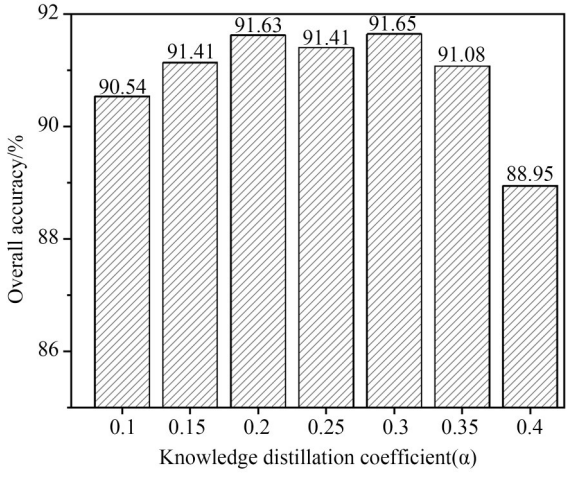


图 7　知识蒸馏系数 α 对总体分类精度的影响

Fig. 7　Effect of KD coefficient α on OAs

3.3　实验结果

3.3.1　消融实验

为评估所提出方法中各个组成部分的有效性，选取教师模型（预训练 ViT）、未经知识蒸馏的学生模型（预训练 VGG16）和知识蒸馏后的学生模型（LDGLKD）在 OHID-SC 和 HSRS-SC 数据集上进行消融实验。

如图 8 所示，因为 ViT 通过自注意力架构能学习到更具鉴别性的上下文特征，教师模型（预训练 ViT）的分类结果比预训练 VGG16 更佳。在 OHID-SC 和 HSRS-SC 数据集上，LDGLKD 比预训练 VGG16 的分类精度分别提高 9.84% 和 2.83%，这是因为 LDGLKD 网络通过知识蒸馏充分接收来自教师模型 ViT 的远程依赖关系，从而进一步提高了分类精度。由于 LDGLKD 网络通过融合长距离上下文信息和局部细节特征显著提升了对特征的判别能力，它在两个数据集上的分类性能甚至超过了教师模型。

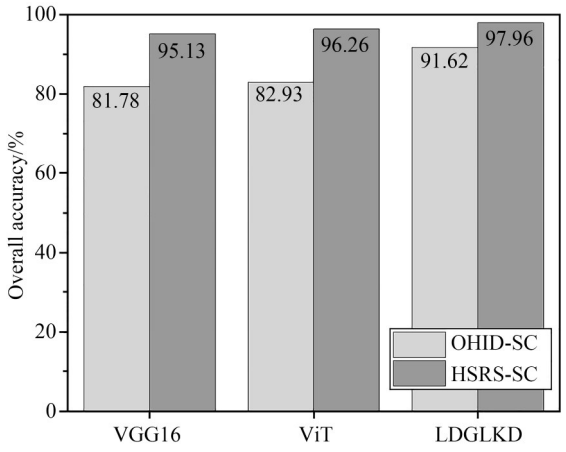


图 8　消融实验结果

Fig. 8　Ablation experiment results

3.3.2　对比实验

为验证 LDGLKD 的分类性能，这里选取 EfficientNet-B3，VGG16，GoogleNet，ResNet18 和 ResNet101 五种常用的 CNNs，且上述网络结构均经过预训练。SKAL[23]和 ARCNet[24]为两种先进的场景分类网络。其中，SKAL 是一种弱监督双流多尺度场景分类方法，SKAL-R，SKAL-V 的骨干网络分别为 ResNet 18 与 VGG16。ARCNet 通过循环注意结构聚焦于关键区域，根据不同的特征提取器 ResNet18，MobileNetV2，ARC-

