

基于分数阶混合几何特征的知识蒸馏算法

刘雅文 周尚波 陈懿佳

(重庆大学计算机学院 重庆 404100)

摘要 目前高光谱图像存在由中心点光谱变异引起的误分类问题。用于解决该问题的高光谱图像分类模型计算成本高、结构复杂,难以部署;而轻量化网络虽计算效率高,却未能有效处理光谱变异,影响分类精度。针对上述问题,本文创新性地提出了一种基于分数阶混合几何特征的知识蒸馏算法(Knowledge Distillation with Fractional-Ordered Mixed Geometric Features, KDFMGF)。本文提出了一种批处理分数阶微分几何特征计算模块,创新性地使用分数阶高斯曲率和分数阶平均曲率定位变异像素点,提出了分数阶微分几何特征蒸馏损失计算模块,让学生模型精准学习教师模型中关于变异像素点的分类知识,设计了一种分数阶几何特征直接损失计算模块,让学生模型从标签中学习关于变异像素点的分类知识,同时首次提出针对不同教师模型的模式:累积模式和消除模式;累积模式用于已解决误分类的教师模型,消除模式适用于未解决误分类问题的教师模型。KDFMGF显著降低了模型计算成本。KDFMGF不同于其他蒸馏算法,经KDFMGF蒸馏后的学生模型的性能能够超越教师模型。实验结果表明,在累积模式和消除模式下,KDFMGF算法显著优于现行的先进知识蒸馏算法,经过KDFMGF蒸馏的学生模型,在计算开销大幅下降的同时,分类性能超越教师模型,特别是在累积模式下,其分类性能超越十种先进的高光谱图像分类算法。

关键词 高光谱图像分类;知识蒸馏;分数阶导数;微分几何

中图分类号 TP391 DOI号 10.11897/SP.J.1016.2026.00535

Knowledge Distillation Based on Fractional-Ordered Mixed Geometric Features

LIU Ya-Wen ZHOU Shang-Bo CHEN Yi-Jia

(Department of Computer Science, Chongqing University, Chongqing 404100)

Abstract At present, there is a misclassification problem in hyperspectral images caused by spectral variation at the center point. The hyperspectral image classification model that solves this problem has high computational cost, complex structure, and is difficult to deploy. Although lightweight networks have high computational efficiency, they fail to effectively handle spectral variations, which affects classification accuracy. This paper proposes an innovative knowledge distillation algorithm based on fractional ordered mixed geometric features (KDFMGF) to address the aforementioned issues. This paper proposes a batch processing fractional order differential geometric feature calculation module, which innovatively uses fractional order Gaussian curvature and fractional order average curvature to locate variant pixels. A fractional order differential geometric feature distillation loss calculation module is proposed to enable student models to accurately learn classification knowledge about variant pixels in teacher models. A direct loss calculation module for fractional order geometric features is designed to enable student models to learn classification knowledge about variant pixels from labels. At the same time, two modes for different teacher models are proposed for the first time: cumulative mode and elimination mode; The cumulative mode is used for teacher models that have resolved misclassifications, while

the elimination mode is suitable for teacher models that have unresolved misclassification problems. KDFMGF significantly reduces the computational cost of the model. KDFMGF is different from other distillation algorithms, as the classification performance of the student model obtained through KDFMGF distillation can surpass that of the teacher model. The experimental results show that in both cumulative and elimination modes, the KDFMGF algorithm outperforms current advanced knowledge distillation algorithms. The student model distilled by KDFMGF not only significantly reduces computational costs, but also outperforms the teacher model in classification performance. Especially in cumulative mode, its classification performance surpasses ten advanced hyperspectral image classification algorithms.

Key words hyperspectral image classification; knowledge distillation; fractional-order derivative; differential geometry

1 引 言

与只捕捉 RGB 三通道的传统视觉系统不同, 高光谱图像包含更宽广的光谱区域, 覆盖了可见光、近红外、短波红外等多个电磁波谱范围^[1]。地面目标包含植被、土壤、水体等多种类别, 它们在不同波长下展现出特有的吸收和反射特征, 因此高光谱图像能够精细识别不同地面物体的差异^[2]。研究人员通过研究高光谱图像, 实现了对多种地面目标的像素级分类。因此, 高光谱图像在环境监测、资源勘探、农业评估、城市规划^[3]等领域中具有重大意义。目前, 基于 CNN 的高光谱图像分类模型存在着由中心点光谱变异引起的误分类问题, 但已经有研究针对这一问题进行优化, 例如, Chen 等人^[4]提出的一种基于分数阶混合几何特征的注意力网络 (Attention Network with Fractional-Ordered Mixed Geometric Features, ANFMGF)。但是这些基于 CNN 的模型在解决误分类问题、提高模型分类性能的同时, 也显著增加了模型参数量^[5], 提高了计算复杂度, 限制了模型在资源受限设备上的部署^[6]。因此, 研发轻量化、高效模型成为亟待解决的关键问题。

目前, 轻量化网络设计、剪枝和知识蒸馏是高光谱图像领域常用的轻量化方法。对于轻量化网络设计, 重点在于构建高效的网络结构和模块。例如, GU 等人^[7]为解决高光谱图像分类中的复杂性问题, 设计了多层级光谱-空间变换网络, 采用紧凑的网络结构, 从而减少参数数量和计算复杂度。如 Chen 等人^[8]提出了轻量级光谱-空间特征提取与融合网络, 该模型使用多种空间特征提取与融合方法, 不仅能有效去除冗余信息, 还显著降低模型复杂度。虽然轻量化网络设计能够提供更快、更节能的模型, 但它需要重新训练模型, 还需要权衡模型的通用性。

网络剪枝方法是一种旨在通过删除神经网络中不重要或者冗余的连接、神经元或者通道, 以减少

模型大小和计算复杂度, 从而加速推理过程并降低模型对存储需求的技术。如 Bai 等人^[9]提出了一种针对多尺度多分枝网络的剪枝方法, 动态调整剪枝比例, 从而减少模型的参数数量和计算复杂度。Lei 等人^[10]为降低模型复杂性和开销, 提出了一种网络协同剪枝方法, 使得模型在小样本条件下仍能在鲁棒性、分类精度、泛化能力等方面表现良好。但剪枝可能会导致模型的精度降低, 还可能对模型的鲁棒性和泛化性产生负面影响^[11]。

知识蒸馏作为一种模型压缩方法, 基于一个具有较高复杂度但分类性能优异的教师模型, 将知识传递给一个小型简单的学生模型^[12], 从而降低模型的参数量和计算成本, 同时保留较高的分类性能^[13]。在高光谱图像分类领域, 知识蒸馏主要有两方面的应用: 首先是利用“教师-学生网络”思想来实现模型轻量化, 降低计算资源。例如, Shi 等人^[14]提出一种尺度蒸馏网络, 单一尺度学生网络从多尺度教师模型中学习知识及其真实标签, 从而提高分类准确性。Xie 等人^[15]提出了一种结合解耦知识蒸馏和空间特征模糊的高光谱图像分类方法, 基于强大的教师网络, 运用有效的知识迁移策略, 提升了学生网络的分类性能, 有效降低网络推理时间。其次, 知识蒸馏通过知识迁移和自蒸馏, 从预训练教师模型中提取有用信息, 从而提高学生模型的分类准确性。Yang 等人^[16]提出一种融合孪生网络与知识蒸馏思想的新高光谱分类网络结构, 通过损失函数来融合教师和学生分支, 提升了模型的分类性能。Wu 等人^[17]提出了一种中心光谱自蒸馏框架, 提取出高光谱图像分块中心的纯光谱信息, 并将其用于自蒸馏学习, 在像素级别上实现了更精确的分类。

然而, 目前高光谱图像分类中的知识蒸馏方法虽然在轻量化部署等方面取得进展, 却没有考虑因光谱变异导致的误分类问题; 同时, 解决了误分类问题的模型, 却忽略了对计算成本和模型复杂度的

控制,导致模型部署困难。因此,亟待提出一种知识蒸馏算法,不仅能够解决模型的计算成本问题,还能优化由光谱变异引起的误分类问题。故,本文提出了一种基于分数阶混合几何特征的知识蒸馏算法(Knowledge Distillation with Fractional-Ordered Mixed Geometric Features, KDFMGF)及其两种模式:累积模式和消除模式。累积模式对优化了误分类问题的模型算法进行蒸馏,在该模式中,本文采用已解决误分类问题的 ANFMGF 作为教师模型,让学生模型充分学习教师模型中关于光谱变异像素点的正确分类知识,从而提高学生模型的分类性能。消除模式则针对未解决误分类问题的教师模型,本文选用面向中心向量的自相似性网络(Central Vector Oriented Self-Similarity Network, CVSSN)作为教师模型进行蒸馏,让学生模型更关注标签中关于光谱变异像素点的正确知识,规避教师模型中的错误分类知识。两种不同的模式适用于不同的教师模型架构,既降低了模型的计算成本,又缓解了误分类问题,提高了学生模型的分类效果。具体来说,本文的贡献如下:

(1)创新提出了一个批处理分数阶微分几何特征计算模块(Batch Fractional-Order Differential Geometry Feature Calculation, BFDGFC),首次将分数阶高斯曲率与分数阶平均曲率用于定位光谱变异像素点位置,识别导致误分类问题的关键像素点。

(2)创造了分数阶微分几何特征蒸馏损失计算模块(Soft Loss with Fractional-Order Differential Geometry Features, SLFDGF),能够分辨出教师模型中与光谱变异像素点分类相关的知识,为后续的知识传递提供针对性的指导。

(3)创造了分数阶微分几何特征直接损失计算模块(Direct Loss with Fractional-Order Differential Geometry Features, DLFDGF),在标签学习过程中,定位与光谱变异像素点分类相关的知识,强化学生模型对这些知识的关注,以便更精确地指导模型。

(4)提出了融合模块的累积模式,用于对优化了光谱变异所导致的误分类问题的算法进行知识蒸馏,继承其对光谱变异像素点的正确分类知识,同时弥补教师模型中的不足。

(5)提出了融合模块的消除模式,针对未考虑误分类问题的算法进行蒸馏,抑制教师模型中错误知识的干扰,转而专注于学习标签中正确的知识。

为了验证 KDFMGF 算法的累积模式和消除模

式的有效性,本文在五个高光谱图像数据集上进行了对比实验,实验结果表明,无论在累积模式还是消除模式下,KDFMGF的蒸馏效果均优于对比的8种先进知识蒸馏方法,且其蒸馏后的学生模型尽管参数量和计算复杂度大幅度下降,其分类性能明显优于教师模型。此外,在KDFMGF的累积模式下,以ANFMGF作为教师模型,蒸馏得到的学生模型分类性能显著优于10种先进高光谱图像分类方法,达到最优性能。

2 相关工作

2.1 知识蒸馏

知识蒸馏是一种简单、有效的新兴模型压缩技术,被广泛应用于工业界,它基于“教师-学生网络框架”,将结构复杂、参数量巨大的模型作为教师模型,将教师模型的知识迁移到简单的学生模型^[18],提升其分类预测能力。该技术的关键在于如何设计、提取并迁移知识,根据知识的来源和传递形式的不同,通常可分为四类:基于输出层特征知识、基于中间层特征知识、基于结构化知识以及基于图表示的知识。

输出特征知识通常是指在模型最后输出的 logits 概率分布中的软化目标信息。基于输出特征知识蒸馏方法的最大优势在于,该方法不依赖各种网络模型的内部结构以及特征表示方法,而是直接利用教师模型的输出来指导学生模型的学习过程。因此,它能够简化蒸馏过程,同时便于与其他蒸馏技术相结合。

中间层知识是指深度神经网络中间层提取出的高维特征。虽然中间层知识蒸馏相比于基于输出特征的蒸馏,传输知识的表征能力和传输的信息量更大^[19],但由于教师-学生模型的网络结构不同,其中间知识层结构难以对齐,导致对学生模型的训练过程艰难^[20-21],同时由于中间层知识蒸馏涉及多个领域任务,缺乏对多种中间层知识的整合^[22],不同层次的特征可能相互干扰^[23],导致训练困难。

结构化知识包括类间关系以及类内关系。虽然结构化知识能够挖掘教师模型中样本特征之间更丰富的关联性^[24-25],但是,结构化知识蒸馏的计算开销较大,且优化成本较高。

图表示知识可视为结构化知识的一种扩展与特殊形式,它能有效地满足对非结构化数据的表示与学习需求。然而,图表示知识蒸馏的特征构建过程

较为复杂^[26],计算开销较大,导致泛化性能下降^[27]。

为了权衡语义意义和蒸馏计算开销两方面需求,本文最终选择输出特征知识蒸馏,具体来说,相比图表示知识蒸馏和结构知识蒸馏,它计算开销更低;相比中间层知识蒸馏,它拥有更高的语义意义,在模型性能、知识理解、泛化能力等方面发挥了重要作用。

2.2 知识蒸馏(KD)算法

知识蒸馏(Knowledge Distillation, KD)是Hinton等人提出的一种方法,通过让学生模型模仿教师模型输出的类间相似性(暗知识),从而提升学生模型的泛化能力。Hinton等人在Soft max函数中引入一个新颖的超参数 T ,表示蒸馏温度,从而产生一个较软的概率分布,即软目标,它能够提供更多的类间相似度信息,其计算过程如式(1)所示。

$$p(z_i, T) = \frac{\exp(z_i / T)}{\sum_j \exp(z_j / T)} \quad (1)$$

其中, z_i 为模型经过Soft max函数之前的输出,即第 i 类的概率, T 代表温度系数,控制输出概率的软化程度。当 $T=1$ 时,公式退化为标准的Soft max函数;当 T 取较大值时,输出的概率分布会变得平滑,且随着 T 取值变大,Soft max函数的输出分布变得更加平缓,学生模型会更加关注到负标签。例如,在手写字符识别任务中,数字2和数字3在外形上相似,如图1(a)所示,若使用硬目标表示,它们会被强制分类为某个类别,如图1(b)中“非0即1”的表示。而引入温度系数 T 后,模型输出的概率分布会变得更加平缓,如图1(c)和图1(d)所示。

