文章编号:1001-9014(XXXX)XX-0001-13

DOI:10. 11972/j. issn. 1001-9014. XXXX. XX. 001

基于知识蒸馏的轻量化遥感图像场景分类

张重阳[1,2], 王 斌[1,2*]

(1. 复旦大学 电磁波信息科学教育部重点实验室,上海,200433;
2. 复旦大学 信息学院图像与智能实验室,上海,200433)

摘要:遥感图像场景分类旨在根据遥感图像的内容为其自动赋予相应的语义标签,已成为当前遥感图像处理领域中的研究热点。基于卷积神经网络(Convolutional Neural Networks, CNNs)的方法和基于自注意力机制的方法则是当前遥感图像场景分类中的两大主流方法。然而,前者不擅长学习长程上下文关系;后者对局部信息的学习能力有限,且具有较大的参数量和运算量。针对上述问题,提议一种基于知识蒸馏的轻量化遥感图像场景分类方法。该方法分别以Swin Transformer和小型CNN网络作为教师模型和学生模型,通过知识蒸馏的方式融合两种模型的优势;更进一步,提出一种新颖的知识蒸馏损失函数,使学生模型能够同时关注遥感图像类间和类内的潜在信息。在两个大规模数据集上的实验结果表明,与现有其它方法相比,所提出方法不仅有高的分类精度,还具有显著降低的参数量和运算量。

关 键 词:遥感图像; 场景分类; 卷积神经网络; 知识蒸馏; 损失函数

中图分类号:TP751 **文献标识码**: A

Lightweight Remote Sensing Scene Classification Based on Knowledge Distillation

ZHANG Chong-Yang[1,2], WANG Bin[1,2*]

(1. Key Laboratory for Information Science of Electromagnetic Waves (MoE), Fudan University, Shanghai 200433, China;
2. Image and Intelligence Laboratory, School of Information Science and Technology, Fudan University, Shanghai 200433, China)

Abstract: Remote sensing image scene classification aims to automatically assign a semantic label to each remote sensing image according to its content, and has become one of the hot topics in the field of remote sensing image processing. Methods based on convolutional neural networks (CNNs) and methods based on self-attention mechanism are two mainstream methods in remote sensing image scene classification. However, the former is less effective in exploring long-range contextual information, and the latter has limitations in learning local information and has a large number of parameters and calculations. In order to address these issues, a lightweight method based on knowledge distillation is proposed to solve the problem of scene classification for remote sensing images. The proposed method uses Swin Transformer and lightweight CNNs as the teacher model and the student models, respectively, and integrates the advantages of the two kinds of models by means of knowledge distillation. Furthermore, a novel distillation loss function is proposed to enable the student models to focus on both inter- and intra-class potential information of remote sensing images simultaneously. The experimental results on two large-scale remote sensing image datasets demonstrate that the proposed method not only achieves high classification accuracy compared to existing methods but also has a significantly reduced number of parameters and calculations.

Key words: Remote sensing images, scene classification, convolutional neural network (CNN), knowledge distillation, loss function

PACS:84. 40. Xb

收稿日期:2023- 11- 16,**Received date**:2023- 11- 16,
基金项目:国家重点研发计划(2022YFB3903404)
Foundation items:Supported by National Key Research and Development Program of China (Grant No. 2022YFB3903404)
作者简介(Biography):张重阳(1999—),男,河南人,硕士研究生,主要研究领域为遥感图像场景分类. E-mail: 21210720041@m. fudan. edu. cn
*通讯作者(Corresponding author): E-mail: wangbin@fudan. edu. cn

引言

随着遥感成像设备和技术的进步，高分辨率遥感图像的获取变得更加容易，为遥感数据的各类应用提供了可靠的数据支持，如环境规划、城市规划、自然灾害监测等[1][2][3]。这些应用都依赖于准确的遥感图像场景分类，即根据遥感图像的内容为其自动赋予相应的语义标签。遥感图像场景分类是当前的研究热点问题之一。然而，相较于自然图像，遥感图像在拍摄角度、空间分布和图像分辨率等方面与自然图像存在显著差异，具有类内差异大而类间差异小的特点，这使得有效的特征提取对于遥感图像场景分类性能的提高尤其重要。

为了解决这一问题，大量的工作致力于为高分辨率遥感图像提取具有辨别性的特征。早期的工作主要使用了手工设计的特征，例如，尺度不变特征变换(Scale-Invariant Feature Transform, SIFT)[4]、梯度方向直方图(Histogram of Oriented Gradients, HOG)[5]和视觉词袋模型(Bag of Visual Words, BoVW)[6]等。但是，这类方法通常不能够充分提取遥感图像中丰富的语义信息，导致其性能难以满足实际应用需求。伴随着人工智能技术的快速发展，基于深度学习的方法可从原始数据中自动学习高阶语义特征，并成为遥感图像场景分类的主流。现阶段，基于深度学习的方法可大致分为两类：基于卷积神经网络(Convolutional Neural Networks, CNNs)的方法和基于自注意力机制的方法。

CNN模型善于捕捉局部信息，并且随着模型层数的增加，网络感受野也会逐渐扩大，使其在遥感图像场景分类任务中取得了良好的结果。基于CNN的方法主要以预训练好的CNN模型作为特征提取器，如VGGNet[7]或ResNet[8]，并对其添加额外的模块，或通过特定的特征融合方法来充分融合CNN中不同层的信息，使其具有更强的多尺度学习能力。在先前工作的基础上，具有八度卷积的多尺度特征融合协方差网络(Multiscale Feature Fusion Covariance Network With Octave Convolution, MF^{2}CNet)[9]通过新颖的多尺度融合方案，进一步增强了模型的多尺度和多频学习能力。然而，受限于卷积运算的局部性，CNN模型容易忽视隐藏在遥感图像中的长程依赖关系，但是，这一关系却对充分而全面地理解遥感图像至关重要。特别是，在遥感图像场景分类应用中，其不同场景图像间的类间相似性往往较高，如果模型只关注于学习局部信息，则一些场景图像可能会被误归类为其它具有相似地物的场景图像，因此，在进行遥感图像场景分类时，应充分考虑长程信息[10]。

针对上述问题，以视觉Transformer(Vision Transformer, ViT)为代表的基于自注意力机制的方法能够充分学习图像子块序列中各子块之间的关系，进一步提取遥感图像中的长程信息，开始被广泛应用于遥感图像场景分类任务[11][12][13]。然而，这一类方法仍然具有一定的局限性：1)大部分基于自注意力机制的方法通常会将输入图像划分为多个图像子块，并利用自注意力机制来建模各个图像子块之间的上下文关系，这导致它们难以充分提取遥感图像中同样至关重要的局部信息；2)基于自注意力机制的方法通常具有较高的时间和空间复杂度，这限制了它们的应用场景，难以满足计算资源受限设备(如移动设备或嵌入式设备)上实时场景识别的需求[14]。

在实际应用中，若能够充分融合CNN和Transformer的优势，使得模型既可充分提取遥感图像中的局部信息，又可充分学习到场景中各个地物间的长程关系，则可显著增强模型对遥感图像的理解；同时，若能进一步降低模型的复杂度，则可增强模型的易用性，使其具备更广泛的应用场景。基于此，我们考虑采用知识蒸馏的策略来实现信息迁移和模型压缩，以融合上述两种模型的优势，并降低模型的计算复杂度。