Net可进一步分为ARCNet-R和ARCNet- M。

如表2所示，LDGLKD网络在OHID-SC和HSRS-SC数据集上均获得的最佳分类精度。与5种常用的CNNs相比，LDGLKD来自于教师模型的远程依赖关系蒸馏到学生模型中，与学生模型提取到的局部细节信息形成有效互补，所以LDGLKD网络的分类结果更具竞争性。值得说明的是，为尽可能模拟实际应用任务，所构建数据集OHID-SC相较于公开数据集HSRS-SC具有空间尺寸更小、观测范围广、地表覆盖状况更加丰富等特点，因此OHID-SC数据集包含的地物信息更加有限，空间布局也更为复杂。SKAL和ARCNet均为基于高分辨率影像的场景分类网络，这两种模型无法通过光谱信息来有效区分视觉感知相似的地物，导致对不同场景类别易造成错分或混淆，所以在OHID-SC数据集上两种模型的分类性能有限。

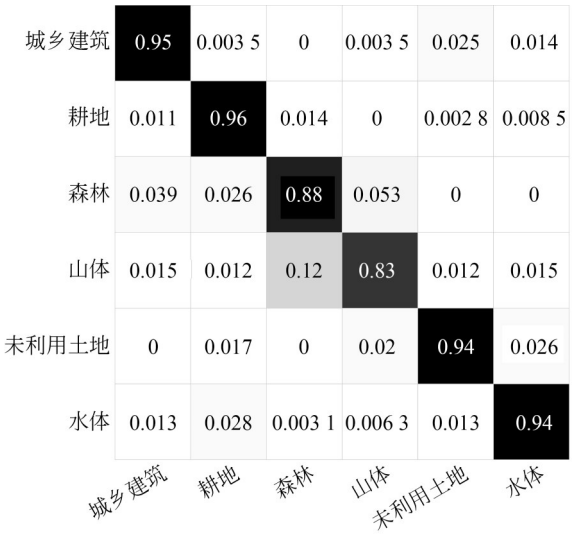
表2　对比算法实验结果[总体分类精度±标准差]

Tab. 2　Results of comparison algorithm [OA±STD]

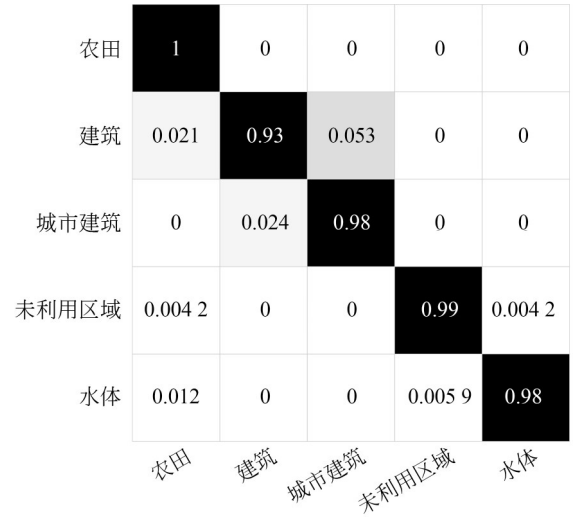
(%)

Method	OHID-SC	HSRS-SC
ResNet101	75.86±1.41	95.95±0.26
ResNet18	80.34±0.36	96.43±0.47
GoogleNet	58.17±3.80	87.81±0.04
EffcientNet	83.81±0.66	97.21±0.64
VGG16	81.78±0.50	95.13±0.22
SKAL-R	55.14±0.42	96.08±0.15
SKAL-V	59.80±0.43	96.01±0.24
ACRNet-R	60.39±0.18	95.75±0.07
ACRNet-M	65.02±0.14	97.96±0.41
LDGLKD	**91.62±0.20**	**97.96±0.04**

