文章编号　1004-924X(2024)07-1087-14

CNN-Transformer结合对比学习的高光谱与LiDAR数据协同分类

吴海滨[1]，戴诗语[1]，王爱丽[1*]，岩堀祐之[2]，于效宇[3]

(1. 哈尔滨理工大学 测控技术与通信工程学院 黑龙江省激光光谱技术及应用重点实验室，黑龙江 哈尔滨 150080；
2. 中部大学 计算机科学学院，日本 爱知 487-8501；
3. 电子科技大学 中山学院 电子信息学院，广东 中山 528400)

摘要：针对高光谱图像(hyperspectral images，HSI)与LiDAR数据多模态分类任务中的跨模态信息表达和特征对齐等问题，提出一种基于对比学习CNN-Transformer高光谱和LiDAR数据协同分类网络(Contrastive Learning based CNN-Transformer Network，CLCT-Net)。CLCT-Net通过由ConvNeXt V2 Block构成的共有特征提取模块，获得不同模态间的共性特征，解决异构传感器数据之间语义对齐的问题。构建了包含空间-通道分支和光谱上下文分支的双分支HSI编码器，以及结合频域自注意力机制的LiDAR编码器，以获取更丰富的特征表示。利用集成对比学习进行分类，进一步提升多模态数据协同分类的精度。在Houston 2013和Trento数据集上的实验结果表明，相较于其他高光谱图像和LiDAR数据分类模型，本文所提模型获得了更高的地物分类精度，分别达到了92.01%和98.90%，实现了跨模态数据特征的深度挖掘和协同提取。

关　键　词：高光谱图像；激光雷达数据；Transformer；卷积神经网络；对比学习

中图分类号：TP394.1；TH691.9　　**文献标识码**：A　　**doi**：10.37188/OPE.20243207.1087

Collaborative classification of hyperspectral and LiDAR data based on CNN-transformer

WU Haibin[1]，DAI Shiyu[1]，WANG Aili[1*]，YUJI Iwahori[2]，YU Xiaoyu[3]

(1. *Heilongjiang Province Key Laboratory of Laser Spectroscopy Technology and Application，College of Measurement and Control Technology and Communication Engineering，Harbin University of Science and Technology，Harbin* 150080，*China*；
2. *Department of Computer Science，Chubu University，Aichi* 487-8501，*Japan*；
3. *College of Electron and Information，University of Electronic Science and Technology of China，Zhongshan Institute，Zhongshan* 528400，*China*)
* *Corresponding author，E-mail：aili*925*@hrbust. edu. cn*

收稿日期：2023-10-23；**修订日期**：2023-11-20.

基金项目：黑龙江省自然科学基金资助项目(No. JJ2023LH1143)；黑龙江省重点研发计划资助项目(No. JD2023SJ19)；“一带一路”创新人才交流外国专家项目(No. G2022012010L)；黑龙江省级领军人才梯队后备带头人资助项目

Abstract: To tackle the challenges in multimodal classification tasks involving hyperspectral images (HSI) and LiDAR data, such as cross-modal information expression and feature alignment, this paper introduces a contrastive learning-based multi-branch CNN-Transformer network (CLCT-Net) for the joint classification of hyperspectral and LiDAR data. Initially, CLCT-Net employs a feature extraction module with a ConvNeXt V2 Block to capture shared features across different modalities, addressing the semantic alignment issue between data from heterogeneous sensors. It then develops a dual-branch HSI encoder with spatial channel and spectral context branches, alongside a LiDAR encoder enhanced by a frequency domain self-attention mechanism, to secure more comprehensive feature representations. Lastly, it leverages ensemble contrastive learning for classification to further refine the accuracy of multimodal collaborative classification. Experimental evaluations on the Houston 2013 and Trento datasets demonstrate that the proposed model excels in extracting and integrating cross-modal data features, achieving superior ground object classification accuracies of 92.01% and 98.90%, respectively, when compared to existing models for classifying hyperspectral images and LiDAR data.