在目前的知识蒸馏中，通常使用大型CNN网络作为教师模型来对小型CNN网络进行蒸馏[15]，这虽然能够在一定程度上提升后者的分类精度，但并没有突破CNN模型在长程信息学习上的限制；并且，以往方法在构建蒸馏目标时，只考虑了类间关系[16]，却没有对类内关系进行建模，这限制了蒸馏模型在遥感图像场景分类任务上精度的进一步提升。

为解决上述问题，本文提出一种基于知识蒸馏的轻量化遥感图像场景分类方法(Knowledge-Distilled Lightweight Networks, KDLNet)，在提升模型易用性的同时，实现高精度的遥感图像场景分类。结合高分辨率遥感图像的特点，本文以Swin Transformer[12]作为教师模型，以充分挖掘遥感图像中的长程上下文信息，并通过知识蒸馏的方式将其传递给一个小的学生模型(如ResNet-18)，使得学生模型能够充分融合两种模型的优势；更进一步，本文

提出一种新颖的知识蒸馏损失函数,它能够促使学生模型在接收教师传递的信息时,同时关注各个类别的类间差异和类内差异,以增进对遥感图像的全面理解。相较于当前的基于深度学习的遥感图像场景分类方法,所提议方法能够同时关注遥感图像中的局部信息和长程信息,实现高精度的遥感图像场景分类,并且具有显著减小的参数量和运算量。

本文的主要贡献可简要总结如下:

1) 以Transformer为教师模型,通过知识蒸馏将潜在知识传递给(小型CNN网络的)学生模型,使得学生模型能兼具两类模型的优势,且大大降低了复杂度;

2) 提出一种新颖的知识蒸馏损失函数,能使学生模型同时关注各类别的类间差异和类内差异,进一步提升了蒸馏效果。

1　相关工作

本节通过介绍知识蒸馏的主要思路,进一步阐述其在信息迁移和模型压缩上的机制。知识蒸馏本质上属于迁移学习的范畴。传统知识蒸馏的主要思路是通过最小化教师模型和学生模型的预测分布间的差异,将知识从预先训练过的教师模型转移到学生模型中[16],如图1所示。从知识蒸馏模型的演化来看,“知识”最先出现在教师模型的输出层,也即模型输出的logits信息,这些信息包含了教师模型对输入的判断,这些判断经过softmax层后便对应着模型对各个类别的预测概率。但是,直接输出类别概率会忽视类别间的相似性,很大程度地影响了模型的泛化性能。因此,引入温度系数T来软化模型softmax输出的分类信息[16]。假设B和N分别表示输入图片的批处理大小和类别的数量,经过软化的教师模型和学生模型的分类预测如下式所示:

$$\boldsymbol{Y}_{i,:}^{(t,T)}=\mathrm{softmax}(\boldsymbol{Z}_{i,:}^{(t)}/T),\quad \boldsymbol{Y}_{i,:}^{(s,T)}=\mathrm{softmax}(\boldsymbol{Z}_{i,:}^{(s)}/T) \tag{1}$$

其中,$\boldsymbol{Z}_{i,:}^{(t)}\in\mathbb{R}^{B\times N}$和$\boldsymbol{Z}_{i,:}^{(s)}\in\mathbb{R}^{B\times N}$分别表示教师模型和学生模型输出的logits信息,$\boldsymbol{Y}_{i,:}^{(t,T)}$和$\boldsymbol{Y}_{i,:}^{(s,T)}$分别表示经过软化的教师模型和学生模型的分类预测信息。

通常,知识蒸馏损失函数为:

$$\mathcal{L}_{\mathrm{KD}}=\frac{T^2}{B}\sum_{i=1}^{B}\mathrm{KL}\left(\boldsymbol{Y}_{i,:}^{(t,T)},\boldsymbol{Y}_{i,:}^{(s,T)}\right)=\frac{T^2}{B}\sum_{i=1}^{B}\sum_{j=1}^{N}\boldsymbol{Y}_{ij}^{(t,T)}\log\left(\frac{\boldsymbol{Y}_{ij}^{(t,T)}}{\boldsymbol{Y}_{ij}^{(s,T)}}\right) \tag{2}$$

其中,KL(·)表示Kullback-Leibler散度,用以衡量两个分布间的差异。

学生模型通常由公式(2)中的软目标和真实标签一同训练。通过共同优化两个目标,不仅可以让学生模型逼近真实标签,还能引导学生学习并拟合教师模型的概率分布。整体损失函数由蒸馏损失$\mathcal{L}_{\mathrm{KD}}$和分类损失$\mathcal{L}_{\mathrm{cls}}$共同构成,即:

$$\mathcal{L}=\alpha\mathcal{L}_{\mathrm{KD}}+(1-\alpha)\mathcal{L}_{\mathrm{cls}} \tag{3}$$

在上式中,α是平衡因子,$\mathcal{L}_{\mathrm{cls}}$通常是学生模型预测值和真实标签之间的交叉熵(Cross Entropy)损失:

$$\mathcal{L}_{\mathrm{cls}}=-\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{N}\boldsymbol{C}_{i,j}\log\left(\boldsymbol{Y}_{i,j}^{(s,1)}\right) \tag{4}$$

其中,$\boldsymbol{C}_{i,j}$是真实标签,$\boldsymbol{Y}_{i,j}^{(s,1)}$表示温度值$T$=1时学生模型的分类预测。

2　模型构建

本文旨在设计一种基于知识蒸馏的轻量化遥感图像场景分类方法,其整体流程如图2所示,由教

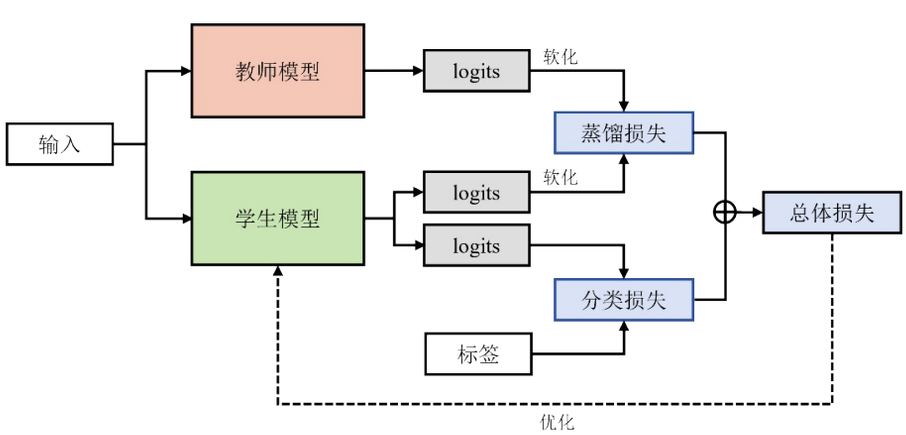


图1　知识蒸馏流程图

Fig. 1　The flowchart of knowledge distillation

师模型、学生模型和知识蒸馏三部分组成。首先，我们对所选用的教师模型和学生模型进行介绍，然后，对知识蒸馏的详细过程进行描述和分析，并给出蒸馏过程中使用的优化目标函数。

2.1 教师模型和学生模型

由于Transformer模型能通过多头自注意力机制学习序列间的长程依赖关系，在图像分类任务中展示出了较为突出的性能，其多种Transformer模型都曾被应用于遥感图像场景分类任务。相较于ViT，Swin Transformer因其特有的分层特征图和转移窗口注意力（Shifted Window Attention）机制而具有更强的多尺度特征学习和整体建模能力，在分类任务上取得了更好的性能。本文以Swin Transformer作为教师模型，将知识传递给学生模型。