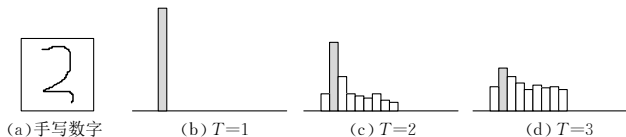


图1 温度系数影响示意图

KD算法的训练流程如图2所示。最终损失由这蒸馏损失和直接损失共同组成,如式(2)所示。

$$L = \alpha L_{hard} + \beta L_{KL} \quad (2)$$

其中, α 和 β 是两个超参数,分别用于控制直接损失

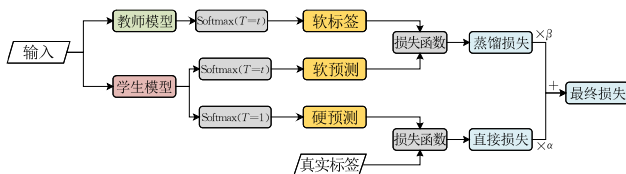


图2 KD算法的训练流程图

L_{hard} 和蒸馏损失 L_{KL} 在总损失中的权重比例。

在模型训练过程中,采用传统交叉熵损失函数来计算直接损失,旨在引导学生模型准确拟合真实标签。采用Kullback-Leibler(KL)散度来计算蒸馏损失,旨在量化教师模型的输出和学生模型的输出概率分布的差异度。

3 基于分数阶混合几何特征的知识蒸馏算法

3.1 整体架构

目前高光谱图像分类中的知识蒸馏方法并未考虑同步解决因光谱变异产生的误分类问题。针对该问题本文探索了在蒸馏过程中同时优化误分类问题的方法,提出了一种基于分数阶混合几何特征的知识蒸馏算法,同时引入了两种模式:累积模式和消除模式。累积模式是针对已优化光谱变异问题的算法。消除模式是针对未考虑光谱变异问题的算法。其中图3展示了基于分数阶混合几何特征的知识蒸馏(KDFMGF)算法的架构图。

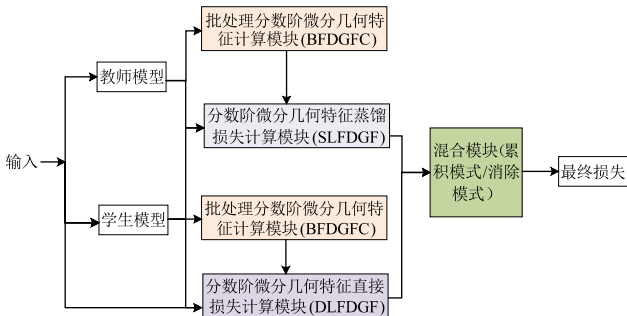


图3 基于分数阶混合几何特征的知识蒸馏算法的架构图

针对优化了误分类问题的教师模型,其大部分相关分类知识是正确的,采用累积模式,通过知识转移,将正确的分类信息有效传递到学生模型。针对未优化误分类问题的教师模型,相关知识大多是错误的,采用消除模式,移除教师模型中的错误知识,防止学生模型被误导。

我们提出了四个模块:批处理分数阶微分几何特征计算(BFDGFC)模块,分数阶微分几何特征蒸馏损失计算(SLFDGF)模块,分数阶微分几何特征直接损失计算(DLFDFG)模块以及混合模块(Integrated Module, IM)模块。

具体来说,BFDGFC模块利用分数阶微分几何特征来定位光谱变异像素所在的区域,相比于传统的整数阶微分方法,分数阶微分能够更细致地捕捉图像中的细节信息,从而提高对光谱变异像素点的

检测精度。

采用分数阶微分几何特征蒸馏损失和分数阶微分几何特征直接损失来构建 SLFDGF 模块和 DLFDGF 模块。SLFDGF 模块通过软目标和软预测的方式，提取教师模型中与光谱变异像素分类相关的知识并传递给学生，不仅保留了教师模型的高级语义信息，还增强了学生对复杂样本的理解能力，从而提高分类性能；DLFDGF 模块采用硬预测和真实标签进行直接损失计算，旨在针对性弥补教师模型在误分类方向的不足，由于直接损失的计算基于真实标签，因此能够避免由于教师模型的失误而导致的错误知识的传播，从而提高分类性能。

最后，两种模式的混合模块：累积模式和消除模式，融合四种损失函数，这种多损失函数结合的方式能够充分利用不同损失函数的优点，实现互补和协同作用，两种模式的设计考虑了多种教师模型架构，能够灵活适应不同架构的教师模型。

KDFMGMF 算法及其两种模式能够在蒸馏过程中实现对模型进行优化，旨在降低计算开销的同时优化误分类问题，从而提升学生模型分类性能。在累积模式下，学生模型能够关注学习教师模型中针对光谱变异像素的正确分类知识，并且弥补教师模型中的不足；在消除模式下，抑制教师模型中的错

误分类知识的同时让学生模型聚焦于标签学习。

3.2 KDFMGMF 总体流程

F 是专为高光谱图像分类设计的知识蒸馏模型，高光谱图像分类模型通常将以像素为中心的 3D 块作为模型输入。图 4 展示了 KDFMGMF 的具体训练流程。

在累积模式下，选择已经解决了误分类问题的教师模型——ANFMGMF 作为例子，详细解析其蒸馏机制与训练策略。

在 ANFMGMF 模型中，需要对高光谱图像输入进行预处理，原始 3D 高光谱图像 $\mathbf{X} \in \mathbb{R}^{h \times w \times b}$ 被划分为一组重叠的小 3D 斑块 \mathbf{I} ， \mathbf{I} 中的每个斑块 $\mathbf{I}_{i,j} \in \mathbb{R}^{s \times s \times b}$ 是以像素 $x_{i,j}$ 为中心的相邻立方体，其中 $s \times s$ 代表 $\mathbf{I}_{i,j}$ 所覆盖空间窗口的面积。ANFMGMF 采用该斑块 $\mathbf{I}_{i,j}$ 作为模型的输入，对中心像素 $x_{i,j}$ 进行分类。

如图 4 所示，教师模型和学生模型通过 BFDGFC 模块精准定位光谱变异像素点，随后将信息分别送入 SLFDGF 模块和 DLFDGF 模块计算分数阶微分几何特征蒸馏损失和分数阶微分几何特征直接损失。最后使用混合模块融合分数阶微分几何特征蒸馏损失、原蒸馏损失、分数阶微分几何特征直接损失、原直接损失，得到最终损失。

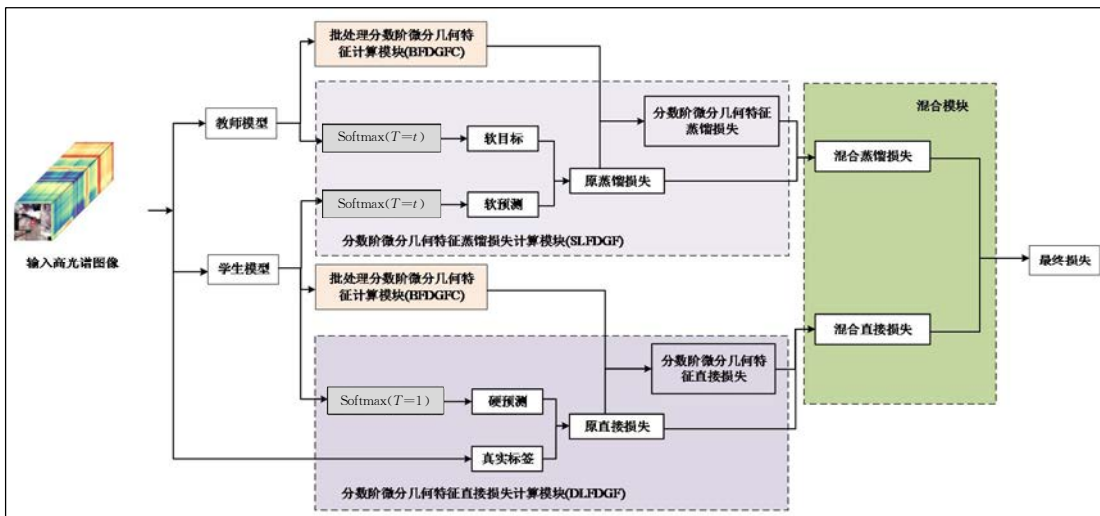


图 4 KDFMGMF 的训练流程

在第 3.3 节~第 3.6 节中，将详细介绍本文涉及的四个模块：BFDGFC 模块、DLFDGF 模块、SLFDGF 模块以及混合模块。

3.3 批处理分数阶微分几何特征计算(BFDGFC) 模块

为实现在知识蒸馏的同时优化由光谱变异产生的误分类问题，则需要精准定位产生光谱变异的像

素点，为此，本文设计 BFDGFC 模块。图 5 展示了 BFDGFC 模块的具体运行流程。

如图 5 所示，首先是对原始高光谱图像 $\mathbf{X} \in \mathbb{R}^{h \times w \times b}$ 进行处理，随机抽取 b 个像素点，并将它们作为中心像素点截取 b 个窗口大小为 $s \times s$ 的高光谱斑块，得到一个样本子集 $\mathbf{P} \in \mathbb{R}^{b \times s \times s \times c}$ ，它由 b 个斑块 $\mathbf{I}_{i,j} \in \mathbb{R}^{s \times s \times c}$ 组成， b 是批量大小，也即一次迭代

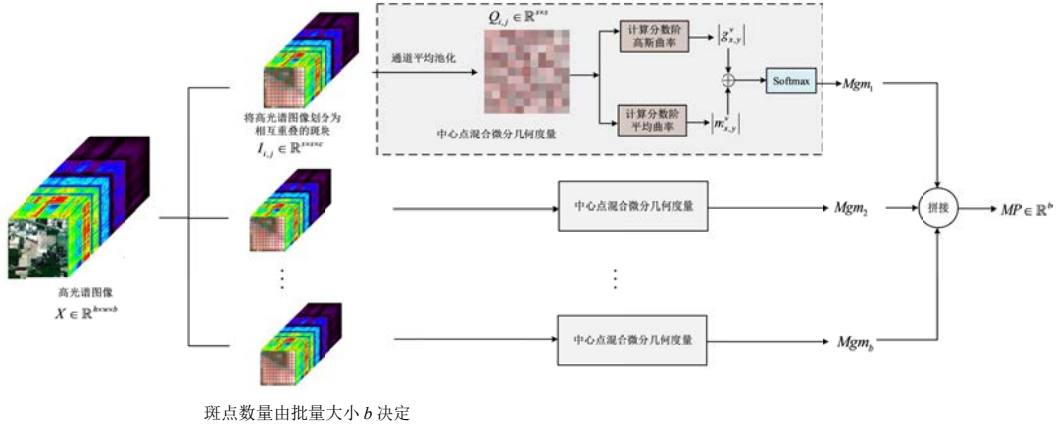


图5 BFDGFC 模块流程图

中所处理的斑块数量, c 是通道数。采用批量大小保证训练效率的同时提高模型鲁棒性。通过批量斑块来近似计算损失函数的梯度, 平衡训练效率与模型收敛的稳定性。在每次迭代中, 算法会计算出每个样本的混合微分几何度量。将所有度量拼接形成分数阶微分几何特征向量作为 BFDGFC 模块的输出。将 BFDGFC 模块的输出向量用于对样本子集中的样本原损失实施加权策略, 提高对光谱变异像素点的关注程度, 从而弥补教师模型的缺陷。其中, 中心点混合微分几何度量需要计算分数阶高斯曲率和分数阶平均曲率。

为计算以像素 $x_{i,j}$ 为中心的斑块 $I_{i,j}$ 的分数阶高斯曲率和分数阶平均曲率, 采用通道平均池化操作将多通道混合为单通道, 从而获得光谱空间块 $Q_{i,j} \in \mathbb{R}^{s \times s}$ 。本文将 $Q_{i,j}$ 视为一个曲面进行参数化, 得到 $\gamma(x,y) = \{x,y,f(x,y)\}$, 其中 $x=1,2,\dots,s$, $y=1,2,\dots,s$ 。在曲面参数公式中, x 和 y 共同表示像素点在斑块窗口中的位置, $f(x,y)$ 则表示位于坐标 (x,y) 的像素通道融合后的像素值。

分数阶高斯曲率 $g_{x,y}^v$ 和分数阶平均曲率 $m_{x,y}^v$ 可以由两个分数阶主曲率计算得到, 如式(3)和(4)。

$$g_{x,y}^v = \kappa_1^v \cdot \kappa_2^v \quad (3)$$

$$m_{x,y}^v = \frac{\kappa_1^v + \kappa_2^v}{2} \quad (4)$$

其中, κ_1^v 和 κ_2^v 分别表示曲面 $Q_{i,j}$ 在非脐点处沿两个相互垂直方向的分数阶主曲率。曲面 $Q_{i,j}$ 在 $P(x,y)$ 处的分数阶主曲率 κ^v 满足式(5)。

$$(EG - F^2)(\kappa^v)^2 - (EN - 2FM + GL)\kappa^v + (LN - M^2) = 0 \quad (5)$$

其中, E 、 F 和 G 表示分数阶第一基本量, L 、 M 和 N 表示分数阶第二基本量。 $g_{x,y}^v$ 和 $m_{x,y}^v$ 可以用分数阶第

一、二基本量来计算, 如式(6)和(7)所示。

$$g_{x,y}^v = \frac{LN - M^2}{EG - F^2} \quad (6)$$

$$m_{x,y}^v = \frac{EN - 2FM + GL}{2(EG - F^2)} \quad (7)$$

其中, 分数阶第一、二基本量可以用曲面 $Q_{i,j}$ 的分数阶一阶、二阶导数表示, 因此曲面 $Q_{i,j}$ 在点 $P(x,y)$ 处的 $g_{x,y}^v$ 和 $m_{x,y}^v$ 可以表示为式(8)和(9), 其中 f_{xx}^v 、 f_{yy}^v 、 f_{xy}^v 表示分数阶二阶混合偏导数, f_x^v 、 f_y^v 表示分数阶一阶偏导数。

当分数阶的阶数 v 取 1 时, 表示曲面 $Q_{i,j}$ 在 $P(x,y)$ 处的分数阶高斯曲率 $g_{x,y}^v$ 和分数阶平均曲率 $m_{x,y}^v$ 退化成正整数阶高斯曲率 $g_{x,y}$ 和整数阶平均曲率 $m_{x,y}$ 。因此曲面 $Q_{i,j}$ 的 $g_{x,y}$ 和 $m_{x,y}$ 可以用整数阶一阶导数和整数阶二阶导数表示, 如式(10)和(11)。

$$g_{x,y}^v = \frac{f_{xx}^v f_{yy}^v - f_{xy}^v f_{xy}^v}{(1 + f_x^v f_x^v + f_y^v f_y^v)^2} \quad (8)$$

$$m_{x,y}^v = \frac{f_{xx}^v (1 + f_y^v f_y^v) - 2f_x^v f_y^v f_{xy}^v + f_{yy}^v (1 + f_x^v f_x^v)}{2(1 + f_x^v f_x^v + f_y^v f_y^v)^{\frac{3}{2}}} \quad (9)$$

$$g_{x,y} = \frac{f_{xx} f_{yy} - f_{xy}^2}{(1 + f_x^2 + f_y^2)^2} \quad (10)$$

$$m_{x,y} = \frac{f_{xx} (1 + f_y^2) - 2f_x f_y f_{xy} + f_{yy} (1 + f_x^2)}{2(1 + f_x^2 + f_y^2)^{\frac{3}{2}}} \quad (11)$$