Key words: hyperspectral image; LiDAR data; transformer; convolutional neural network; contrastive learning

1 引　言

高光谱图像由同一区域数百个连续波段的光谱组成,具有光谱分辨率高、“图谱合一”的独特优势,其丰富的光谱信息可以用于识别不同地物的组成材质与内在结构[1-2]。近年来,深度学习端到端的特征学习框架,如卷积神经网络(Convolutional Neural Network, CNN)[3-4],3D CNN[5]等深度模型,能自动学习图像中的复杂特征表示,为高光谱图像分类提供了新的方法路径。

激光雷达(LiDAR)可以生成数字表面模型(Digital Surface Model, DSM),反映地表目标的三维立体信息[6-7],地物在高程形态特征上的差异被广泛应用于分类任务中。高光谱图像和LiDAR数据作为两种不同的遥感模态,存在明显的异质性。充分利用两种数据之间的互补信息,提取更丰富的特征表达是当前限制异构遥感数据深层次协同的关键难题之一。

相较于使用单传感器数据源,协同后的高光谱和LiDAR数据集成了光谱特征、空间结构以及高程信息,能够从更多维度全面描述地物。具体来说,高光谱数据提供细致的光谱信息,在识别和表达地物光谱差异性方面具有明显优势;而LiDAR数据提供高精度的空间分辨率和高程信息,能够准确反映地物的空间分布特征。两种数据源在表征地物方面呈现互补性,深层次协同可以增强地物类别的可分离性,进而提高分类的准确率。

ConvNeXt[8-9], ViT[10], DaViT[11]和SpectFormer[12]等网络架构,通过自注意力机制和全局上下文信息的建模,能够更好地捕捉图像的关键特征,实现更准确的视觉推理和分析。基于深度学习的CNN和Transformer模型,也被引入遥感图像的多源数据协同分类任务,取得了令人满意的协同分类效果[13-18]。例如,采用形态学扩展的属性剖面[13]、IP-CNN[14],Transformer[15]、双分支卷积网络(Two-Branch CNN)[16]、深度编码器-解码器网络(EndNet)[17]、多源特征中间层融合网络(MDL-Middle)[18]和多注意力分层稠密融合网络(MAHiDFNet)[19]。

对比学习作为一种自监督表示学习方法,可以学习到具有强大区分能力的特征表达,在多模态领域得到了广泛的应用[20]。通过使用同一样本在不同视角下的代理任务进行训练,对比学习能够获得具有语义对齐的特征表达[21-22]。为解决异构多模态数据特征表达能力不足的问题,本文提出了基于对比学习CNN-Transformer高光谱和LiDAR数据协同分类网络(Contrastive Learning based CNN-Transformer Network, CLCT-Net),结合ConvNeXt V2 Block设计了共有特征提取网络,增强模型对异构多模态数据的表征能力,实现跨模态数据之间的特征对齐。然后,充

分发挥CNN的局部特征学习和Transformer的全局上下文建模能力，构建了包含空间-通道分支和光谱上下文分支的双分支HSI编码器，以及结合频域自注意力机制的LiDAR编码器，挖掘不同模态数据之间的互补信息。最后，利用集成对比学习进行分类，进一步推动模态特征对齐，提升多模态数据协同分类的精度。