作为使用最为广泛的深度学习模型，CNN模型通常由卷积层、池化层和全连接层等组成，通过逐层堆叠的形式，逐步提取出图像中抽象的高阶语义特征，且CNN模型具有良好的局部特征学习能力。将Transformer模型所具有的长程关系学习能力传递给CNN模型，能够使得CNN同时关注遥感图像中的局部信息和全局信息，从而可使模型更全面地认知和理解遥感图像。本文以ResNet-18[8]，MobileNetV3[17]，EfficientNet[18]三个轻量化网络作为学生模型。其中，ResNet-18是ResNet系列中参数最少、层数最浅的模型；MobileNetV3是Google团队继MobileNetV2之后的改进版本，也是当前应用最广泛的轻量化网络之一；EfficientNet运用了复合缩放、双线性缩放、深度可分离卷积和自动化选择网络结构参数等，是一种性能强大且高效的轻量化网络。

对于每个模型的最后一个特征图，应用全局平均池化（Global Average Pooling，GAP）和全连接层来获取模型输出的logits信息。GAP能够将特征图 $U \in \mathbb{R}^{W\times H\times C}$ 转换为特征向量 $[v_1, v_2, \ldots, v_C] \in \mathbb{R}^C$，

$$v_c = \frac{1}{W\times H}\sum_{i}^{W}\sum_{j}^{H} u_{i,j,c} \quad (c = 1,2,...,C) \tag{5}$$

其中，W，H和C分别表示输出特征图的宽度、高度和通道数。

2.2 优化目标

相较于自然图像，遥感图像通常具有类内差异大而类间差异小的特点，这对遥感图像的场景分类带来了较大的挑战。我们考虑，在每个遥感图像场景类别中，多个实例预测分数的分布也是信息丰富且有用的，这些分数能够反映多个示例与某个类别的相似度。例如，一个批次输入了3张类别分别为“公园”、“学校”和“飞机场”的遥感图像，它们在“学校”类别上有3个预测分数，分别记为 y_1, y_2, y_3。那么，“学校”图片对“学校”类别的得分最高，而“飞机场”对该类别的得分最低，因为它与“学校”之间的相似地物最少。这种“$y_2 > y_1 > y_3$”的关系也可以传递到学生身上。此外，即使是同一类别的图像，教师模型对它们的预测得分也可能是不同的，这反映了教师模型的先验知识对图像可靠性的判断，即哪一张图片更可能属于该类别。因此，我们认为，在学习类间关系的同时，学习类内关系能使学生模型更好地关注到遥感图像的细致差异。然而，如公式(2)和(3)所示，目前的知识蒸馏方法通常只考虑不同类别间的类间关系，却不考虑这种类内信息。

为使学生同时学习类间关系和类内关系，我们首先将公式(2)中的原始蒸馏损失记为类间（inter-class）蒸馏损失 $\mathcal{L}_{\text{inter}}$，

$$\mathcal{L}_{\text{inter}} = \frac{T^2}{B}\sum_{i=1}^{B}\mathrm{KL}\left(\boldsymbol{Y}_{i,:}^{(t,T)}, \boldsymbol{Y}_{i,:}^{(s,T)}\right) \tag{6}$$

该目标使学生逼近教师的输出在每一行上的分布，对应于类间关系。

我们在公式(6)的基础上进行改写，使学生模型关注于教师模型的输出在每一列上的分布，得到用于蒸馏类内关系的类内（intra-class）蒸馏损失 $\mathcal{L}_{\text{intra}}$，其数学表示为：

$$\mathcal{L}_{\text{intra}} = \frac{T^2}{N}\sum_{j=1}^{N}\mathrm{KL}\left(\boldsymbol{Y}_{:,j}^{(t,T)}, \boldsymbol{Y}_{:,j}^{(s,T)}\right) \tag{7}$$

从而，蒸馏损失函数 $\mathcal{L}_{\text{KD}}$ 可由类间蒸馏损失 $\mathcal{L}_{\text{inter}}$ 和类内蒸馏损失 $\mathcal{L}_{\text{intra}}$ 两部分构成：

$$\mathcal{L}_{\text{KD}} = \mathcal{L}_{\text{inter}} + \mathcal{L}_{\text{intra}} \tag{8}$$

此外，由于蒸馏损失与分类损失所关注的信息并不一致，同时优化分类损失与蒸馏损失会限制学生模型在遥感图像场景分类任务上的性能。具体而言，蒸馏损失会使学生更加关注于教师模型所传递的潜在信息，如类间类内差异，对应着“软标签”。而分类损失会迫使学生向真实标签逼近，真实标签通常是一个独热（One-hot）码，信息单一，与设立软标签的目标相悖。相对于纯标签的学习，遥感图像场景分类任务更期望学生模型能够学习遥感图像中复杂的潜在信息，这有助于帮助学生模型更全面充分地理解遥感图像。因此，我们弃用公式(3)中的同时优化蒸馏损失和分类损失的策略，而只在蒸馏的过程中利用软标签对学生模型进行优化。在后续实验中，我们充分验证了这种策略的有效性。

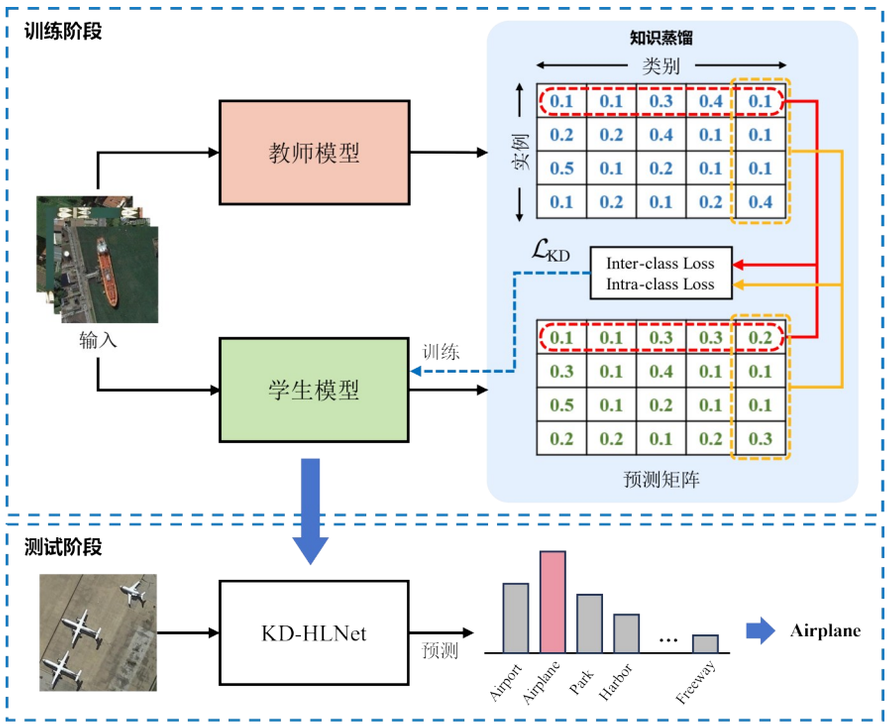


图2　所提出KDLNet的框架。

Fig. 2　The framework of the proposed KDLNet.