为更加直观地评估LDGLKD网络分类结果的类间误差，其在两个数据集上的混淆矩阵如图9所示。由于LDGLKD网络有效融合了长距离依赖关系和局部细节特征，LDGLKD网络对HSRS-SC数据集中的4类场景的分类精度均超过98%，对“农田”的分类准确度更达到了100%。在OHID-SC数据集上，LDGLKD对“城乡建筑”“耕地”“未利用土地”和“水体”四类场景的分类精度超过94%。虽然“森林”和“山体”中目标地物在颜色、光谱曲线等方面存在相似性，但LDGLKD可通过从ViT学习到的远程依赖关系对场景空间结构进一步探索挖掘，在两类场景上的分类精度达到88%和83%，在很大程度上缓解了错分和混淆现象。



(a) OHID-SC数据集
(a) OHID-SC dataset



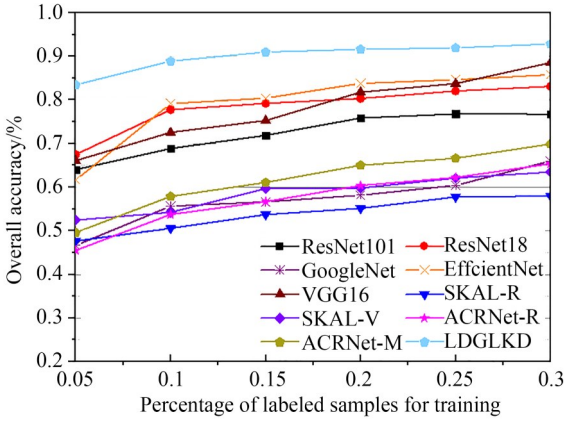
(b) HSRS-SC数据集
(b) HSRS-SC dataset

图9　LDGLKD网络的混淆矩阵结果

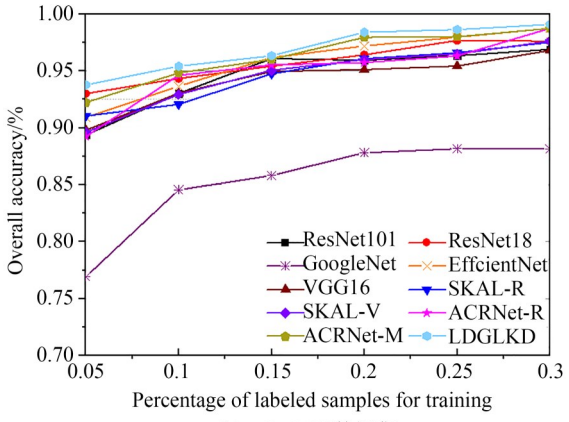
Fig. 9　Confusion matrix results of LDGLKD network

在两个高光谱场景分类数据集中，本文随机选取{5%，10%，15%，20%，25%，30%}的标记样本作为训练数据，以验证所提出方法的鲁棒性和泛化能力。如图10所示，随着训练样本百分比的增加，各方法的分类精度均明显提高。此外，

现有的场景分类模型通常针对高分辨率遥感影像所设计，如SKAL，ARCNet等，难以有效学习到高光谱影像丰富的空-谱信息。在不同数量的训练样本情况下，LDGLKD网络与其他方法相比，其分类结果具有较大优势。因为LDGLKD网络能够有效利用知识蒸馏，促使学生模型同时提取到场景图像中的远程依赖关系和局部细节特征。由于LDGLKD网络的学生模型单独进入测试阶段，并不会额外增加计算复杂度，可进一步缓解过拟合现象。



(a) OHID-SC数据集
(a) OHID-SC dataset



(b) HSRS-SC数据集
(b) HSRS-SC dataset

图10　在不同比例的训练样本下对比算法的OAs

Fig. 10　OAs with different training data percentages for comparison methods

为定量分析LDGLKD网络的运算效率，不同方法在OHID-SC和HSRS-SC数据集上的运行时间如表3所示。需要说明的是，所有方法均在同一台计算机上进行计算。基于高分辨率遥感影像的SKAL和ARCNet难以充分利用高光谱影像中丰富的空-谱信息，所以这两种模型的运行时间较短。EfficientNet-B3的矩阵计算复杂，它在OHID-SC数据集上的运行时间最长。在HSRS-SC数据集上，VGG16包含的大量可训练参数导致其运算时间较长。此外，ResNet18和ResNet101的运行时间随着网络深度的加深而增加。综合考虑时间成本和分类精度，本文选择去除全连接层的VGG16作为LDGLKD网络的学生模型，在经过知识蒸馏后单独进入测试阶段。由表3可知，LDGLKD网络具有较高的计算效率，在时间和硬件方面节省了计算资源。