在本文中, 整数阶一阶导数的计算采用中心差分方法, 如式(12)。

$$f'(t) = \lim_{h \rightarrow 0} \frac{f(t+h) - f(t-h)}{2h} \quad (12)$$

当分数阶阶数 v 取 $0 < v < 1$, $g_{x,y}^v$ 和 $m_{x,y}^v$ 的公式如式(8)和(9)。

$$D_{c_1}^\nu f(t) = \lim_{h \rightarrow 0} \frac{\Gamma(\nu+1)}{h^\nu} \sum_{k \in \mathbb{Z}} \frac{(-1)^k f(t-kh)}{\Gamma\left(\frac{\nu+2}{2}-k\right)\Gamma\left(\frac{\nu+2}{2}+k\right)} \quad (13)$$

$$D_{c_2}^\nu f(t) = \lim_{h \rightarrow 0} \frac{\Gamma(\nu+1)}{h^\nu} \sum_{k \in \mathbb{Z}} \frac{(-1)^k f\left(t-kh+\frac{h}{2}\right)}{\Gamma\left(\frac{\nu+3}{2}-k\right)\Gamma\left(\frac{\nu+1}{2}+k\right)} \quad (14)$$

经分析,常用的由后向差分推导而来的G-L分数阶微分定义方法计算曲率会导致对光谱变异像素点定位出现偏差。因此,本文采用由中心差分得出的G-L分数阶中心导数^[28]。G-L分数阶中心导数为1型分数阶中心导数和2型分数阶中心导数。如式(13)和(14)。其中 $h \in \mathbb{R}^+$, Γ 表示Gamma函数。

当 10ν 是偶数时,分数阶导数符合1型分数阶中心导数, 10ν 为分数阶阶数 ν 的十倍。如果函数 $f(t)$ 的连续周期 t 按步长 $h=1$ 等分,则 $f^\nu(\nu)$ 的计算如式(15)所示。

$$f^\nu(\nu) = D_{c_1}^\nu f(t) \approx \Gamma(\nu+1) \sum_{k \in \mathbb{Z}} \frac{(-1)^k f(t-k)}{\Gamma\left(\frac{\nu+2}{2}-k\right)\Gamma\left(\frac{\nu+2}{2}+k\right)} \quad (15)$$

为便于实现,本文中仅取式(15)的中间三项进行近似,近似结果如式(16)所示。

$$f^\nu(\nu) \approx \Gamma(\nu+1) \left[\frac{-1}{\Gamma\left(\frac{\nu}{2}+2\right)\Gamma\left(\frac{\nu}{2}\right)} f(t+1) + \frac{1}{\Gamma\left(\frac{\nu}{2}+1\right)\Gamma\left(\frac{\nu}{2}+1\right)} f(t) + \frac{-1}{\Gamma\left(\frac{\nu}{2}\right)\Gamma\left(\frac{\nu}{2}+2\right)} f(t-1) \right] \quad (16)$$

当 10ν 是奇数时,分数阶导数符合2型分数阶中心导数, 10ν 为分数阶阶数 ν 的十倍。此时为了使 $t-kh+h/2$ 取整数,函数 $f(t)$ 的连续周期 t 按步长 $h=2$ 等分,则 $f^\nu(\nu)$ 的计算如式(17)所示。

$$f^\nu(\nu) = D_{c_2}^\nu f(t) \approx \frac{\Gamma(\nu+1)}{2^\nu} \sum_{k \in \mathbb{Z}} \frac{(-1)^k f(t-2k+1)}{\Gamma\left(\frac{\nu+3}{2}-k\right)\Gamma\left(\frac{\nu+1}{2}+k\right)} \quad (17)$$

为方便实现,本文取式(17)的中间两项进行近似计算,结果如式(18)所示。

$$f^\nu(\nu) \approx \frac{\Gamma(\nu+1)}{2^\nu} \left[\frac{1}{\Gamma\left(\frac{\nu+3}{2}\right)\Gamma\left(\frac{\nu+1}{2}\right)} f(t+1) + \frac{-1}{\Gamma\left(\frac{\nu+1}{2}\right)\Gamma\left(\frac{\nu+3}{2}\right)} f(t-1) \right] \quad (18)$$

计算出曲面 $Q_{i,j}$ 在点 $P(x,y)$ 处的分数阶高斯曲率 $g_{x,y}^\nu$ 和分数阶平均曲率 $m_{x,y}^\nu$ 后,将 $g_{x,y}^\nu$ 和 $m_{x,y}^\nu$ 进行融合,并使用Soft max函数进行归一化操作得到融合微分几何度量 $Mgm_{x,y}$,如式(19)所示。

$$Mgm_{x,y} = \text{Soft max}(|g_{x,y}^\nu| + |m_{x,y}^\nu|) \quad (19)$$

通过分数阶高斯曲率和分数阶主曲率能够区分平坦区域、边缘和噪点。 $|g_{x,y}^\nu|$ 和 $|m_{x,y}^\nu|$ 大小随图像区域变化,在平坦区域往往较小,在边缘和噪点处往往较大。

假设 κ_1^ν 是 κ_1^ν 、 κ_2^ν 中较大的分数阶主曲率。主曲率是点 $P(x,y)$ 在曲面 $Q_{i,j}$ 中两个正交方向上的最大和最小法曲率,反映了该点在曲面上沿不同方向的弯曲程度,根据式(3)和(4),在平坦区域,中心点像素和周围像素值差别小,则 κ_1^ν 和 κ_2^ν 都较小,因此 $|g_{x,y}^\nu|$ 和 $|m_{x,y}^\nu|$ 的值都较小,甚至可能为0;在边缘区域,垂直边缘方向像素值差距较大,平行边缘方向像素值差距较小,则 κ_1^ν 较大, κ_2^ν 较小,因此 $|g_{x,y}^\nu|$ 和 $|m_{x,y}^\nu|$ 的值比在平坦区域中的大;在噪点区域,中心点作为异常点与周围像素值相差均较大,则 κ_1^ν 和 κ_2^ν 都较大,因此 $|g_{x,y}^\nu|$ 和 $|m_{x,y}^\nu|$ 的值更大。

因此在平坦区域中, $Mgm_{x,y}$ 很小,在边缘处 $Mgm_{x,y}$ 较大,在孤立噪点区域, $Mgm_{x,y}$ 最大。故,曲面 $Q_{i,j}$ 中心点 $P(i,j)$ 的 $Mgm_{i,j}$ 代表斑块 $I_{i,j}$ 中心坐标像素点与周围坐标像素点之间的差异程度,值越大差异越显著,由此可定位产生光谱变异的像素点。

BFDGFC模块输出分数阶微分几何特征向量 $MP \in \mathbb{R}^b$,它由 b 个斑块 $I_{i,j}$ 得到的混合微分几何度量 $Mgm_{i,j}$ 拼接而成。

3.4 分数阶微分几何特征直接损失计算(DLFDGF)模块

原直接损失的计算采用交叉熵损失函数,每个斑块 $I_{i,j}$ 的原直接损失函数如式(20)所示。

$$\mathbf{lm}_{hard} = -\sum_{k=1}^K \ln(y_m = k) \log(\hat{y}_m(i_k)) \quad (20)$$

其中, y_m 是单次迭代中第 m 个高光谱斑块 \mathbf{I}_m 对应的标签, $\ln(y_m = k)$ 表示如果 y_m 的类别为 k , 则 $\ln(y_m = k)$ 的值为 1, 否则为 0; $\hat{y}_m(i_k)$ 表示经过 Soft max 函数后得到的 \mathbf{I}_m 属于 k 类的概率。

直接损失矩阵 $\mathbf{LM}_{hard} \in \mathbb{R}^b$ 由 b 个斑块 $\mathbf{I}_{i,j}$ 得到的直接损失 \mathbf{lm}_{hard} 拼接得到。将样本子集的直接损失矩阵 \mathbf{LM}_{hard} 与第 3.3 节中 BFDGFC 模块的输出即分数阶微分几何特征向量 $\mathbf{MP} \in \mathbb{R}^b$ 相乘, 得到分数阶微分几何特征直接损失矩阵 $\mathbf{LM}_{GM_hard} \in \mathbb{R}^b$, 如式(21)所示。

$$\mathbf{LM}_{GM_hard} = \mathbf{MP} \otimes \mathbf{LM}_{hard} \quad (21)$$

其中, \otimes 表示逐元素乘法。由于中心像素点产生光谱变异会导致分数阶微分几何算子的值偏大, 因此在 \mathbf{MP} 中, 光谱变异中心像素点所处斑块的混合微分几何度量 $\mathbf{Mgm}_{i,j}$ 偏高, 导致 \mathbf{LM}_{GM_hard} 在产生光谱变异像素点所在的斑块处的损失同样偏大, 这样能在标签学习阶段对光谱变异像素点给予更高的关注权重, 从而弥补教师模型对光谱变异像素点学习不足的缺陷。

随后, 计算 \mathbf{LM}_{GM_hard} 的平均值得到分数阶微分几何特征直接损失 \mathbf{L}_{GM_hard} , 计算过程如式(22)所示。

$$\mathbf{L}_{GM_hard} = \frac{\text{SUM}(\mathbf{LM}_{GM_hard})}{b} \quad (22)$$

样本子集的原直接损失 \mathbf{L}_{hard} 计算过程如式(23)所示。

$$\mathbf{L}_{hard} = \frac{\text{SUM}(\mathbf{LM}_{hard})}{b} \quad (23)$$

3.5 分数阶微分几何特征蒸馏损失计算(SLFDGF)模块

本文采用 KL 散度方法来计算原蒸馏损失, 衡量教师模型与学生模型输出结果概率分布的差异。每个斑块 $\mathbf{I}_{i,j}$ 的直接蒸馏损失 \mathbf{lm}_{KL} 计算过程如式(24)所示。

$$\mathbf{lm}_{KL} = \sum_{i=1}^N p_i^T \log\left(\frac{p_i^T}{q_i^T}\right) \quad (24)$$

其中, N 代表样本类别数量, p_i^T , q_i^T 分别表示教师和学生模型的第 i 个斑块的预测结果经过温度 T 蒸馏后的概率分布, 计算如式(25)和(26)所示。

$$p_i^T = \text{Soft max}\left(\frac{f_T(x_i)}{T}\right) \quad (25)$$

$$q_i^T = \text{Soft max}\left(\frac{f_S(x_i)}{T}\right) \quad (26)$$

其中, f_T 为教师模型输出, f_S 为学生模型输出。

蒸馏损失矩阵 $\mathbf{LM}_{KL} \in \mathbb{R}^b$ 由 b 个斑块 $\mathbf{I}_{i,j}$ 计算得到的 \mathbf{lm}_{KL} 拼接得到。将样本子集的原蒸馏损失矩阵 \mathbf{LM}_{KL} 与第 3.3 节中 BFDGFC 模块的分数阶微分几何特征向量 $\mathbf{MP} \in \mathbb{R}^b$ 相乘, 得到分数阶微分几何特征蒸馏损失矩阵 $\mathbf{LM}_{GM_KL} \in \mathbb{R}^b$, 如式(27)所示。

$$\mathbf{LM}_{GM_KL} = \mathbf{MP} \otimes \mathbf{LM}_{KL} \quad (27)$$

\mathbf{LM}_{GM_KL} 的变化与 \mathbf{LM}_{GM_hard} 相同, 光谱变异像素点所在的斑块处的蒸馏损失同样偏大, 从而可以精准定位教师模型对光谱变异像素点的知识点, 学生模型可以针对性地优化和处理。

随后, 对 \mathbf{LM}_{GM_KL} 取平均值计算得到分数阶微分几何特征蒸馏损失 \mathbf{L}_{GM_KL} , 如式(28)所示。

$$\mathbf{L}_{GM_KL} = \frac{\text{SUM}(\mathbf{LM}_{GM_KL})}{b} \quad (28)$$

样本子集的原蒸馏损失 \mathbf{L}_{KL} 的计算如式(29)。

$$\mathbf{L}_{KL} = \frac{\text{SUM}(\mathbf{LM}_{KL})}{b} \quad (29)$$

3.6 两种模式下的混合模块

经过分数阶微分几何特征直接损失计算(DLFDGF)模块以及分数阶微分几何特征蒸馏损失计算(SLFDGF)模块后, 模型获得了四个损失函数, 即原直接损失 \mathbf{L}_{hard} 、分数阶微分几何特征直接损失 \mathbf{L}_{GM_hard} 、原蒸馏损失 \mathbf{L}_{KL} 、分数阶微分几何特征蒸馏损失 \mathbf{L}_{GM_KL} 。本文设计了两种模式下的混合模块有效集成四个损失函数, 确保学生模型能精准地从教师模型中学习正确的分类知识, 同时, 根据教师模型是否优化光谱变异问题选择不同模式下的混合模块从而弥补教师模型中的不足, 有效提升学生模型的性能。

3.6.1 累积模式

针对已经优化了误分类问题的教师模型, 如 ANFMGF, 本文创新性设计了累积模式下的混合模块。对于此类教师模型, 学生模型需要针对性地学习教师模型中对光谱变异像素点的正确分类知识, 因此将原蒸馏损失与分数阶微分几何特征蒸馏损失融合得到混合蒸馏损失 \mathbf{L}_{MixKL} , 计算过程如式(30)所示。

$$\mathbf{L}_{MixKL} = \alpha_1 \mathbf{L}_{KL} + \beta_1 \mathbf{L}_{GM_KL} \quad (30)$$

α_1 和 β_1 均为超参数, 默认值为 0.5。

针对教师模型可能存在的学习错误和学习不足等问题, 本文提出一种标签学习的方式, 针对性学

习标签中光谱变异像素点的分类知识，将原直接损失和分数阶微分几何特征直接损失融合得到混合直接损失 $L_{Mixhard}$ ，如式(31)所示。

$$L_{Mixhard} = \alpha_2 L_{hard} + \beta_2 L_{GM_hard} \quad (31)$$

其中， β_2 均为超参数，默认值设为 0.5。

最后，融合混合直接损失 $L_{Mixhard}$ 和混合蒸馏损失 L_{MixKL} 得到最终损失 L ，融合过程如式(32)所示

$$L = \alpha_3 L_{Mixhard} + \beta_3 \times T^2 \times L_{MixKL} \quad (32)$$

其中， T 是蒸馏温度系数，系数 T^2 的作用在于即使蒸馏温度系数发生变化，也保持混合蒸馏损失和混合直接损失对最终损失的贡献相同。 α_3 和 β_3 均为超参数，默认值设为 0.5。