最终,所提出模型的整体损失函数如下所示:

$$\mathcal{L} = \mathcal{L}_{KD} \tag{9}$$

为将教师模型中的潜在信息迁移到学生模型中,选用离线蒸馏的策略来对学生模型进行优化。具体而言,首先,采用交叉熵损失 $\mathcal{L}_{cls}$ 来对教师模型进行优化;教师模型达到收敛后,将其用于对学生模型的蒸馏。蒸馏过程中,教师模型只进行推理而不更新参数,学生模型在每个训练周期都从教师模型获取固定不变的潜在信息。这种处理方式的好处是,在蒸馏的过程中,只需要关注学生模型的学习,使得训练过程的部署简单可控,大大降低了训练成本和资源消耗。

3　实验结果与分析

本节在两个大规模的公开数据集上对所提议方法的性能进行评估。首先,我们对数据集、评价指标和实验设置的具体细节进行介绍。然后,将所提出的KDLNet与现存的遥感图像场景分类方法的分类结果进行对比,并对所提出的KDLNet进行了参数实验和消融实验。最后,本节对所提议方法进行可视化分析,并给出了其与其它常见网络参数量和运算量的对比。

3.1　实验数据

实验数据采用了 Aerial Image dataset (AID)[①]和NWPU-RESISC45[②]两个经典大规模数据集[19][20],它们采集于不同国家和地区,拍摄于不同的成像平台及气候条件下,包含了丰富的场景类别。其中,AID数据集包含10000张尺寸为600×600像素的图像,空间分辨率从8 m到0.5 m不等,每类样本有220到420张图片。NWPU-RESISC45数据集包含45个场景类别,每类有700张图像,共31500张,每幅图像的尺寸为256×256像素,空间分辨率从30 m到0.2 m不等。这两个数据集中的图像均具有空间布局复杂、纹理信息丰富的特点。相较于AID数据集,NWPU-RESISC45数据集的空间分辨率变化范围更大,类内多样性和类间相似性更显著,因而更具挑战性。

上述两个数据集的样例图像如图3所示,两个数据集中的图像均同时具有空间信息丰富、地物分布复杂的特点。在实验中,根据已有工作的数据集划分方式[13][21][22],对于AID数据集,每类场景图像中分别随机选取20%和50%的样本作为训练集,其余

①https://captain-whu.github.io/AID/

②http://www.escience.cn/people/JunweiHan/NWPU-RESISC45.html

作为测试集；对于NWPU-RESISC45数据集，训练集比例分别设为10%和20%，其余作为测试集。

3.2 评价指标和实验设置

在实验中，采用总体准确率（Overall Accuracy，OA）和混淆矩阵作为评价指标来评估模型的分类精度。其中，OA为测试集中被正确分类的样本数占总样本数的比例；混淆矩阵则以表格的形式通过百分比反映每个场景类别中正确分类和错误分类的图像数目比例，是一种全面且直观的评价指标。此外，采用总体参数量和浮点运算次数（Floating Point Operations，FLOPs）来分别衡量模型的参数量（空间复杂度）和运算量（计算复杂度）。为保证结果的可靠性，所有实验均重复5次，并在实验结果中报告其平均值和标准差。

本文所提出的模型在Pytorch框架上构建，并使用GeForce RTX 3090的单个GPU对模型进行训练，其GPU具有24GB的内存。教师模型和学生模型均采用在ImageNet数据集上预训练的权重来初始化。模型训练过程中，采用自适应矩估计优化器（Adaptive Moment Estimation optimizer，Adam）来进行优化，学习率为5.0×10^{-5}，批处理大小设置为32，总训练轮次为50个epoch，并采用具有线性预热的余弦衰减学习率调度器（预热期为10个epoch）来动态调整学习率，以保证训练过程的稳定。为满足不同网络的尺寸需求，所有输入图像的尺寸都被调整为224×224。训练时，对数据集进行了数据增广，对每张图像进行了水平翻转、垂直翻转和随机旋转。此外，我们将T设置为20，该参数值的具体确定方法将在超参数分析实验中进行说明。

3.3 与其它方法的对比

为验证所提出方法的分类精度，基于AID和NWPU-RESISC45数据集，将所提出的KDLNet的分类结果与其它遥感图像场景分类方法进行对比。对比方法均为基于CNN或自注意力机制的深度模型，包括五种基线模型（VGGNet-16、ResNet-50、ResNet-101、ResNet-152、ViT）和几种现存的用于遥感图像场景分类的SOTA方法（SCViT[13]，T-CNN[21]和MGSNet[22]）。其中，SCViT充分考虑了高分辨率遥感图像中详细的几何信息和用于分类的token中不同通道的贡献，进一步提升了ViT模型的分类精度。而T-CNN和MGSNet均是在CNN的基础上进行改进的方法。T-CNN提出了一种用于迁移CNN模型的自适应学习策略，能够使模型适应源域和目标域间的差异，其分类精度相较于基线方法取得了较大提高。MGSNet通过背景信息的利用、对比正则化和自导网络，缓解了遥感图像场景分类中背景和目标之间不平衡的特征差异和目标内部表现不一致等问题，在遥感图像场景分类任务中取得了优异的分类性能。

表1给出了在两个大规模数据集上不同方法的分类精度，其中，为了保证公平对比，上述五种基线模型的分类结果均是在本文实验条件下得出的，其它方法的分类结果来自其原始论文。通过比较所提出方法与其它方法，可以发现KDLNet在分类精度上实现了较大幅度的提升。无论学生模型是ResNet-18、MobileNetV3还是EfficientNet，KDLNet均取得了优异的分类准确率，例如，KDLNet（EfficientNet）在AID数据集和NWPU-RESISC45数据集的不同训练比例下上分别取得了96.01%、97.32%、93.18%和95.12%的分类精度。这表明：KDLNet可融合来自教师模型传递的潜在知识和自身属性，实现高精度的遥感图像场景分类。

3.4 分析

3.4.1 所提议方法的有效性

本文所提议的方法是一种基于知识蒸馏策略的方法，目的在于将基于注意力机制的教师模型中所包含的潜在信息迁移到轻量化的学生模型中，提升学生模型的分类精度。因此，本节考察所提议方

图3 遥感图像样例图像：(a) AID数据集 (b) NWPU-RESISC45数据集

Fig. 3 Samples of remote sensing images: (a) AID dataset (b) NWPU-RESISC45 dataset

表 1　各种方法在 AID 数据集和 NWPU-RESISC45 数据集上的分类精度

Table 1　OA of different methods on AID dataset and NWPU-RESISC45 dataset with different training ratios

Method	AID		NWPU-RESISC45	
	Tr=20%	Tr=50%	Tr=10%	Tr=20%
Fine-tuned VGGNet-16	92.75±0.38	95.32±0.19	90.11±0.09	93.27±0.15
Fine-tuned ResNet-50	94.28±0.27	96.25±0.24	91.41±0.22	93.83±0.11
Fine-tuned ResNet-101	94.12±0.40	96.35±0.31	91.50±0.25	94.07±0.09
Fine-tuned ResNet-152	94.99±0.24	96.90±0.13	92.52±0.14	94.40±0.17
Fine-tuned ViT-B	93.54±0.28	95.24±0.24	89.58±0.20	91.89±0.09
SCViT[13]	95.56±0.17	96.98±0.16	92.72±0.04	94.66±0.10
T-CNN[21]	94.55±0.27	96.27±0.23	90.25±0.14	93.05±0.12
MGSNet[22]	95.46±0.21	97.18±0.16	92.40±0.16	94.57±0.12
KDLNet (ResNet-18)	95.25±0.29	96.81±0.20	92.11±0.15	94.05±0.23
KDLNet (MobileNetV3)	**95.68　0.17**	97.11±0.22	**92.87　0.13**	**94.88　0.06**
KDLNet (EfficientNet)	**96.01　0.25**	**97.32　0.18**	**93.18　0.20**	**95.12　0.13**