表3　对比算法的运行时间

Tab. 3　Running time of comparison algorithms　(s)

Method	Dataset	OHID-SC	HSRS-SC
ResNet101	Train	9	52
	Test	45	126
ResNet18	Train	8	49
	Test	19	121
GoogleNet	Train	9	50
	Test	27	123
EfficientNet	Train	10	51
	Test	56	125
VGG16	Train	7	52
	Test	15	129
SKAL-R	Train	5	11
	Test	14	34
SKAL-V	Train	5	15
	Test	15	36
ACRNet-R	Train	2	4
	Test	5	13
ACRNet-M	Train	2	5
	Test	5	14
LDGLKD	Train	8	43
	Test	13	108

4　结　论

本文提出了一种端到端基于高光谱影像的场景分类方法LDGLKD。在LDGLKD方法中，将教师模型ViT所提取的远程上下文信息通过知识蒸馏转移到学生模型(预训练VGG16)。上

述两个模型在训练阶段实现协同优化后，逐渐减小教师模型的学习率直至为零，并增加知识蒸馏损失函数的权重系数。因此，LDGLKD方法中的学生模型能同时提取高光谱影像中远距离依赖关系和局部几何结构。

由于高光谱场景分类研究当前仍处于前期探索阶段，并且现有高光谱场景分类数据集存在波段未统一、规模较小和观测范围有限等问题，本文基于欧比特珠海一号高光谱卫星数据构建了欧比特高光谱遥感场景分类数据集（OHID-SC）。在OHID-SC数据集和公开数据集HSRS-SC上的实验表明，与一些先进方法相比，LDGLKD方法的分类精度更具竞争性，泛化性能也更佳。后续，将进一步减少LDGLKD网络的训练时间成本，并且从语义类别及样本数上对所构建数据集进行扩展。

参考文献：

[1] 闫敬文，陈宏达，刘蕾. 高光谱图像分类的研究进展[J]. 光学 精密工程，2019，27(3)：680-693.
YAN J W，CHEN H D，LIU L. Overview of hyperspectral image classification [J]. *Opt. Precision Eng.*，2019，27(3)：680-693. (in Chinese)

[2] 丁国绅，乔延利，易维宁，等. 基于高光谱图像的改进SIFT特征提取与匹配[J]. 光学 精密工程，2020，28(4)：954-962.
DING G S，QIAO Y L，YI W N，*et al*. Improved SIFT feature extraction and matching technology based on hyperspectral image [J]. *Opt. Precision Eng.*，2020，28(4)：954-962. (in Chinese)

[3] 陈静，张静. 改进高斯过程回归的高光谱空谱联合分类算法[J]. 光学 精密工程，2019，27(7)：1649-1660.
CHEN J，ZHANG J. Spectral-spatial joint classification of hyperspectral image algorithm based on improved Gaussian process regression [J]. *Opt. Precision Eng.*，2019，27(7)：1649-1660. (in Chinese)

[4] CAI Y M，LIU X B，CAI Z H. BS-nets：an end-to-end framework for band selection of hyperspectral image [J]. *IEEE Transactions on Geoscience and Remote Sensing*，2020，58(3)：1969-1984.

[5] HU F，XIA G S，YANG W，*et al*. Mining deep semantic representations for scene classification of high-resolution remote sensing imagery [J]. *IEEE Transactions on Big Data*，2020，6(3)：522-536.