累积模式下，学生模型能够最大程度捕获教师模型中的正确分类知识，同时通过标签学习弥补教师模型的缺陷，以提升学生模型分类性能。

3.6.2 消除模式

在消除模式下，针对未解决光谱变异导致的误分类问题的教师模型，可以在蒸馏过程中解决误分类问题。未解决误分类问题的教师模型对于产生光谱变异像素点的分类知识大多错误，因此为了消减教师模型中此类知识对学生模型的负面影响，本文在将原蒸馏损失与分数阶微分几何特征蒸馏损失相融合得到融合蒸馏损失中，对超参数进行特殊设置，以达到目的，计算过程如式(33)所示

$$L_{MixKL} = \alpha_4 L_{KL} + \beta_4 L_{GM_KL} \quad (33)$$

其中， α_4 和 β_4 均为超参数， α_4 默认值设为 1.5， β_4 的默认值设为 -0.5。

减少教师模型中光谱变异像素点的分类知识对学生模型的负面影响，学生模型还是需要学习对光谱变异像素点的正确分类知识，因此将 L_{hard} 与 L_{GM_hard} 按照式(31)进行融合得到混合直接损失 $L_{Mixhard}$ 。随后，将混合直接损失和混合蒸馏损失按照式(32)进行融合得到消除模式下的最终损失 L 。

在消除模式下，抑制了暗知识对学生模型的影响，转而从标签上学习正确知识。

综上所述，不同模式下的混合模块针对不同的教师模型所设计，针对未解决误分类问题的教师模型，选择消除模式，减少教师模型中关于变异像素点错误知识对学生模型的影响，让学生模型从光谱变异像素点标签中学习正确的知识。对于解决了误分类问题的教师模型，由于其对光谱变异像素点的分类知识是大部分正确的，因此，可以通过知识迁移的方法，把正确的分类知识高效传递给学生模型。

4 实验结果以及分析

4.1 数据集

为了验证 KDFMGF 算法分别在累积模式和消除模式下的有效性，实验在五个高光谱图像分类数据集上进行验证：Indian Pines(IP)数据集、Pavia University(UP)数据集、Kennedy Space Center(KSC)数据集、University of Houston(UH)数据集和 WHU-Hi-LongKou(LK)数据集。

对于数据集的划分，本文考虑采用两种方式：随机划分和空间不相交划分。

随机划分是对于每个数据集，随机从每类总标记像素中选择一定比例的像素作为训练集和验证集，剩余像素作为测试集。本文随机训练集和验证集的比例为 10% 和 1%，剩余为测试集。在随机划分中选择 IP 数据集、UP 数据集、KSC 数据集和 UH 数据集。随机划分会忽略训练集和测试集之间的空间依赖性导致分类性能被高估，因此补充空间不相交划分方式，确保算法的有效性。

空间不相交划分，选择 UP、LK 和 UH 数据集。UP 数据集的空间不相交的训练和测试样本可在 GRSSDASE 网站上访问，而 LK 和 UH 数据集的训练和测试样本由官方提供。

4.2 实验设置

实验在搭载了 Intel Core i9-10920X CPU、64 GB RAM、Nvidia GeForce RTX 3090 GPU(24 GB 显存)的工作站上运行，使用 Ubuntu 20.04.4 LTS 64 位系统。在训练过程中，采用 Adam 优化器，批量大小 b 设置为 32，学习率和训练轮次分别设置为 0.001 和 100。此外，学生模型和教师模型对于输入都采用零填充策略，在高光谱图像边缘填充 $(s-1)/2$ 个零像素，其中设置为 9。

在 IP、KSC、UP、UH 数据集上分别将分数 ν 阶数设置为 0.8、0.2、0.9、0.5。

在累积模式下，损失函数中的 α_1 、 β_1 、 α_2 、 β_2 、 α_3 、 β_3 都设置为 0.5。在消除模式下， α_2 、 β_2 、 α_3 、 β_3 都设置为 0.5， α_4 、 β_4 设置为 1.5 和 -0.5。

实验采用三种常见的评价指标：总体准确率 (Overall Accuracy, OA)、平均准确率 (Average Accuracy, AA)、kappa 系数 (Kappa)。对于模型本身，使用模型参数量 (Parameters, Params) 和浮点数计算量 (Floating Point Operations, FLOPs) 反映模型复杂度。

4.3 累积模式下 KDFMGF 的实验结果

4.3.1 教师网络和学生网络的设置

如图 6 所示, 教师模型选择已经解决了误分类问题的 ANFMGF 模型。本文设计的学生网络, 称为特征提取网络(Feature Extraction Network, FEN), 它在 ANFMGF 模型的基础上去掉 FGFSVS 模块和 ED-FVSS 模块, 只剩下 SSIF 模块、CSS-Conv 模块、SIC-Conv 模块和分类层。

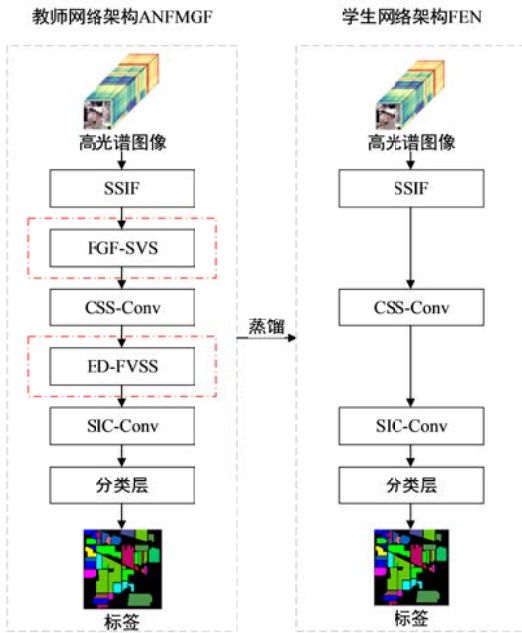


图 6 教师网络和学生网络架构对比图

4.3.2 随机选择样本的对比实验

累积模式下, 为验证 KDFMGF 模型的有效性,

在本节实验中采取随机划分的方法, 将 IP、KSC、UP、UH 数据集划分为训练集和测试集, 和现行的蒸馏算法进行对比, 对比的算法包括经典知识蒸馏 KD^[29]算法、基于注意力机制的 AT^[30]算法、基于结构化知识的 RKD^[31]算法、CCKD^[32]算法、ICKD^[33]算法、DIST^[34]算法、Sun 等人^[35]提出的基于输出特征知识算法, 在本文中用 LS 代指, 以及 Miles 等人^[36]提出的基于中间层知识的算法, 在本文中用 PJ 代指。同时将蒸馏后的学生模型 FEN 与教师模型 ANFMGF 和其余十个高光图像分类算法进行对比: SVM^[37]、DSNet^[38]、CNC^[39]、RSSAN^[40]、SSTN^[41]、SSAN^[42]、SSAN^[43]、SSAtt^[44]、A²S²K-ResNet^[45]和 CVSSN^[46]。训练时, 重复 10 次得到分类结果然后取平均值。

表 1 展示了不同数据集下教师模型和学生模型参数和浮点数计算量。可以看到蒸馏得到的学生模型 FEN 和教师模型 ANFMGF 相比较, FLOPs 和 Params 有所下降。

表 1 不同数据集下教师模型 ANFMGF 和学生模型 FEN 的计算成本比较

数据集	ANFMGF 的 FLOPs(M)	FEN 的 FLOPs(M)	ANFMGF 的 Params(M)	FEN 的 Params(M)
IP	674.92	546.97(↓)	0.26	0.21(↓)
KSC	666.63	538.74(↓)	0.26	0.21(↓)
UP	641.11	513.41(↓)	0.25	0.20(↓)
UH	655.26	527.46(↓)	0.25	0.20(↓)

表 2 展示了在 IP、KSC、UP、UH 数据集下, 经过 KDFMGF 算法蒸馏后的学生模型 FEN 与教师模型以及多种高光图像分类算法的分类结果。

表 2 不同数据集下使用 ANFMGF 蒸馏得到的学生模型 FEN 和教师模型以及不同算法的分类结果对比

数据集	类别	SVM ^[37]	DSNet ^[38]	CNC ^[39]	RSSAN ^[40]	SSTN ^[41]	SSAN ^[42]	SSAN ^[43]	SSAtt ^[44]	A ² S ² K-ResNet ^[45]	CVSSN ^[46]	教师模型 (ANFMGF ^[41])	蒸馏后 FEN(Ours)
IP	OA	80.31	93.88	87.71	87.54	94.46	91.25	95.51	95.81	97.89	98.11	98.23	98.80
	AA	79.43	89.71	89.65	87.90	92.40	90.81	95.11	95.91	96.25	97.36	97.65	99.09
	Kappa	77.40	92.99	85.93	85.74	93.68	90.00	94.87	95.22	97.59	97.85	97.99	98.63
KSC	OA	88.20	95.83	92.48	94.18	96.07	92.63	96.38	97.09	97.14	98.62	99.00	99.01
	AA	83.24	90.08	88.08	90.80	93.60	88.40	94.12	95.23	95.56	97.82	98.48	98.64
	Kappa	86.84	95.33	91.62	93.52	95.63	91.79	95.97	96.77	96.81	98.47	98.89	98.90
UP	OA	93.28	99.27	99.15	99.06	98.90	99.53	99.74	99.63	99.85	99.84	99.90	99.94
	AA	93.01	98.42	98.88	98.68	98.49	99.40	99.70	99.56	99.80	99.79	99.86	99.93
	Kappa	91.02	99.02	98.87	98.75	98.55	99.37	99.66	99.51	99.80	99.78	99.87	99.92
UH	OA	86.15	95.24	94.40	93.02	94.98	95.07	95.26	95.35	97.94	98.22	98.64	98.65
	AA	88.17	93.56	95.09	93.85	95.45	95.65	95.77	95.95	98.21	98.28	98.61	98.83
	Kappa	85.01	94.85	93.95	92.45	94.57	94.67	94.88	94.97	97.77	98.07	98.53	98.54

结合表 1 和表 2, 可以看到在 IP、KSC、UP、UH 数据集下, 经 KDFMGF 蒸馏后的模型参数量和浮点数计算量均有所下降, 具体来说, IP 数据集中, 在浮点数计算量和参数量分别下降了 18.96%、19.23% 的情况下, 蒸馏后的学生模型分类性能达到了 98.80% 的 OA、99.09% 的 AA、98.63% 的 Kappa, 不

仅超越了教师模型还优于十种先进高光图像分类算法; KSC 数据集中, 蒸馏后学生模型 FLOPs 和 Params 分别下降了 19.18%、19.23%, 但分类性能在 OA、AA 和 Kappa 上达到了 99.01%、98.64%、98.90%, 同样超越了教师模型和其余十种先进的高光谱图像分类算法; 在 UP 数据集上, 蒸馏后学生模型的

FLOPs 和 Params 分别下降了 19.92%、20.00%，但分类性能超越了 ANFMGF 和其余十种先进高光谱图像分类算法，在 UH 数据集上，学生模型的 FLOPs 和 Params 分别下降了 19.50%、20.00%，但它的分类性能达到了 98.65% 的 OA、98.83% 的 AA 和 98.54% 的 Kappa，超越了教师模型 ANFMGF 和其余十种先进高光谱图像分类算法。

表 3 不同数据集下基于 ANFMGF 教师模型的不同知识蒸馏算法效果对比

数据集	类别	KDFMGF(Ours)	教师模型	学生模型	KD ^[29]	AT ^[30]	RKD ^[31]	CCKD ^[32]	ICKD ^[33]	DIST ^[34]	LS ^[35]	PJ ^[36]
IP	OA	98.80	98.23	97.90	97.02	95.94	98.27	97.34	97.29	97.78	98.51	97.75
	AA	99.09	97.65	96.40	98.08	97.39	98.50	97.65	97.87	97.38	98.87	95.57
	Kappa	98.63	97.99	97.60	96.60	95.39	98.03	96.97	96.91	97.47	98.30	97.44
KSC	OA	99.01	99.00	83.84	84.83	89.28	90.23	90.62	89.93	84.29	96.66	97.44
	AA	98.64	98.48	92.39	93.10	93.21	93.27	93.50	87.95	92.81	95.21	96.06
	Kappa	98.90	98.89	82.25	83.33	88.16	89.20	89.63	88.76	82.75	96.29	97.15
UP	OA	99.94	99.90	98.81	99.26	97.85	98.63	97.80	96.39	98.88	99.36	99.62
	AA	99.93	99.86	99.08	98.61	96.67	97.63	96.77	95.31	98.02	98.91	99.45
	Kappa	99.92	99.87	98.41	99.02	97.16	98.18	97.08	95.23	98.52	99.15	99.50
UH	OA	98.65	98.64	96.17	96.37	96.94	98.19	93.91	96.24	97.12	97.71	95.07
	AA	98.83	98.61	96.57	97.07	97.67	98.08	95.68	96.83	97.55	97.37	96.53
	Kappa	98.54	98.53	95.86	96.08	96.69	98.04	93.41	95.93	96.88	97.52	94.67

在 IP 数据集下，KDFMGF 蒸馏算法相比其他蒸馏算法，表现最佳。与次优蒸馏算法 LS 相比，KDFMGF 算法分别提升了 0.29% 的 OA、0.22% 的 AA 和 0.33% 的 Kappa。

在 KSC 数据集下，使用 KDFMGF 算法进行蒸馏得到的学生模型相较于教师模型，在模型参数量和计算成本显著下降的同时，OA、AA 和 Kappa 分别提升了 0.01%、0.16% 和 0.01%；同时和单独使用学生模型相比，蒸馏后得到的学生模型在 OA、AA 和 Kappa 上分别提升了 15.17%、6.25% 和 16.65%。这种显著提升主要归因于未蒸馏的学生模型在 KSC 数据集上的原始分类正确率太低，提升空间较大，同时使用 ANFMGF 作为教师模型给原始学生模型提供了大量有效知识。使蒸馏后的学生模型分类性能大幅提升。

在 UP 数据集上，使用 KDFMGF 算法得到的学生模型和未蒸馏学生模型相比，在 OA、AA 和 Kappa 性能指标上提高了 1.13%、0.85%、1.51%；在模型参数和计算开销显著降低的同时，比教师模型在 OA、AA、Kappa 上提高了 0.04%、0.07%、0.05%；与次优知识蒸馏方法 PJ 相比，KDFMGF 模型在 OA、AA 和 Kappa 上提升了 0.32%、0.48% 和 0.42%。

在 UH 数据集上，KDFMGF 蒸馏得到的学生模型和未蒸馏学生模型相比，在 OA、AA 和 Kappa 上提高了 2.48%、2.26%、2.68%；在模型参数和计算开销显著降低的同时，比教师模型在 OA、AA、Kappa 上提高了 0.01%、0.22%、0.01%；与次优知识蒸馏方法 RKD 相比，KDFMGF 模型在 OA、AA 和 Kappa