法对模型分类精度提升的有效性。

最常见的训练策略是基于预训练参数的微调(Fine-tune),即:首先,在大规模数据集(如 ImageNet[23])上对模型进行预训练,使得模型学习到通用的特征表示;然后,在下游任务上进行微调,以适应特定的应用需求。通过这种策略,可以大幅减少训练时间,同时提升模型的性能和泛化能力。本文将基于微调所得到的结果作为基线,并与其进行对比。

表 2 和表 3 分别给出了所提议方法与基线方法在 AID 和 NWPU-RESISC45 数据集上的分类结果。相比于基线方法,所提议方法能大幅提高分类精度。以 EfficientNet 为例,在 AID 数据集的两种训练比例下,所提议模型的 OA 提升分别为 2.06% 和 0.74%;在 NWPU-RESISC45 数据集上,所提议方法的 OA 提升分别为 1.95% 和 1.14%。值得注意的是,经过蒸馏后,学生模型甚至能够取得超越教师的模型的分类精度。这是因为,擅长提取局部信息的 CNN 模型接收了 Transformer 模型传递的长程信息,使得其同时兼具了局部信息和长程信息的学习能力,可更加全面地理解遥感图像。综上,在两个数据集上的实验结果表明,所提议方法在遥感图像场景分类任务中是有效的。

此外,图 4 给出了 KDLNet (EfficientNet) 与 Fine-tuned EfficientNet 训练过程中在测试集上分类精度的变化趋势。可以看到,所提出的 KDLNet 与基线方法具有相似的收敛速度,且分类精度始终保持较为显著的提高。

除了 OA 结果,混淆矩阵可进一步验证 KDLNet 的性能。以 KDLNet (EfficientNet) 为例,图 5 展示了所提议的 KDLNet 在 AID 和 NWPU-RESISC45 数据

表 2　所提议方法在 AID 数据集上对基线方法的提升

Table 2　The increase of the proposed method over baseline methods on the AID dataset

Teacher	OA		Student	OA			
	Tr=20%	Tr=50%		FT, Tr=20%	Ours, Tr=20%	FT, Tr=50%	Ours, Tr=50%
Swin-Transformer	95.95±0.35	97.18±0.23	ResNet-18	94.22±0.24	95.25±0.29 (1.03↑)	96.24±0.12	96.81±0.20 (0.57↑)
			MobileNetV3	93.46±0.36	95.68±0.17 (2.22↑)	96.21±0.27	97.11±0.22 (0.90↑)
			EfficientNet	93.95±0.36	96.01±0.25 (2.06↑)	96.58±0.18	97.32±0.18 (0.74↑)

表 3　所提议方法在 NWPU-RESISC45 数据集上对基线方法的提升

Table 3　The increase of the proposed method over baseline methods on the NWPU-RESISC45 dataset

Teacher	OA		Student	OA			
	Tr=10%	Tr=20%		FT, Tr=10%	Ours, Tr=10%	FT, Tr=20%	Ours, Tr=20%
Swin-Transformer	93.23±0.17	94.90±0.14	ResNet-18	91.02±0.24	92.11±0.15 (1.09↑)	93.31±0.15	94.05±0.23 (0.74↑)
			MobileNetV3	90.69±0.37	92.87±0.13 (2.18↑)	93.56±0.09	94.88±0.06 (1.32↑)
			EfficientNet	91.23±0.21	93.18±0.20 (1.95↑)	93.98±0.13	95.12±0.13 (1.14↑)

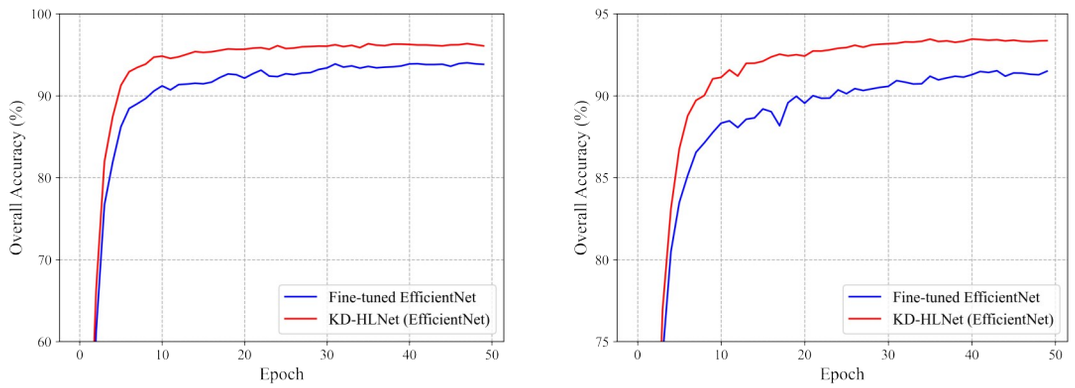


图4 所提出的KDLNet和Fine-tuned EfficientNet在AID数据集(20% 训练样本)和NWPU-RESISC45数据集(10% 训练样本)上的测试集精度: (a) AID数据集 (b) NWPU-RESISC45数据集

Fig. 4 Test accuracy on AID dataset (20% training images) and NWPU-RESISC45 (10% training images) with the proposed KDLNet and the Fine-tuned EfficientNet: (a) AID dataset (b) NWPU-RESISC45 dataset

上的混淆矩阵。从图5(a)中可以看到,KDLNet在绝大多数场景类别上都能取得优越的分类性能(≥98%),例如“Beach”、“Farmland”和“Pond”等类别;在AID数据集的两种训练比例下,整体分类准确率均高于95%。图5(b)表明,尽管NWPU-RESISC45数据集更加具有挑战性,KDLNet依旧能够在绝大多数类别上取得优良的准确率(≥95%)。除此之外,该数据集中的“Palace”和“Church”类别,因其具有相似的单体对象和结构特征而易被混淆;总体上,在这两个类别上,无论训练比例为10%还是20%,KDLNet均取得了可接受的分类结果(≥65%)。