[6] LUO F L，DU B，ZHANG L P，*et al*. Feature learning using spatial-spectral hypergraph discriminant analysis for hyperspectral image [J]. *IEEE Transactions on Cybernetics*，2019，49(7)：2406-2419.

[7] TONG X Y，XIA G S，LU Q，*et al*. Land-cover classification with high-resolution remote sensing images using transferable deep models [J]. *Remote Sensing of Environment*，2020，237：111322.

[8] YU Y T，LI Y Y，WANG J，*et al*. C^2-CapsViT：cross-context and cross-scale capsule vision transformers for remote sensing image scene classification [J]. *IEEE Geoscience and Remote Sensing Letters*，2022，19：1-5.

[9] 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.

[10] Bi Q，QIN K，LI Z L，*et al*. A multiple-instance densely-connected ConvNet for aerial scene classification [J]. *IEEE Transactions on Image Processing*，2020，9：4911-4926.

[11] CHENG G，CAI L M，LANG C B，*et al*. SPNet：Siamese-prototype network for few-shot remote sensing image scene classification [J]. *IEEE Transactions on Geoscience and Remote Sensing*，2022，60：1-11.

[12] LI E Z，XIA J S，Du P J，*et al*. Integrating multilayer features of convolutional neural networks for remote sensing scene classification [J]. *IEEE Transactions on Geoscience and Remote Sensing*，2017，55(10)：5653-5665.

[13] 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.

[14] XU K J，HUANG H，DENG P F，*et al*. Two-stream feature aggregation deep neural network for scene classification of remote sensing images [J]. *Information Sciences*，2020，539：250-268.

[15] FANG J，YUAN Y，LU X Q，*et al*. Robust space - frequency joint representation for remote sensing image scene classification [J]. *IEEE Transactions on Geoscience and Remote Sensing*，2019，57

(10): 7492-7502.

[16] HE X, CHEN Y S, LIN Z H. Spatial-spectral transformer for hyperspectral image classification [J]. *Remote Sensing*, 2021, 13(3): 498.

[17] BAZI Y, BASHMAL L, AL RAHHAL M M, *et al*. Vision transformers for remote sensing image classification [J]. *Remote Sensing*, 2021, 13(3): 516.

[18] ZHANG J R, ZHAO H W, LI J. TRS: Transformers for remote sensing scene classification [J]. *Remote Sensing*, 2021, 13(20): 4143.

[19] SHI C, FANG L, LV Z Y, *et al*. Explainable scale distillation for hyperspectral image classification [J]. *Pattern Recognition*, 2022, 122: 108316.

[20] HU Y T, HUANG X, LUO X Y, *et al*. Variational Self-Distillation for Remote Sensing Scene Classification [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 1-13.

[21] 刘康, 周壮, 李盛阳, 等. 天宫一号高光谱遥感场景分类数据集及应用[J]. 遥感学报, 2020, 24(9): 1077-1087.
LIU K, ZHOU Z, LI S Y, *et al*. Scene classification dataset using the Tiangong-1 hyperspectral remote sensing imagery and its applications [J]. *Journal of Remote Sensing*, 2020, 24(9): 1077-1087. (in Chinese)

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

[23] WANG Q, HUANG W, XIONG Z T, *et al*. Looking closer at the scene: multiscale representation learning for remote sensing image scene classification [J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2022, 33(4): 1414-1428.

[24] WANG Q, LIU S T, CHANUSSOT J, *et al*. Scene classification with recurrent attention of VHR remote sensing images [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2019, 57(2): 1155-1167.

作者简介：

刘英旭(2000—),男,吉林长春人,硕士研究生,2022年于重庆大学获得学士学位,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

通讯作者：

黄　鸿(1980—),男,湖南新宁人,教授,博士生导师,2003,2005,2008年于重庆大学分别获得学士、硕士和博士学位,主要从事流形学习、模式识别、遥感影像智能化处理等方面的研究。E-mail:hhuang@cqu.edu.cn