表 3 展示了在 IP、KSC、UP 和 UH 数据集下，以 ANFMGF 作为教师模型，采用不同的蒸馏算法训练出的学生模型 FEN 的分类结果。可以看到，KDFMGF 蒸馏算法在四个数据集上，与其他先进的蒸馏算法相比，表现最佳，KDFMGF 蒸馏后得到的学生模型分类性能优于教师模型。

上提升了 0.46%、0.75% 和 0.50%。图 7、图 10、图 11 和图 14 分别展示了在 IP、KSC、UP、UH 数据集上，将 ANFMGF 作为教师模型，采用不同的知识蒸馏算法得到的学生模型的整体效果。

图 8、图 9、图 12 和图 13 分别展示了在 IP、

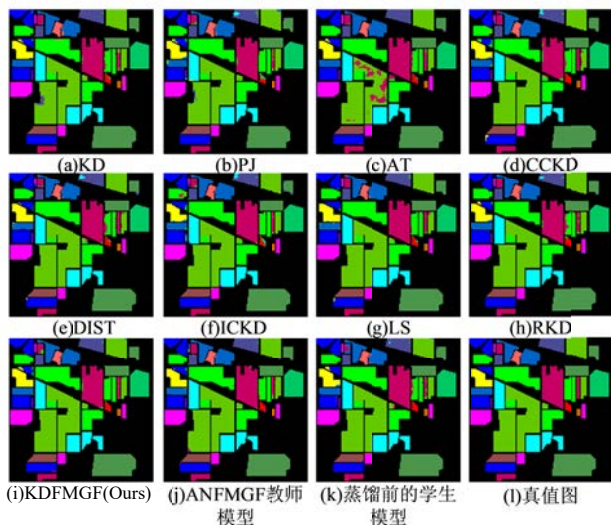


图 7 IP 数据集下基于 ANFMGF 教师模型的不同蒸馏算法的分类效果

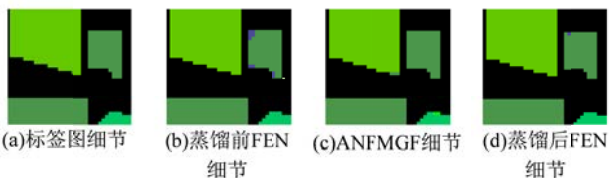


图 8 IP 数据集下基于 ANFMGF 教师模型 KDFMGF 算法的分类效果细节



图 9 KSC数据集下基于 ANFMGF 教师模型 KDFMGF 蒸馏算法蒸馏前后细节对比

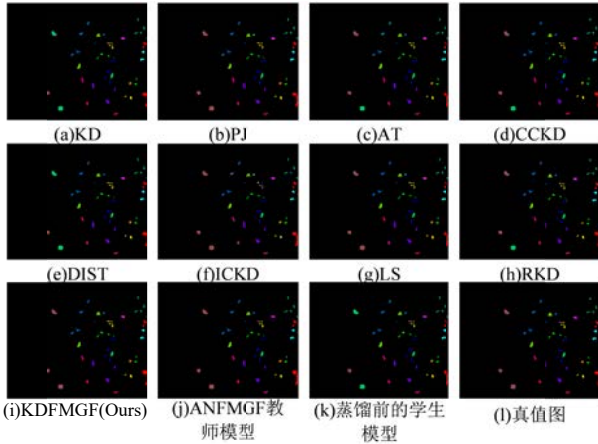


图 10 KSC数据集下基于 ANFMGF 教师模型的不同知识蒸馏算法的分类效果图

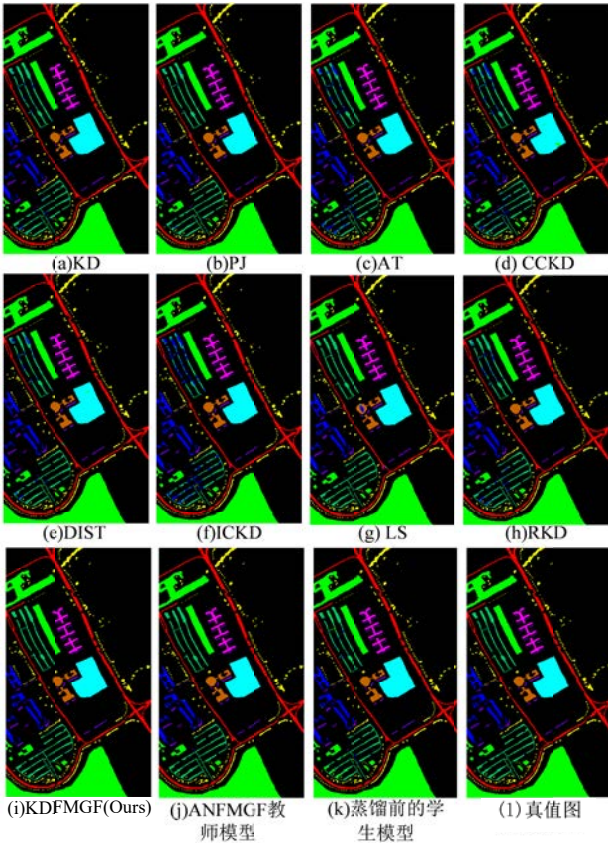


图 11 UP数据集下基于 ANFMGF 教师模型的不同知识蒸馏算法的分类效果图

KSC、UP、UH数据集上，使用 KDFMGF 蒸馏算法得到的学生模型分类效果细节图。从图 7、图 10、图 11 和图 14 中可以看到，使用 KDFMGF 蒸馏后的

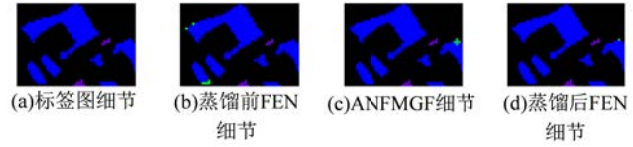


图 12 UP数据集下基于 ANFMGF 教师模型 KDFMGF 蒸馏算法蒸馏前后细节对比

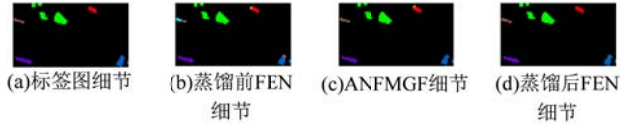


图 13 UH数据集下基于 ANFMGF 教师模型 KDFMGF 蒸馏算法蒸馏前后细节对比

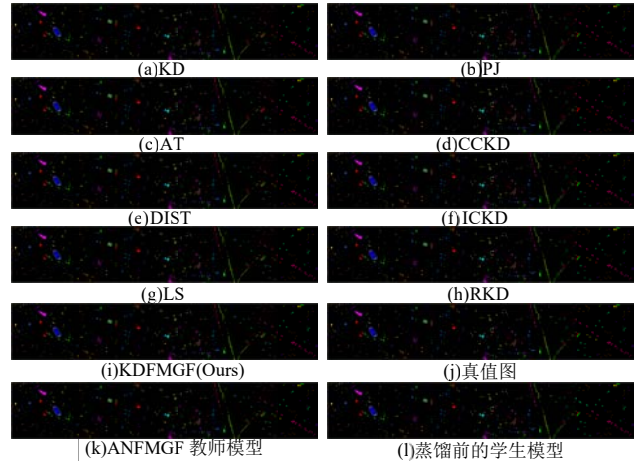


图 14 UH数据集下基于 ANFMGF 教师模型的不同知识蒸馏算法的分类效果图

学生模型 FEN 分类的效果最好。从图 8、图 9、图 12 和图 13 中可以看出，在四个数据集上，蒸馏后的学生模型不仅吸收了教师模型的正确分类知识，同时还对教师模型中的错误进行了纠正。

综上所述，在四个数据集上，KDFMGF 蒸馏算法都优于现有先进的知识蒸馏算法，并且 KDFMGF 蒸馏后的学生模型分类性能优于现在先进高光谱图像分类算法。KDFMGF 蒸馏后的学生模型不仅能从教师模型中学习正确分类知识，还能纠正教师模型中的错误。

本节将深入研究分数阶微分几何特征直接损失 (DLFDGF) 和分数阶微分几何特征蒸馏损失 (SLFDGF) 对算法的贡献。

4.3.3 随机选择样本的消融实验

本实验采用随机划分方法将 IP、KSC、UP 和 UH 数据集划分为训练集和测试集，进行消融实验。在本实验中，将本文提出的 KDFMGF 与其余五个变体进行比较。表 4 展示了在四个数据集下对损失模块的消融实验结果。

以 IP 数据集为例，与 KDFMGF 相比，变体 2 在 OA、AA 和 Kappa 性能指标上分别下降了 0.21%、

表4 四个数据集下基于 ANFMGF 教师模型的损失模块消融实验

数据集/组成	KDFMGF	变体 1	变体 2	变体 3	变体 4	变体 5	%
DLFDGF	✓	✓	✓				
SLFDGF	✓			✓			
原蒸馏损失	✓		✓	✓			✓
原直接损失	✓	✓	✓	✓	✓	✓	✓
IP	OA	98.80	98.51	98.59	98.60	97.90	97.02
	AA	99.09	97.18	98.58	98.38	96.40	98.08
	Kappa	98.63	98.30	98.39	98.40	97.60	96.60
KSC	OA	99.01	89.09	98.56	98.77	83.84	84.83
	AA	98.64	93.33	97.96	98.35	92.39	93.10
	Kappa	98.90	87.96	98.39	98.63	82.25	83.33
UP	OA	99.94	99.64	99.92	99.92	98.81	99.26
	AA	99.93	99.26	99.89	99.88	99.08	98.61
	Kappa	99.92	99.52	99.90	99.90	98.41	99.02
UH	OA	98.65	96.53	98.48	98.31	96.17	96.37
	AA	98.83	97.25	98.59	98.56	96.57	97.07
	Kappa	98.54	96.25	98.36	98.18	95.86	96.08

0.51%、0.24%；对比变体 3 和变体 5 可以发现，变体 5 的 OA、AA、Kappa 下降了 1.58%、0.30%、1.8%。这表明分数阶微分几何特征蒸馏损失模块可从 ANFMGF 教师模型中学习与光谱变异像素点分类有关的知识，从而提高分类的准确性。

与 KDFMGF 相比，变体 3 的 OA、AA 和 Kappa 下降了 0.2%、0.71% 和 0.23%；变体 2 和变体 5 相比，变体 5 的 OA、AA 和 Kappa 系数分别下降了 1.57%、0.50% 和 1.79%。在其他三个数据集上，性能变化趋势和 IP 数据集相同。

这表明分数阶微分几何特征直接损失使学生模型 FEN 纠正教师模型在光谱变异像素点相关的错误分类知识，从正确标签中学习知识，进一步补充教师模型学习不足的部分，从而提高模型分类性能。

综上所述，分数阶微分几何特征蒸馏损失使学生模型充分学习教师模型中关于光谱变异像素点分类的正确知识，而分数阶微分几何特征直接损失使学生模型从标签中进行学习，纠正教师模型中潜在误分类信息，补充教师模型学习不足等问题，从而显著提升分类效果。因此，分数阶微分几何特征蒸馏损失和分数阶微分几何特征直接损失在 KDFMGF 算法中各有作用，相辅相成，两者共同作用，让模型取得的分类性能最好。

4.3.4 空间不相交样本的对比实验

本节旨在训练集与测试集空间相互独立情况下，验证 KDFMGF 累积模式的有效性。本实验采用空间不相交划分方法将 LK、UP、UH 数据集拆分为训练集和测试集，并与现行的知识蒸馏算法进行对比：KD^[29]、AT^[30]、RKD^[31]、CCKD^[32]、ICKD^[33]、DIST^[34]、LS^[35]、PJ^[36]。并在空间不相关数据集上，将蒸馏后的学生模型 FEN 与其余十个高光谱图像分类算法进行对比：SVM^[37]、DSNet^[38]、CNC^[39]、RSSAN^[40]、SSTN^[41]、SSAN^[42]、SSAN^[43]、SSAtt^[44]、A²S²K-ResNet^[45]和 CVSSN^[46]。

表 5 展示了在 UP、LK、UH 数据集下，以 ANFMGF 为教师模型，采用不同的蒸馏算法得到的学生模型分类效果。从表中可以看到，与随机划分方法相比，采用空间不相交划分方法的学生模型整体分类性能均有所下降。这是因为随机划分模型忽略了训练集和测试集之间的空间依赖性，从而导致分类性能被高估。

表 6 展示了使用 KDFMGF 蒸馏后得到的 FEN 和教师模型 ANFMGF 以及其他十种高光谱图像分类算法的分类效果。其中 A²S²K-ResNet 方法在 LK 数据集上对内存资源需求较高，在本文实验环境中无法实现，导致表 6 中缺少数据。从两张表可知，尽管在不相交数据集上的模型性能有所下降，但使用

表5 在各空间不相交数据集下基于 ANFMGF 教师模型的不同知识蒸馏算法效果对比

数据集	性能指标	KDFMGF	教师模型	学生模型	KD ^[29]	AT ^[30]	RKD ^[31]	CCKD ^[32]	ICKD ^[33]	DIST ^[34]	LS ^[35]	PJ ^[36]	%
UP	OA	96.53	91.93	89.58	91.83	93.48	93.00	94.17	92.22	91.05	93.62	94.50	
	AA	96.30	92.42	93.22	94.21	94.35	92.78	94.92	94.16	93.15	94.03	95.01	
	Kappa	95.29	89.01	85.49	88.77	91.10	90.47	92.04	89.29	87.71	91.36	92.47	
LK	OA	99.52	99.36	98.49	98.71	98.66	98.78	98.32	98.04	98.88	98.99	98.15	
	AA	98.37	97.75	93.61	94.97	94.68	94.56	93.84	93.21	95.36	95.61	93.54	
	Kappa	99.37	99.16	98.03	98.31	98.24	98.40	97.80	97.44	98.53	98.67	97.59	
UH	OA	85.99	82.81	81.38	83.98	82.27	85.06	80.63	83.56	82.52	84.13	84.99	
	AA	89.30	85.98	82.36	84.78	85.43	84.74	83.90	86.98	85.08	88.10	84.99	
	Kappa	84.86	81.43	79.90	82.67	80.84	83.86	79.03	82.23	81.13	82.84	83.78	

KDFMGF 蒸馏后的学生模型 FEN 的性能显著优于目前先进高光谱图像分类算法,同时 KDFMGF 的蒸馏算法优于现在先进的知识蒸馏算法。

表 7 展示了教师模型 ANFMGF 和蒸馏后的学生模型 FEN 之间的模型参数量和计算开销。可以看到,在 LK 数据集上, KDFMGF 方法蒸馏后的学生模型以 81.48% 的参数量和 81.68% 的浮点数计算量取