3.4.2 超参数分析

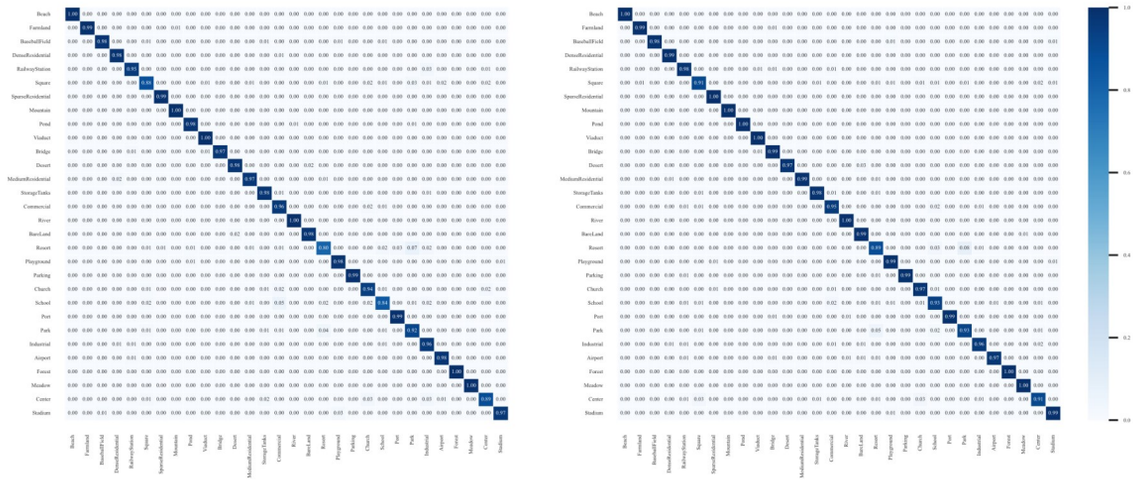
在所提议的KDLNet中,损失函数中温度系数T的确定对于准确分类至关重要。本节考察$\mathcal{L}_{KD}$中温度系数T对于分类结果的影响。在实验中,选择EfficientNet作为实验对象,在AID数据集(Tr=20%)和NWPU-RESISC45数据集(Tr=10%)上分析T的不同取值对分类结果的影响,从而确定合适的参数取值,其实验结果如表4所示。由表4可见,随着T的增大,OA呈现“先增加后趋于稳定”的趋势。当T在1~40之间变化时,模型的准确率在初期急剧增加,而后趋于稳定。这是因为,当T=1时,学生接收到的信息是来自于教师的硬标签,不能充分向学生传递潜在信息(如长程信息);随着T的增大,教师所传递的标签被软化,更多潜在的上下文信息得以充分传递。实验结果表明,KDLNet可以在一个较宽的温度系数T变化范围内取得良好的结果,当T=20时,模型在两个数据集上都能够取得最好的分类结果。因此,在本文的所有实验中,我们都将T设置为20。

进一步,我们验证了公式(3)中平衡因子α对分类结果的影响,以证明在蒸馏的过程中,舍弃分类损失能使学生模型在遥感图像场景分类任务中取得更好的结果。图6展示了KDLNet(EfficientNet)的分类结果随α的变化。从图6中可以看到,随着α的增大,分类准确率均逐渐上升。当α=1时,两个数据集上的分类准确率均达到最大值。在蒸馏的过程中,$\mathcal{L}_{cls}$占比越小,蒸馏所得的效果越好,这表明其分类损失$\mathcal{L}_{cls}$的存在对学生模型的训练产生了负面影响。当$\mathcal{L}_{cls}$不参与优化过程时,能使学生最充分地学到教师传递的知识,最大程度上提升遥感图像场景分类的精度。

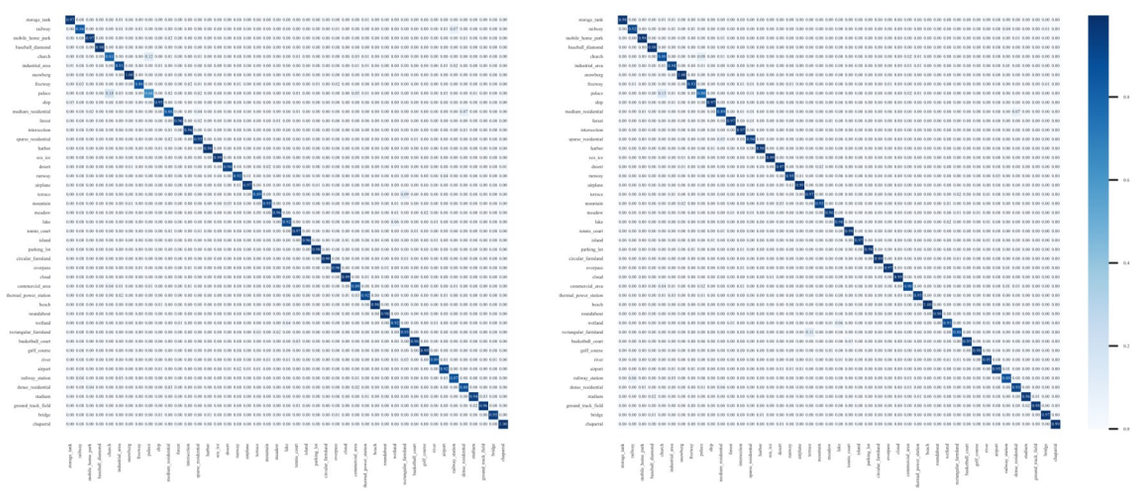
3.4.3 消融实验

为了评估所提出模型中各个部分对结果带来的提升效果,在有分类损失和无分类损失两种情况下进行了消融实验,学生模型选择为EfficientNet。

对于有分类损失的情况,从基线模型(Baseline)开始,分别对其在训练的过程中加入类间蒸馏损失$\mathcal{L}_{inter}$、类内蒸馏损失$\mathcal{L}_{intra}$和改进的蒸馏损失($\mathcal{L}_{inter}$+$\mathcal{L}_{intra}$),平衡因子α设置为0.5,以验证蒸馏损失能够在基线模型的基础上进一步提升分类效果;对于无分类损失的情况,即平衡因子α为1时,我们分别采用$\mathcal{L}_{inter}$、$\mathcal{L}_{intra}$和$\mathcal{L}_{inter}$+$\mathcal{L}_{intra}$对学生模型进行训练。具体分类结果如表5所示。



(a)



(b)

图5　KDLNet (EfficientNet) 在 AID 和 NWPU-RESISC45 数据集上的混淆矩阵: (a) AID 数据集,左侧为 20% 训练样本,右侧为 50% 训练样本 (b) NWPU-RESISC45 数据集,左侧为 10% 训练样本,右侧为 20% 训练样本

Fig. 5　Confusion matrices of the proposed KDLNet based on EfficientNet on the AID and NWPU-RESISC45 datasets: (a) AID dataset, Tr=20% (left), and Tr=50% (right), (b) NWPU-RESISC45 dataset, Tr=10% (left), and Tr=20% (right).

从表5中可以得出如下结论:1)类内蒸馏损失 $\mathcal{L}_{\text{intra}}$ 能够向学生模型传递潜在的类内关系,大幅提升分类精度,且无论是否包含分类损失,在多数情况下,类内蒸馏损失展示出了比类间蒸馏损失更好的效果;2)蒸馏过程中,同时优化类间蒸馏损失和类内蒸馏损失,能最大程度地将教师模型具备的潜在知识传递给学生模型,进一步提升学生模型的分类精度;3)两组实验的对比,进一步验证了当 $\mathcal{L}_{\text{cls}}$ 不参与优化过程时,学生模型能够充分地学习到教师模型传递的潜在信息。以上表明,所提议模型能充分关注遥感图像场景类别间的关系和类内的差异,取得更好的分类结果。

3.4.4　参数量和运算量分析

本节对所提出模型的参数量和运算量进行分析。表6分别对比了目前常见模型及本文模型的参数量和 FLOPs。结果表明,VGGNet-16 网络具有最大的参数量和较大的运算量,包括 ViT-B 和 Swin-Transformer (Base)在内的 Transformer 模型也都具有较大的参数量和运算量。而本文所提出的方法中,学生模型均是轻量化网络,在参数量和 FLOPs 上都有大幅度的降低,例如,KDLNet (EfficientNet)的参数量和 FLOPs 仅为 5.3M 和 0.42G,远远小于本文所