得了优于 ANFMGF 模型的性能。在 UP 数据集上, KDFMGF 方法蒸馏后的学生模型以 80.00% 的参数量和 80.08% 的浮点数计算量取得了优于 ANFMGF 模型的性能。在 UH 数据集上,使用 KDFMGF 方法蒸馏得到的学生模型以 80.00% 的参数量和 80.50% 的浮点数计算量取得了优于 ANFMGF 模型的性能。

表 6 在各空间不相交数据集上使用 ANFMGF 蒸馏得到的学生模型 FEN 和教师模型以及不同算法的分类结果对比 /%

数据集	性能指标	SVM ^[37]	DSNet ^[38]	CNC ^[39]	RSSAN ^[40]	SSTN ^[41]	SSAN ^[42]	SSAN ^[43]	SSAtt ^[44]	A ² S ² K-ResNet ^[45]	CVSSN ^[46]	教师模型 (ANFMGF ^[4])	KDFMGF 蒸馏后 FEN
UP	OA	75.86	87.80	85.49	85.91	87.64	88.17	87.61	89.47	85.49	91.07	91.93	96.53
	AA	76.58	86.92	83.35	83.28	87.70	86.44	88.15	89.72	88.36	91.89	92.42	96.30
	Kappa	69.76	83.59	80.54	80.97	83.62	84.00	83.34	85.66	81.26	87.85	89.01	95.29
LK	OA	91.23	96.54	96.35	96.98	98.09	97.27	98.37	97.95		99.30	99.36	99.52
	AA	78.08	97.28	88.48	89.70	93.51	90.45	93.97	92.42		97.61	97.75	98.37
	Kappa	88.72	95.49	95.24	96.06	97.50	96.43	97.86	97.32		99.08	99.16	99.37
UH	OA	76.76	78.40	74.70	76.11	79.25	77.24	81.07	80.28	82.58	82.22	82.81	85.99
	AA	78.49	79.60	77.07	77.60	80.68	79.93	83.67	82.75	85.90	85.15	85.98	89.30
	Kappa	74.98	76.66	72.65	74.19	77.61	75.39	79.55	76.88	81.22	80.78	81.43	84.86

表 7 在各空间不相交样本下教师模型 ANFMGF 和学生模型 FEN 计算成本对比

数据集	模型	FLOPs(M)	Params(M)
LK	教师模型 ANFMGF	699.37	0.27
	蒸馏后 FEN	571.23(↓)	0.22(↓)
UP	教师模型 ANFMGF	641.11	0.25
	蒸馏后 FEN	513.41(↓)	0.20(↓)
UH	教师模型 ANFMGF	655.26	0.25
	蒸馏后 FEN	527.46(↓)	0.20(↓)

在 UP 数据集上, KDFMGF 蒸馏后的学生模型 FEN 相比于教师模型 ANFMGF 提升显著,其原因在于 UP 样本量较少却分布集中,导致在空间不相关划分的训练集中边缘与光谱变异像素较少, ANFMGF 教师模型中的模块对边缘以及光谱变异像素点的学习不够全面。而 KDFMGF 通过标签学习弥补教师模型学习不足的问题,显著提升分类性能。

因此,在累积模式下,采用空间不相交样本, KDFMGF 算法不仅能够减少参数量和计算开销,还能让蒸馏得到的学生模型分类性能超越教师模型且优于现有先进高光谱图像分类算法。

4.3.5 空间不相交样本的消融实验

本节旨在确保训练集和测试集空间独立的前提下,将深入分析 KDFMGF 各损失模块包含分数阶微分几何特征直接损失和分数阶微分几何特征蒸馏损失对算法性能贡献。本实验采用空间不相交划分方法将 LK、UP、UH 数据集划分为训练集和测试集。

在三个数据集上进行消融实验,具体而言将 KDFMGF 与其他 5 个模型变体进行比较。表 8 展示了各不相交样本下对各损失模块的消融实验结果。

由表 8 可知,与 KDFMGF 相比,变体 2 在 LK

表 8 在各空间不相交样本下 KDFMGF 的消融实验结果 /%

数据集/组成	KDFMGF	变体 1	变体 2	变体 3	变体 4	变体 5	
DLFDGF	✓	✓	✓				
SLFDGF	✓			✓			
原蒸馏损失	✓		✓	✓		✓	
原直接损失	✓	✓	✓	✓	✓	✓	
LK	OA/%	99.52	98.75	99.50	99.36	98.49	98.71
	AA/%	98.37	94.48	97.92	97.57	93.61	94.97
	Kappa/%	99.37	98.36	99.35	99.16	98.03	98.31
UP	OA/%	96.53	92.55	94.83	95.65	89.58	91.83
	AA/%	96.30	94.13	95.43	95.75	93.22	94.21
	Kappa/%	95.29	89.74	92.95	94.07	85.49	88.77
UH	OA/%	85.99	83.68	85.82	85.41	81.38	83.98
	AA/%	89.30	89.11	88.03	86.05	82.36	84.78
	Kappa/%	84.86	82.36	84.63	84.25	79.90	82.67

数据集上,其OA、AA和Kappa下降了0.02%、0.45%、0.02%,在UP数据集上,下降了1.7%OA、0.87%AA和2.43%Kappa,在UH数据集上,下降了0.17%OA、1.27%AA和0.23%Kappa;变体3和变体5进行比较,变体5在LK数据集上OA、AA、Kappa性能指标分别下降了0.65%、2.6%、0.85%,在UP数据集上OA、AA、Kappa性能指标分别下降了3.82%、1.54%、5.3%,在UH数据集上OA、AA、Kappa性能指标分别下降了1.43%、1.27%、1.58%。这说明分数阶微分几何特征蒸馏损失能让学生模型从教师模型上充分学习光谱变异像素点分类的相关知识,从而提高分类正确率。

同样,表8可知,在所有数据集下,与KDFMGF相比,变体3的性能指标均有所下降;变体2和变体5相比,变体5在去掉分数阶微分几何特征直接损失后,OA、AA和Kappa均有所下降。这表明分数阶微分几何特征直接损失,能够纠正教师模型ANFMGF关于光谱变异像素点分类的错误知识,同时通过标签学习弥补教师模型可能存在的学习不足等问题,从而提高分类性能。因此,在累积模式下,采用空间不相交样本,分数阶微分几何特征直接损失和分数阶微分几何特征蒸馏损失在KDFMGF算法中各有作用,协同提高分类性能。

4.4 消除模式下KDFMGF的实验结果

本节选择的数据集以及划分方法和实验设置已在第4.1和第4.2节中详细介绍,这里不再赘述。

4.4.1 教师网络和学生网络的设置

在消除模式下,本章节在实验时选择CVSSN^[47]模型作为教师模型,选择图6中的FEN模型作为学生模型。

4.4.2 随机选择样本的对比实验结果

本节旨在消除模式下,采用随机划分方法评估KDFMGF算法的有效性,IP、KSC、UP、UH数据

集被划分为训练集和测试集,与现行的先进蒸馏算法进行比较,包括:KD^[29]算法、AT^[30]算法、RKD^[31]算法、CCKD^[32]算法、ICKD^[33]算法、DIST^[34]算法、LS^[35]和PJ^[36]。

表9展示了不同数据集下教师模型CVSSN和学生模型FEN的模型参数量和浮点数计算量,数据显示,学生模型的Params和FLOPs都有所降低。

表10展示了在IP、KSC、UP、UH数据集下,将CVSSN作为教师模型,采用KDFMGF蒸馏算法以及其他蒸馏算法得到的学生模型FEN的分类性能。数据显示,KDFMGF蒸馏算法在四个数据集上均优于其他蒸馏算法。经KDFMGF蒸馏的学生模型比原学生模型的性能更优,同时还优于教师模型CVSSN。

结合表9和表10可以看到,在IP数据集下,学生模型FEN在Params和FLOPs上分别下降18.89%和19.23%的同时,蒸馏后学生模型的OA、AA和Kappa分别为98.73%、98.21%和98.55%,比未蒸馏学生模型在OA、AA和Kappa上分别提高了0.83%、1.81%和0.95%,比教师模型在OA、AA和Kappa上提高了0.62%、0.85%和0.70%。与次优知识蒸馏方法RKD相比,KDFMGF模型进一步提升了0.24%的

表9 所有数据集下教师模型CVSSN和学生模型FEN的计算成本对比分析

数据	模型	FLOPs(M)	P
IP	教师模型 CVSSN	674.39	0
	学生模型 FEN	546.97(t)	0
KSC	教师模型 CVSSN	666.17	0
	学生模型 FEN	538.74(t)	0
UP	教师模型 CVSSN	640.83	0
	学生模型 FEN	513.41(t)	0
UH	教师模型 CVSSN	654.89	0
	学生模型 FEN	527.46(t)	0

表10 不同数据集下基于CVSSN教师模型的不同知识蒸馏算法效果对比

数据集	类别	KDFMGF	教师模型	学生模型	KD ^[29]	AT ^[30]	RKD ^[31]	CCKD ^[32]	ICKD ^[33]	DIST ^[34]	LS ^[35]	PJ ^[36]
IP	OA	98.73	98.11	97.90	98.08	98.26	98.49	97.82	97.35	97.74	98.17	97.97
	AA	98.21	97.36	96.40	96.54	97.18	97.93	96.10	97.14	96.34	96.85	96.55
	Kappa	98.55	97.85	97.60	97.82	98.02	98.28	97.52	96.98	97.43	97.91	97.69
KSC	OA	98.69	98.62	83.84	87.91	89.07	88.34	84.38	90.08	92.94	94.77	97.42
	AA	98.17	97.82	92.39	94.06	92.76	93.08	92.82	86.05	94.60	94.14	94.92
	Kappa	98.54	98.47	82.25	86.67	87.92	87.11	82.83	88.93	92.17	94.19	97.13
UP	OA	99.89	99.84	98.81	99.12	97.87	98.12	97.64	98.32	99.28	99.67	99.48
	AA	99.89	99.79	99.08	98.35	96.74	98.47	96.65	97.25	98.66	99.43	99.10
	Kappa	99.85	99.78	98.41	98.84	97.18	97.53	96.87	97.78	99.05	99.57	99.31
UH	OA	98.46	98.22	96.17	97.19	96.43	97.77	93.58	96.16	97.50	97.45	94.60
	AA	98.56	98.28	96.57	97.11	97.30	97.53	94.74	97.06	97.17	98.02	96.27
	Kappa	98.34	98.07	95.86	96.96	96.14	97.59	93.06	95.85	97.30	97.25	94.16

OA、0.28%的AA和0.27%的Kappa。

在KSC数据集下,相较于教师模型,KDFMGF蒸馏后学生模型,在Params和FLOPs上分别降低了19.13%和19.23%,OA、AA、Kappa上分别提高了0.07%、0.35%和0.07%;相较于未蒸馏学生模型,提高了14.85%的OA、5.78%的AA和16.29%的Kappa。蒸馏后学生模型相较于未蒸馏学生模型在KSC数据集上提升显著,主要原因在于未蒸馏学生模型在KSC数据集上的原始分类性能较低,给蒸馏算法留下了较大的优化空间;同时,高性能的教师模型CVSSN提供了除光谱变异像素点分类信息外的丰富且正确的分类知识,促使蒸馏后学生模型分类性能显著提升。与次优知识蒸馏方法PJ相比,KDFMGF算法提升了1.27%的OA、3.25%的AA和1.41%的Kappa。

在UP数据集下,KDFMGF蒸馏后学生模型在Params和FLOPs上分别降低了20%和19.88%的情况下,性能优于教师模型,分别提高了0.05%的OA、0.1%的AA和0.07%的Kappa;相较于未蒸馏学生模型,在OA、AA、Kappa上提升了1.08%、0.81%和1.44%;与次优知识蒸馏算法LS相比,分别提升了0.22%的OA、0.46%的AA和0.28%的Kappa。

在UH数据集上,KDFMGF蒸馏后学生模型以80.00%的Params和80.54%的FLOPs取得了98.46%、98.56%和98.34%的OA、AA和Kappa,相较于教师模型,提高了0.24%的OA、0.28%的AA和0.27%的Kappa,与未蒸馏学生模型相比,在OA、AA和Kappa上提高了2.29%、1.99%和2.48%。

图15、图18、图19和图21分别展示了在IP、KSC、UP、UH数据集下,基于CVSSN教师模型,采用不同的蒸馏算法得到的学生模型FEN的分类效果;可以看出KDFMGF蒸馏后学生模型分类效果最好。图16、图17、图20和图22单独展示了在IP、KSC、UP、UH数据集上,使用KDFMGF算法进行蒸馏前后学生模型分类性能的可视化细节部分;可以看到蒸馏后的FEN学生模型主要从标签中学习光谱变异像素点的正确分类知识,抑制CVSSN中错误信息的传递,从而提高分类效果。

4.4.3 随机选择样本的消融实验结果

本节旨在进一步探索在消除模式下,两个损失模块:分数阶微分几何特征蒸馏损失和分数阶微分

几何直接损失对算法的贡献,除此之外,还探索两个模式下融合模块的贡献。在本节中,采用随机划分方法将IP、KSC、UP和UH数据集拆分为训练集和测试集,并进行两个消融实验。

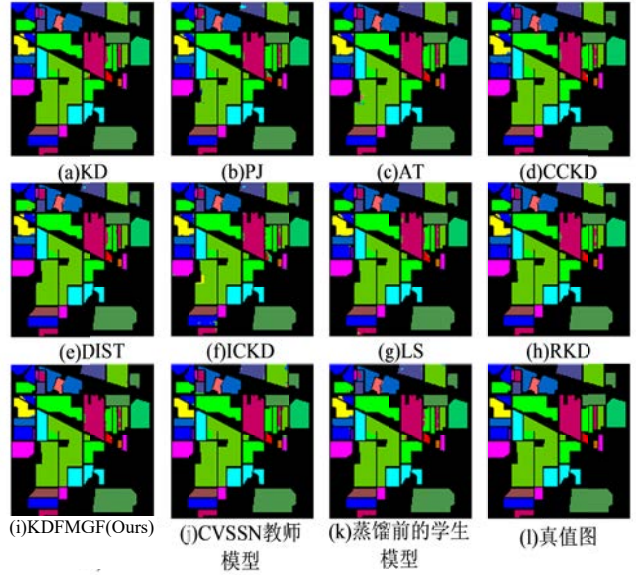


图15 IP数据集下基于CVSSN教师模型的不同蒸馏算法的效果图

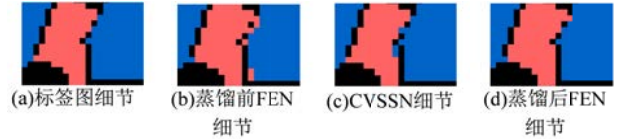


图16 IP数据集下基于CVSSN教师模型KDFMGF蒸馏算法的蒸馏效果细节

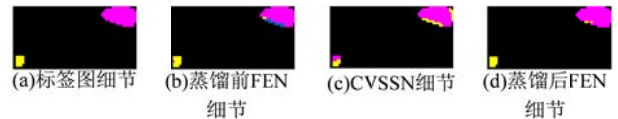


图17 KSC数据集下基于CVSSN教师模型KDFMGF蒸馏算法的蒸馏效果细节图

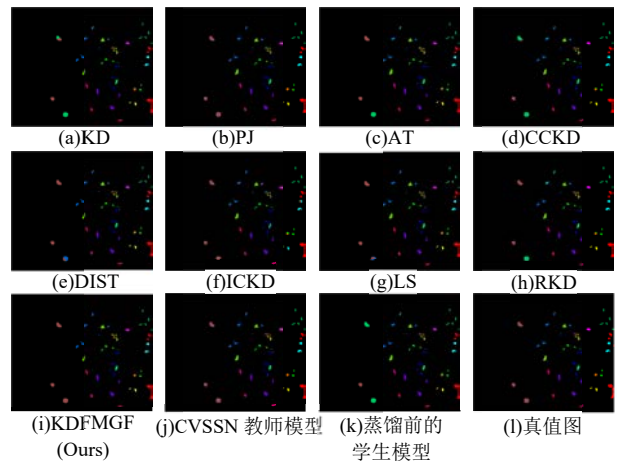


图18 KSC数据集下基于CVSSN教师模型的不同蒸馏算法的效果图

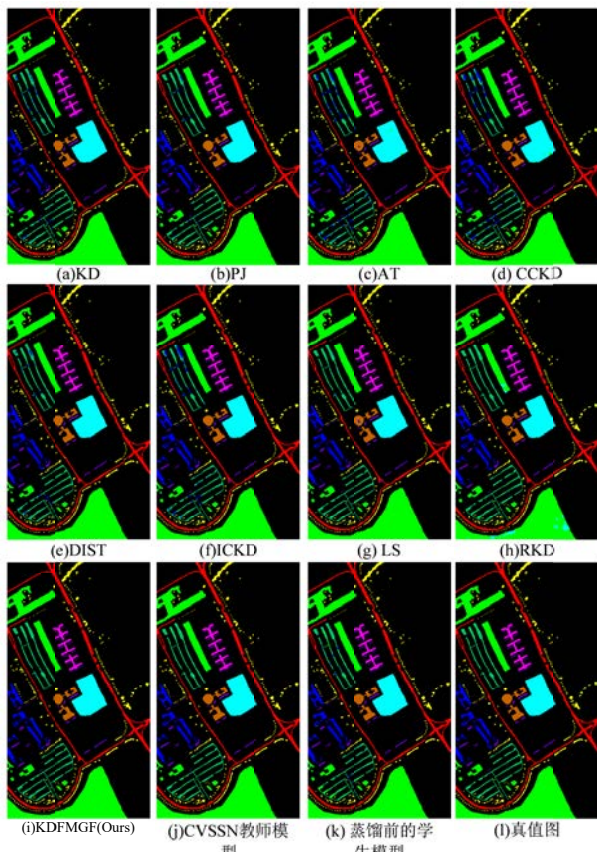


图 19 UP 数据集下基于 CVSSN 教师模型的不同蒸馏算法的效果图

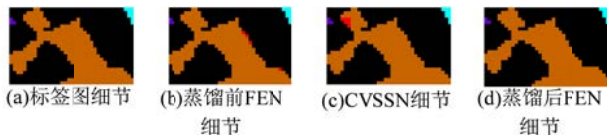


图 20 UP 数据集下基于 CVSSN 教师模型 KDFMGF 蒸馏算法的蒸馏效果细节图

表 11 展示了在不同数据集下，基于 CVSSN 教师模型的 KDFMGF 算法消除模式下的各损失模块

表 11 所有数据集下基于 CVSSN 教师模型的损失模块消融实验

数据集/组成		KDFMGF	变体 1	变体 2	变体 3	变体 4	变体 5	%
DLFDGF	DLFDGF	✓	✓	✓				
	SLFDGF	✓			✓			
	原蒸馏损失	✓		✓	✓			✓
	原直接损失	✓	✓	✓	✓	✓	✓	✓
IP	OA(%)	98.73	98.55	98.62	98.51	97.90	98.08	
	AA(%)	98.21	98.20	97.18	98.16	96.40	96.54	
	Kappa(%)	98.55	98.35	98.43	98.30	97.60	97.82	
KSC	OA(%)	98.69	93.61	98.52	98.54	83.84	87.91	
	AA(%)	98.17	95.05	97.80	98.13	92.39	94.06	
	Kappa(%)	98.54	92.92	98.35	98.37	82.25	86.67	
UP	OA(%)	99.89	99.77	99.87	99.86	98.81	99.12	
	AA(%)	99.89	99.60	99.84	99.82	99.08	98.35	
	Kappa(%)	99.85	99.69	99.83	99.82	98.41	98.84	
UH	OA(%)	98.46	98.16	98.28	98.19	96.17	97.19	
	AA(%)	98.56	98.19	98.36	98.53	96.57	97.11	
	Kappa(%)	98.34	98.01	98.14	98.05	95.86	96.96	

的对分类性能的影响。数据显示，KDFMGF 与变体 2 相比，变体 2 的分类性能在所有数据集上都有所下降，变体 3 与变体 5 相比，变体 5 在所有数据集上的分类性能都有所下降，这表明分数阶微分几何特征蒸馏损失能够抑制 CVSSN 教师模型中关于光谱变异像素点错误分类知识，提高分类的准确性。

对比 KDFMGF 和变体 3 可以发现先，变体 3 在所有数据集下的分类性能都有所下降，对比变体 2 和变体 5 可以发现，变体 5 在所有数据集下的分类性能均有所下降，说明分数阶微分几何特征直接损失能够从标签学习光谱变异像素点的正确分类知识，从而提高分类性能。

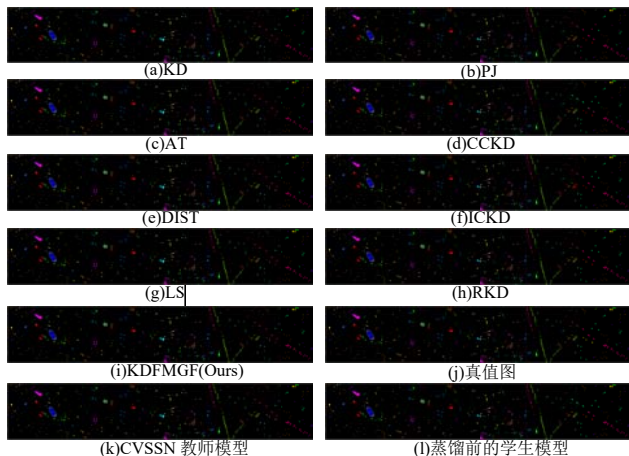


图 21 UH 数据集下基于 CVSSN 教师模型的不同蒸馏算法的效果图



图 22 UH 数据集下基于 CVSSN 教师模型 KDFMGF 蒸馏算法的蒸馏效果细节图

综合来说,分数阶微分几何特征直接损失使学生模型能够从标签中学习正确的光谱变异像素点的分类知识,分数阶微分几何特征蒸馏损失能够抑制学生模型从教师模型 CVSSN 中学习关于光谱变异像素点的错误分类知识,让学生模型主要关注教师模型中的正确分类知识。分数阶微分几何特征直接

损失和分数阶微分几何特征蒸馏损失相辅相成,共同作用于 KDFMGF 算法中,提高模型分类性能。

表 12 展示了在不同数据集下,基于 CVSSN 教师模型的不同混合模块模式的消融实验。CVSSN 教师模型并未解决由于光谱变异引起的误分类问题。

表 12 所有数据集下基于 CVSSN 教师模型的混合模块消融实验

	DLFDGF	混合模块消除模式	混合模块累积模式	IP			KSC			UP			UH		
				OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa
消除模式	✓	✓		98.73	98.21	98.55	98.69	98.17	98.54	99.89	99.89	99.85	98.46	98.56	98.34
累积模式	✓		✓	98.67	97.94	98.49	98.00	96.42	97.77	99.63	99.32	99.51	98.09	98.21	97.93
变体 1		✓		98.51	98.16	98.30	98.54	98.13	98.37	99.86	99.82	99.82	98.19	98.53	98.05
变体 2			✓	98.40	97.75	98.18	85.65	92.27	84.20	98.60	97.67	98.15	98.08	97.51	97.93

与 KDFMGF 消除模式相比, KDFMGF 累积模式下所有数据集上的分类性能均有所下降,这表明,对于未解决误分类问题的教师模型,在考虑从标签中学习光谱变异像素点的正确分类知识的前提下,消除模式比累积模式的优势更大,消除模式能够抑制教师模型对错误知识的传播。变体 1 和变体 2 相比较可以发现,变体 2 的分类性能在不同数据集上均有不同程度的下降,这表明,在不考虑从标签中学习光谱变异像素点的正确分类知识的前提下,消除模式更适合未解决误分类问题的教师模型。

综上所述,针对未解决光谱变异所导致的误分类问题的教师模型,消除模式更适合,消除模式能够有效剔除教师模型关于光谱变异像素点的错误知识点,使学生模型聚焦于正确的分类知识,从而纠正因光谱变异导致的误分类问题,提高分类性能。

4.4.4 空间不相交样本的对比实验结果

在随机划分方法中,不能确定训练集和测试集之间空间独立性,从而导致分类性能虚高。因此,在本节中,我们采用空间不相交划分方法,将 LK、UP、UH 数据集划分为训练集和测试集,与现行知识蒸馏

算法进行比较: KD^[29]、AT^[30]、RKD^[31]、CCKD^[32]、ICKD^[33]、DIST^[34]、LS^[35]、PJ^[36]。

表 13 展示了在各空间不相交数据集下教师模型 CVSSN 和学生模型 FEN 之间的计算成本的差别。表 14 展示了在各空间不相交数据集下,基于 CVSSN 教师模型,使用不同知识蒸馏算法得到的学生模型的性能。

由表 14 可知,在消除模式下,与随机划分相比,在空间不相交数据集中,各种蒸馏方法得到的学生模型的性能都有所下降,不可否认的是,即使存在分类性能的总体下降,但 KDFMGF 蒸馏方法仍然比现行蒸馏方法性能更优,同时 KDFMGF 蒸馏得到的学生模型比原学生模型和教师模型都有所提升。

表 13 在各空间不相交数据集下教师模型 CVSSN 和学生模型 FEN 计算成本对比

数据集	模型	FLOPs(M)	Params(M)
LK	教师模型 CVSSN	698.66	0.27
	蒸馏后 FEN	571.23(↓)	0.22(↓)
UP	教师模型 CVSSN	640.83	0.25
	蒸馏后 FEN	513.41(↓)	0.20(↓)
UH	教师模型 CVSSN	654.89	0.25
	蒸馏后 FEN	527.46(↓)	0.20(↓)

表 14 各空间不相交数据集下基于 CVSSN 教师模型的不同知识蒸馏算法效果对比

数据集	指标	KDFMGF	教师模型	学生模型	KD ^[29]	AT ^[30]	RKD ^[31]	CCKD ^[32]	ICKD ^[33]	DIST ^[34]	LS ^[35]	PJ ^[36]
UP	OA	94.58	91.07	89.58	93.02	93.93	91.53	93.35	90.13	91.15	93.32	92.64
	AA	95.52	91.89	93.22	93.82	93.24	92.93	94.59	91.39	92.00	93.77	93.44
	Kappa	92.60	87.85	85.49	90.40	91.78	88.43	90.83	86.80	87.89	90.97	89.83
LK	OA	99.51	99.30	98.49	98.77	98.50	98.86	98.63	97.26	98.62	98.35	98.15
	AA	97.76	97.61	93.61	94.86	94.53	95.37	94.67	91.52	94.28	94.60	93.44
	Kappa	99.35	99.08	98.03	98.39	98.03	98.51	98.20	96.43	98.20	97.85	97.57
UH	OA	85.39	82.22	81.38	82.03	82.38	84.94	82.46	82.48	83.73	83.93	82.32
	AA	88.25	85.15	82.36	83.14	82.65	86.39	83.64	85.88	85.08	87.72	86.81
	Kappa	84.24	80.78	79.90	80.60	80.99	83.72	81.06	81.06	82.43	82.63	80.89

在表 14 中,在 UP、UH 数据集下,蒸馏后的学生模型相较于教师模型和未蒸馏学生模型的性能提

升最大,这是由于 UP、UH 的样本量较少,训练集和测试集空间相互独立时,训练集中产生光谱变异

的像素点较少，对于光谱变异像素点的标签学习不足。而 KDFMGF 中的 DLFDGF 模块通过标签对光谱变异像素点进行针对性学习，因此，在 UP 和 UH 数据集上经过 KDFMGF 蒸馏后的学生模型性能提升显著。

结合由表 13 可知，在消除模式下，与随机划分相比，在空间不相交数据集中，各种蒸馏方法得到的学生模型的性能都有所下降，不可否认的是，即使存在分类性能的总体下降，但 KDFMGF 蒸馏方法仍然比现行蒸馏方法性能更优，同时 KDFMGF 蒸馏得到的学生模型比原学生模型和教师模型都有所提升。

结合表 13 可知，在消除模式下，在各空间不相交样本中，蒸馏后的 FEN 相较于教师模型 CVSSN 的浮点数计算量和模型参数量都有不同程度的下降，但蒸馏后学生模型的性能并未下降，甚至在各空间不相交样本中都取得了比教师模型 CVSSN 更好的分类性能。

4.4.5 空间不相交样本的消融实验结果

本节将在训练集和测试集空间不相交的前提下，深入探索分数阶微分几何特征直接损失和分数阶微分几何特征蒸馏损失在 KDFMGF 当中的贡献，同时探索不同的混合模块在 KDFMGF 中的贡献。本实验采用空间不相交划分方法，将 UP、UH 和 LK 数据集划分为训练集和测试集，进行两个消融实验。

表 15 展示了消除模式下，各空间不相交样本中，将 CVSSN 作为教师模型，分数阶微分几何特征直接损失和分数阶微分几何特征蒸馏损失对分类性能的影响，具体来说就是将 KDFMGF 与五个变体进行比较。从表 15 中可知，将 KDFMGF 与变体 2 进行比较，变体 2 在各空间不相交样本中的三个性能指标上都有不同程度的下降，将变体 3 与变体 5 进行比较，变体 5 在不同数据集上的分类性能都有所下降，这说明在消除模式下，分数阶微分几何特征蒸馏损失能够抑制教师模型 CVSSN 中光谱变异像素点错误分类知识的传播，从而提高分类的准确性。