表4 超参数T的取值分析

Table 4 Analysis of hyper-parameter T

T	OA	
	AID (Tr=20%)	NWPU-RESISC45 (Tr=10%)
1	95.19±0.33	92.58±0.15
5	95.90±0.29	93.16±0.19
10	95.97±0.29	93.13±0.19
15	95.98±0.25	**93.18 0.20**
20	**96.01 0.25**	**93.18 0.20**
25	95.92±0.19	93.15±0.24
30	95.95±0.24	93.16±0.24
35	95.94±0.22	93.13±0.17
40	95.95±0.26	93.12±0.22

采用的教师模型Swin Transformer的87.70M和15.17G。结合表1-2和表3的结果可知,所提出的网络不仅具有显著减小的参数量和运算量,还能取得比许多大型网络更好的分类精度,这对实际应用更具应用潜力。

3.4.5 可视化结果

模型在遥感图像上所关注的区域可通过类别激活映射(Class Activation Mapping, CAM)[24]以热图的形式进行直观的展示。为了更直观地展示所提出方法对遥感图像的理解效果,图7以ResNet-18为例,给出了基线方法(Fine-tuned ResNet-18)与所提出方法的热图输出。图7的第一行均为来自NWPU-RESISC45数据集的原始图像,第二、三行分别展示了基线方法和所提出方法对应的热图输出。

表5 不同数据集上损失函数中各部分提升效果比较

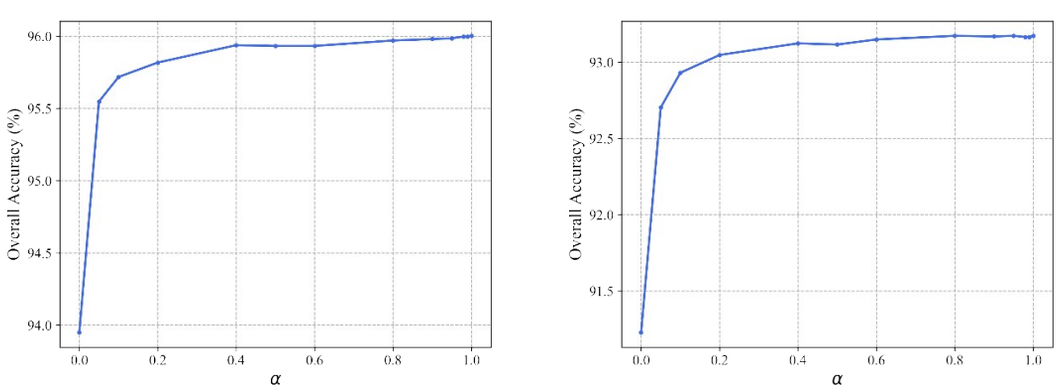
Table 5 Detialed performance comparison of each component of the loss function on different datasets

Method	OA	
	AID (Tr=20%)	NWPU-RESISC45 (Tr=10%)
Baseline ($\mathcal{L}_{cls}$)	93.95±0.36	91.23±0.21
$+\mathcal{L}_{inter}$	95.75±0.10	93.02±0.21
$+\mathcal{L}_{intra}$	95.82±0.17	93.05±0.26
$+\mathcal{L}_{inter}+\mathcal{L}_{intra}$	**95.93 0.23**	**93.11 0.22**
$\mathcal{L}_{inter}$	95.81±0.17	93.13±0.19
$\mathcal{L}_{intra}$	95.86±0.14	93.12±0.20
$\mathcal{L}_{inter}+\mathcal{L}_{intra}$	**96.01 0.25**	**93.18 0.20**

可以看到,在田径场(Ground Track Field)场景中,KDLNet比基线方法所关注的区域更加准确,能够完整地感知到图中分散的两个田径场,而基线方法只能够完整地感知到其中一个田径场。这表明所提议方法具有更强的长程上下文信息学习能力,能更全面地理解遥感图像。此外,从图7中还可以观察到,所提议方法普遍拥有更大范围的激活区域,这表明所提出的网络能更有效地捕捉到图像中不同子区域的信息,进而取得更高的分类精度。

3.4.6 小结

由以上实验结果可知,本文所提出的KDLNet在两个大规模数据集上均具有良好的分类性能,其主要特点可总结如下:



(a) (b)

图6 平衡因子α在AID数据集(20%训练样本)和NWPU-RESISC45数据集(10%训练样本)上对分类精度的影响: (a) AID数据集 (b) NWPU-RESISC45数据集

Fig. 6 The impact of the balancing factorαon the classification accuracy on the AID dataset (20% training images) and NWPU-RESISC45 dataset (10% training images): (a) AID dataset (b) NWPU-RESISC45 dataset

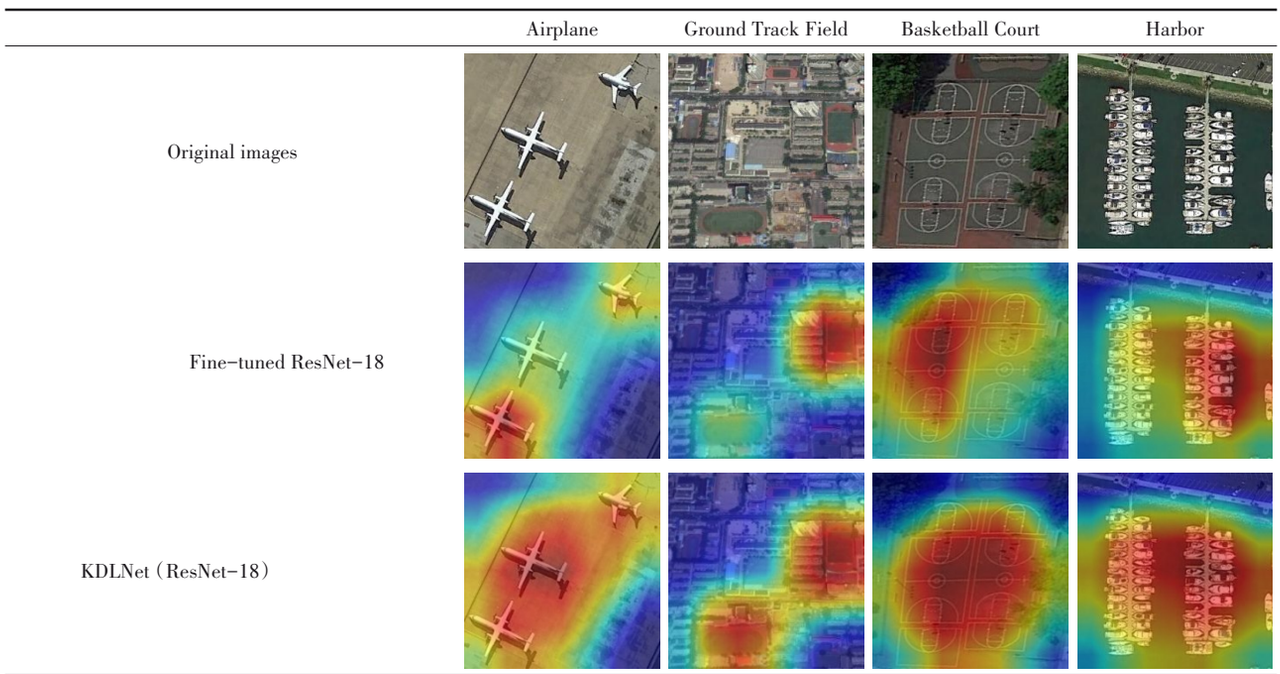
表6　不同网络的参数量和浮点运算次数

Table 6　Parameters and FLOPs of different networks

Methods	Parameters	FLOPs
VGGNet-16[7]	134.38M	15.47G
ResNet-50[8]	23.57M	4.13G
ResNet-101[8]	44.55M	7.87G
ResNet-152[8]	60.19M	11.60G
ViT-B[11]	85.68M	16.86G
Swin-Base[12]	87.70M	15.17G
KDLNet (ResNet-18)	11.7M	1.82G
KDLNet (MobileNetV3)	4.2M	0.23G
KDLNet (EfficientNet)	5.3M	0.42G