表 15 各空间不相交样本下基于 CVSSN 教师模型的损失模块消融实验

数据集/组成		KDFMGF	变体 1	变体 2	变体 3	变体 4	变体 5	%
	DLFDGF	✓	✓	✓				
	SLFDGF	✓			✓			
	原蒸馏损失	✓		✓	✓		✓	
	原直接损失	✓	✓	✓	✓	✓	✓	
LK	OA	99.51	98.88	99.38	99.19	98.49	98.77	
	AA	97.76	95.11	97.61	96.75	93.61	94.86	
	Kappa	99.35	98.53	99.19	98.94	98.03	98.39	
UP	OA	94.58	93.35	94.45	93.84	89.58	93.02	
	AA	95.52	94.23	95.25	95.00	93.22	93.82	
	Kappa	92.60	90.98	92.42	91.59	85.49	90.40	
UH	OA	85.39	82.48	85.16	85.29	81.38	82.03	
	AA	88.25	87.77	87.88	88.22	82.36	83.14	
	Kappa	84.24	81.07	83.97	84.11	79.90	80.60	

将 KDFMGF 与变体 3 进行比较，变体 3 在所有数据集下，分类性能都有所下降；将变体 2 与变体 5 进行比较，变体 5 的性能指标在所有数据集中都有所下降，这说明分数阶微分几何特征直接损失能够让学生模型从标签中学习光谱变异像素点的正确分类知识，从而提高分类性能。

综上所述，分数阶微分几何特征蒸馏损失能够让学生模型减少对教师模型关于光谱变异像素点的错误分类知识的关注，重点关注正确分类知识，分

数阶微分几何特征直接损失使学生模型从标签中学习关于光谱变异像素点的正确分类信息，弥补单从教师模型中进行学习的不足之处。两个损失模块相辅相成共同作用于 KDFMGF，提高蒸馏后学生模型的性能。

表 16 验证了混合模块的两个模式对算法性能的影响。具体来说，本消融实验基于 CVSSN 教师模型，对 KDFMGF 的消除模式、KDFMGF 的累积模式和两个 KDFMGF 的变体进行对比分析。

表 16 各空间不相交数据集下基于 CVSSN 教师模型的混合模块消融实验

	DLFDGF	混合模块 消除模式	混合模块 累积模式	LK			UP			UH			%
				OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa	
消除模式	✓	✓		99.51	97.76	99.35	94.58	95.52	92.60	85.39	88.25	84.24	
累积模式	✓		✓	99.23	96.86	98.98	93.61	94.92	91.25	85.30	86.81	84.15	
变体 1		✓		99.19	96.75	98.94	93.84	95.00	91.59	85.29	88.22	84.11	
变体 2			✓	98.87	95.07	98.52	93.31	94.66	90.78	84.68	84.66	83.46	

在各空间不相交样本中,与 KDFMGF 累积模式相比, KDFMGF 消除模式在 OA、AA 和 Kappa 上都有所提升,在排除分数阶微分几何特征直接损失对算法的影响后,变体 1(消除模式)和变体 2(累积模式)对比发现,前者在 OA、AA 和 Kappa 上表现更优。这表明,在空间不相交样本下,消除模式在 KDFMGF 算法中的有效性进一步得到验证。

5 总 结

为了降低模型复杂度的同时,解决模型的误分类问题,本文创新性提出了一种专为高光谱图像分类算法设计的知识蒸馏方法 KDFMGF 和两种模式:混合模式和消除模式。针对教师模型是否解决光谱变异误分类问题来灵活调整蒸馏方式。实验结果表明,在两种模式下蒸馏算法均优于现行的八种先进蒸馏算法,蒸馏得到的学生模型性能均超越教师模型。特别在累积模式下,蒸馏得到的学生模型分类性能超越十种先进高光谱图像分类算法。此外,针对融合模块的两种模式和两种分数阶微分几何特征损失模块,均设计了消融实验,验证这些模块的有效性,进一步证明了 KDFMGF 的优越性。

KDFMGF 算法虽然在简化模型的同时还提升了模型的性能,但 KDFMGF 仍然存在改进空间,由于学生模型保留了包含大量卷积操作的 CSS-Conv 和 SIC-Conv 等核心模块,即使采用了深度可分离卷积技术,但学生模型的计算成本下降幅度仍然较低,因此在后续的工作中,在保证模型分类性能的前提下,考虑进一步简化学生模型,例如尝试采用将动态修剪的迭代特性与类似于静态修剪的预定义方面相结合的级联剪枝,对学生模型进行剪枝操作,或者探索更好的量化感知训练(QAT)策略或者是训练后量化(PTQ)策略^[48],进一步轻量化学生模型。

参 考 文 献

- [1] Xu Y, Du B, Zhang L. Beyond the patchwise classification: Spectral-spatial fully convolutional networks for hyperspectral image classification. *IEEE Transactions on Big Data*, 2020, 6(3): 492-506
- [2] Li C, Rasti B, Tang X, et al. Channel-layer-oriented lightweight spectral-spatial network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, 62: 1-14
- [3] Niedzielko J, Kopeć D, Wylazłowska J, et al. Airborne data and machine learning for urban tree species mapping: Enhancing the legend design to improve the map applicability for city greenery management. *International Journal of Applied Earth Observation and*
- [4] Chen Y, Zhou S, Huang Y. Attention network with fractional-ordered mixed geometric features for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, 62: 1-15
- [5] Waheed Z, Khalid S, Riaz SM, et al. Resource-restricted environments based memory-efficient compressed convolutional neural network model for image-level object classification. *IEEE Access*, 2023, 11: 1386-1406
- [6] Yang D, Luo Z. A parallel processing CNN accelerator on embedded devices based on optimized MobileNet. *IEEE Internet of Things Journal*, 2023, 10(21): 18844-18852
- [7] Gu Q, Luan H, Huang K, et al. Hyperspectral image classification using multi-scale lightweight transformer. *Electronics*, 2024, 13(5): 949
- [8] Chen L, Wei Z, Xu Y. A lightweight spectral-spatial feature extraction and fusion network for hyperspectral image classification. *Remote Sensing*, 2020, 12(9): 1395
- [9] Bai Y, Xu M, Zhang L, et al. Pruning multi-scale multi-branch network for small-sample hyperspectral image classification. *Electronics*, 2023, 12(3): 674
- [10] Lei Y, Wang D, Yang S, et al. Network collaborative pruning method for hyperspectral image classification based on evolutionary multi-task optimization. *Remote Sensing*, 2023, 15(12): 3084
- [11] An S, Shin J, Kim J. Quantization-aware training with dynamic and static pruning. *IEEE Access*, 2025, 13: 57476-57484
- [12] Neeralgi A A, Avarsekar R R, Bhajantri M, et al. Knowledge distillation using deep learning techniques: A survey//*Proceedings of the 2024 IEEE Conference on Engineering Informatics (ICEI)*. Melbourne, Australia, 2024: 1-10
- [13] Wang H, Dong M, Zhu G, et al. Decoupled classifier knowledge distillation. *PLOS ONE*, 2025, 20(2): e0314267
- [14] Shi C, Fang L, Lv Z, et al. Explainable scale distillation for hyperspectral image classification. *Pattern Recognition*, 2022, 122: 108316
- [15] Xie W, Zhang Z, Jiao L, et al. Decoupled knowledge distillation via spatial feature blurring for hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024, 17: 8938-8955
- [16] Yang Z, Cao Y, Zhou X, et al. Random shuffling data for hyperspectral image classification with siamese and knowledge distillation network. *Remote Sensing*, 2023, 15(16): 4078
- [17] Wu H, Xue Z, Zhou S, et al. Overcoming granularity mismatch in knowledge distillation for few-shot hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2025, 63: 1-17
- [18] Wang F, Pan C, Huang J. Application of model compression technology based on knowledge distillation in convolutional neural network lightweight//*Proceedings of the 2022 China Automation Congress (CAC)*. 2022: 4305-4310
- [19] Passalis N, Tefas A. Learning deep representations with probabilistic knowledge transfer//*Proceedings of the European conference on computer vision (ECCV)*. 2018: 268-284
- [20] Sepahvand M, Abdali-Mohammadi F, Taherkordi A. Teacher-student knowledge distillation based on decomposed deep feature representation for intelligent mobile applications. *Expert Systems with*

- Applications, 2022, 202: 117474
- [21] Liu Y, Cao J, Li B, et al. Cross-architecture knowledge distillation. *International Journal of Computer Vision*, 2024, 132(8): 2798-2824
- [22] Long Z, Ma F, Sun B, et al. Diversified branch fusion for self-knowledge distillation. *Information Fusion*, 2023, 90: 12-22
- [23] Wang D, Yuan Z, Ouyang W, et al. Adversarial learning based intermediate feature refinement for semantic segmentation. *Applied Intelligence*, 2023, 53(12): 14775-14791
- [24] Guermazi E, Mdhaffar A, Jmaiel M, et al. MulKD: Multi-layer knowledge distillation via collaborative learning. *Engineering Applications of Artificial Intelligence*, 2024, 133: 108170
- [25] Wang Y, Dai T, Chen B, et al. Attribute structured knowledge distillation//Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN). Shenzhen, China, 2021: 1-8
- [26] Tian Y, Xu S, Li M. Class-view graph knowledge distillation: A new idea for learning MLPs on graphs. *Neurocomputing*, 2025, 637: 130035
- [27] Yu T, Zhao X, An Y, et al. Knowledge distillation dealing with sample-wise long-tail problem//Proceedings of the Asian Conference on Computer Vision. Singapore: Springer Nature Singapore, 2025: 411-427
- [28] Ortigueira M D. Riesz potential operators and inverses via fractional centre derivatives. *International Journal of Mathematics and Mathematical Sciences*, 2006, 2006(1): 048391
- [29] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. *arXiv, arXiv preprint arXiv:1503.02531*, 2015
- [30] Zagoruyko S, Komodakis N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer//Proceedings the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, arXiv preprint arXiv:1612.03928, 2016
- [31] Park W, Kim D, Lu Y, et al. Relational knowledge distillation//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR). Long Beach, USA, 2019: 3962-3971
- [32] Peng B, Jin X, Li D, et al. Correlation congruence for knowledge distillation//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Republic of Korea, 2019: 5006-5015
- [33] Liu L, Huang Q, Lin S, et al. Exploring inter-channel correlation for diversity-preserved knowledge distillation//Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV). Montreal, Canada, 2021: 8251-8260
- [34] Huang T, You S, Wang F, et al. Knowledge distillation from a stronger teacher//Advances in Neural Information Processing Systems (NIPS). 2022,35: 33716-33727
- [35] Sun S, Ren W, Li J, et al. Logit standardization in knowledge distillation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2024: 15731-15740
- [36] Miles R, Mikolajczyk K. Understanding the role of the projector in knowledge distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, 38(5): 4233-4241
- [37] Melgani F, Bruzzone L. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, 2004, 42(8): 1778-1790
- [38] Han Z, Yang J, Gao L, et al. Dual-branch subpixel-guided network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, 62: 1-13
- [39] Lee H, Kwon H. Going deeper with contextual CNN for hyperspectral image classification. *IEEE Transactions on Image Processing*, 2017, 26(10): 4843-4855
- [40] Zhu M, Jiao L, Liu F, et al. Residual spectral-spatial attention network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2021, 59(1): 449-462
- [41] Zhong Z, Li Y, Ma L, et al. Spectral-spatial transformer network for hyperspectral image classification: A factorized architecture search framework. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 1-15
- [42] Sun H, Zheng X, Lu X, et al. Spectral-spatial attention network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2020, 58(5): 3232-3245
- [43] Zhang X, Sun G, Jia X, et al. Spectral-spatial self-attention networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 1-15
- [44] Hang R, Li Z, Liu Q, et al. Hyperspectral image classification with attention-aided CNNs. *IEEE Transactions on Geoscience and Remote Sensing*, 2021, 59(3): 2281-2293
- [45] Roy SK, Manna S, Song T, et al. Attention-based adaptive spectral-spatial kernel ResNet for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2021, 59(9): 7831-7843
- [46] Li M, Liu Y, Xue G, et al. Exploring the relationship between center and neighborhoods: Central vector oriented self-similarity network for hyperspectral image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023, 33(4): 1979-1993
- [47] Feng S, Zhang H, Xi B, et al. Cross-domain few-shot learning based on decoupled knowledge distillation for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, 62: 1-14
- [48] Saha S, Xu L. Vision transformers on the edge: A comprehensive survey of model compression and acceleration strategies. *Neurocomputing*, 2025, 643: 130417



LIU Ya-Wen, M.S. candidate. Her research interests include computer vision, hyperspectral image classification, and remote sensing image processing.

ZHOU Shang-Bo, Ph.D., professor, Ph.D. supervisor. His research interests include artificial neural networks, physical engineering simulation, visual object tracking, and nonlinear dynamical systems.

CHEN Yi-Jia, M.S. candidate. Her research interests include computer vision, deep learning, and hyperspectral image classification.

Background

In the field of hyperspectral image classification, CNN based classification models have achieved excellent classification results while optimizing the misclassification problem caused by spectral variations. However, while the hyperspectral image classification model achieves excellent classification results, the parameter count and computational cost of the model continue to increase, limiting its deployment on resource constrained devices. Therefore, the knowledge distillation method is chosen to lightweight the model. At present, the model that solves the problem of misclassification caused by spectral variation has high computational complexity, and the model after knowledge distillation has not solved the misclassification problem. To address this problem, we explore a knowledge distillation algorithm that can solve the misclassification problem during the distillation process.

This paper proposes KDFMGF, a knowledge distillation algorithm tailored for hyperspectral image classification. It introduces the BFDGFC module to locate spectral variation pixels and identify those responsible for misclassification. The SLFDGF module captures teacher knowledge related to spectral variation classification, enabling targeted knowledge transfer.

Meanwhile, the DLFDGF module guides label learning using spectral variation features for more accurate model training. Additionally, the method incorporates a dual-mode fusion strategy: cumulative mode is designed for well-performing teacher models, allowing students to fully absorb error-correction knowledge, while elimination mode is used when teachers underperform, preventing students from learning incorrect knowledge and helping them focus on accurate classification of spectral variation pixels based on labels.

The experimental results show that the distillation algorithm proposed in this paper has the best performance compared to the other 8 advanced knowledge distillation algorithms. At the same time, in cumulative mode, the student model distilled by the teacher model (based on fractional order mixed geometric features attention network) using this method not only outperforms the teacher model, but also surpasses 10 advanced hyperspectral image classification algorithms.

This project has received support from the National Natural Science Foundation of China (Research on Guided Collaborative Deep Learning Model Based on Differential Features, 6227072538).