图7　所提出方法(KDLNet)与基线方法(Fine-tuned ResNet-18)在NWPU-RESISC45数据集上热图结果对比。第一行为数据集中的原始图像,第二行和第三行分别为基线方法和KDLNet的热图。

Fig. 7　Visual comparison of heatmaps between the proposed method (KDLNet) and the baseline method (Fine-tuned ResNet-18) for NWPU-RESISC45 dataset. Original scenes from the NWPU-RESISC45 dataset are given in the first row. The second and third rows correspond to the heatmaps of the baseline and the proposed KDLNet, respectively.



1) **高精度:**KDLNet (EfficientNet)在AID数据集的50%训练样本和NWPU-RESISC45数据集的20%训练样本下,分别取得了97.32%和95.12%的分类准确率,超过了过去主流的深度学习基线方法和当前先进的方法,具有较高的分类精度。

2) **轻量化:**相较于过去的方法,KDLNet在参数量和FLOPs上均有显著降低。以KDLNet (EfficientNet)为例,其参数量和运算量(FLOPs)仅为Swin-Base的6.04%和2.77%,并且与Swin-Base具有相近的分类准确率。

3) **超参数易调节:**在KDLNet中,仅有温度系数T是需要调节的超参数。实验中,我们在多个数据集和不同的学生模型上均采用了相同的参数设置,取得了良好的分类结果;并且KDLNet可以在一个较宽的温度系数T变化范围内保持相近的结果,表明KDLNet对超参数变化不敏感。

另外,可视化结果也表明,尽管经过蒸馏的轻量级网络感知范围有所扩大,但它们的表达能力和

泛化能力仍然有限。这也意味着,它们能识别并聚焦于对分类决策最重要的区域,却仍不能足够精确地捕捉到目标对象的细微特征和边缘信息,其感知精度仍显不足,这也导致其可能不利于直接应用于那些需要高精度定位和细粒度识别的下游任务(如目标检测、语义分割等)。

4 结论

本文提出了一种基于知识蒸馏的轻量化遥感图像场景分类方法(即KDLNet)。所提议方法能够融合Transformer模型和CNN模型的优点,不仅能充分提取遥感图像的局部信息,还能充分挖掘遥感图像中的长程信息,具有高精度和轻量化的特点。具体地,本文以Swin Transformer和三种常见的轻量化CNN模型分别作为教师模型和学生模型,前者能够提取丰富的长程上下文信息,后者能够充分学习局部特征,并且具有参数量小、运算量低的特点;然后,通过知识蒸馏的方式将教师模型中的潜在知识转移到学生模型,提升后者对遥感图像的全面理解。更进一步,本文提出了一种新颖的知识蒸馏损失函数,舍弃了传统知识蒸馏过程中的分类损失,引入了类内蒸馏损失,使得学生模型在蒸馏过程中进一步学习遥感图像场景类别间和类别内的关系。在两个大规模公开数据集上的实验结果表明,所提出方法能够更好地学习到遥感图像的潜在特征,大幅提升基线方法的分类精度,甚至超过教师模型;更进一步,所提出的方法相较于目前的SOTA遥感图像场景分类方法也有显著的精度提升,与此同时,具有显著减小的参数量和运算量。所提议方法的这些特点对于实际应用具有重要的意义。

在未来工作中,我们计划将所提出方法与特征蒸馏相结合,同时向轻量级网络传递教师模型的中间层特征和输出特征,以使得学生模型能够更加精确地捕捉到特征细节,并提升泛化能力。

References

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

[2] Lv Z, Shi W, Zhang X, *et al*. Landslide inventory mapping from bitemporal high-resolution remote sensing images using change detection and multiscale segmentation [J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2018, **11**(5): 1520-1532.

[3] Longbotham N, Chaapel C, Bleiler L, *et al*. Very high resolution multiangle urban classification analysis [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2012, **50**(5): 1155-1170.

[4] Lowe D. Distinctive image features from scale-invariant keypoints [J]. *International Journal of Computer Vision*, 2004, **60**(2): 91-110.

[5] Dalal N, Triggs B. Histograms of oriented gradients for human detection [C]. IEEE Conference on Computer Vision & Pattern Recognition, 2005.

[6] Yang Y, Newsam S. Bag-of-visual-words and spatial extensions for land-use classification [C]. Sigspatial International Conference on Advances in Geographic Information Systems, 2010.

[7] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [C]. International Conference on Learning Representations, 2014.

[8] He K, Zhang X, Ren S, *et al*. Deep residual learning for image recognition [C]. IEEE Conference on Computer Vision & Pattern Recognition, 2016.

[9] Bai L, Liu Q, Li C, *et al*. Remote Sensing Image Scene Classification Using Multiscale Feature Fusion Covariance Network With Octave Convolution [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, **60**: 1-14.

[10] Tand X, Li M, Ma J, *et al*. EMTCAL: Efficient Multiscale Transformer and Cross-Level Attention Learning for Remote Sensing Scene Classification [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, **60**: 1-15.

[11] Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale [C]. International Conference on Learning Representations, 2021.

[12] Liu Z, Lin Y, Cao Y, *et al*. Swin transformer: Hierarchical vision transformer using shifted windows [C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[13] Lv P, Wu W, Zhong Y, *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**: 1-12.

[14] Xu K, Deng P, 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**: 1-15.

[15] Chen G, Zhang X, Tan X, *et al*. Training small networks for scene classification of remote sensing images via knowledge distillation [J]. *Remote Sensing*, 2018, **10**(5): 719.

[16] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network [J]. *arXiv preprint arXiv*: 1503.02531, 2015.

[17] Howard A, Sandler M, Chu G, *et al*. Searching for mobilenetv3 [C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.

[18] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks [C]. International Conference on Machine Learning, 2019.

[19] Xia G, Hu J, 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.

[20] Cheng G, Han J, Lu X. Remote sensing image scene clas-

sification: benchmark and state of the art [J]. *Proceedings of the IEEE*, 2017, **105**(10): 1865-1883.
[21] Wang W, Chen Y, Ghamisi P. Transferring CNN with adaptive learning for remote sensing scene classification [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, **60**: 1-18.
[22] Wang J, Li W, Zhang M, *et al*. Remote Sensing Scene Classification via Multi-Stage Self-Guided Separation Network [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2023, **61**: 1-12.
[23] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [J]. *Advances in Neural Information Processing Systems*, 2012, 25.
[24] Zhou B, Khosla A, Lapedriza A, *et al*. Learning deep features for discriminative localization [C]. IEEE Conference on Computer Vision & Pattern Recognition, 2016.