2　原　理

2.1　CLCT-Net模型架构

图1展示了CLCT-Net模型的架构框图。该模型主要包含以下三个部分：共有特征提取网络、HSI编码器、LiDAR编码器和集成对比学习损失函数。CLCT-Net模型首先经过共有特征提取网络进行共有特征提取，共有特征提取网络由ConvNeXt V2 Block组成，通过全局响应归一化、深度可分离卷积等实现共性特征提取。提取后的共有特征分别输入HSI编码器和LiDAR数据编码器中，HSI编码器由空间-通道子分支、光谱上下文子分支组成，其中空间-通道子分支利用局部空间窗口多头双注意力（Spatial Window Multi-headed Self-attention，SW-MHSA）机制和通道组多头双注意力（Channel Group Multi-headed Self-attention，CG-MHSA）机制，LiDAR编码器利用频域注意力机制（Spectrum Former）。HSI编码器学习图像的空间结构和光谱信息，LiDAR编码器学习数据中的高程信息以及其空间结构依赖性。最后，两种模态特征通过基于集成对比学习的联合分类器，其中损失函数同时包含对比损失和分类损失，可以增强特征的判别能力，将同源数据特征距离最小化，异源数据特征距离最大化，实现高光谱图像和LiDAR数据的协同分类。

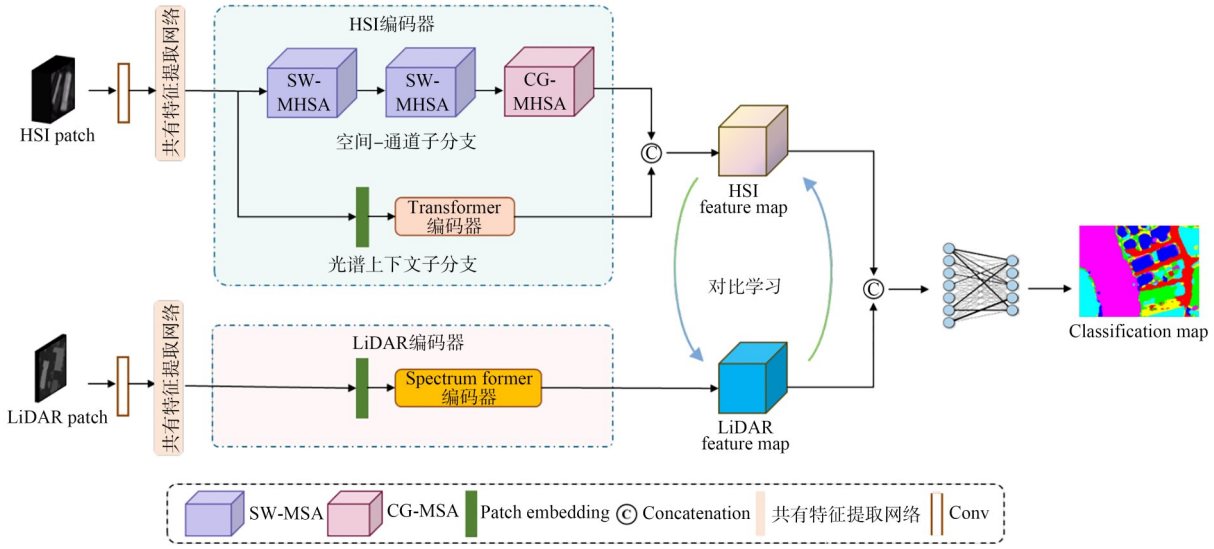


图1　CLCT-Net模型架构

Fig. 1　Model architecture of CLCT-Net

2.2　共有特征提取网络

由于异构多模态特征分布存在差异性，这给模型建模跨模态的相关信息对齐带来困难，模型难以直接学习到不同模态间隐含的联系规律。为解决异构多模态特征的对齐问题，本文设计了基于ConvNeXt V2 Block的共有特征提取网络，通过深度可分离卷积高效提取两种模态数据的低频共性特征，再通过全局响应归一化（Global Response Normalization，GRN）更好地传导共有信息，进而实现异构数据的深层协同，从而在多层抽象程度上挖掘不同模态之间的语义关联信息，实现跨模态数据之间的特征对齐，如图2所示。其中，24维7×7深度可分离卷积层（d7×7，24）用于聚合全局特征信息，深度可分离卷积将标准卷积操作分解为深度卷积和点卷积，大大减少了计算量而保持了等效的建模能力。在可分离卷积后进行层规范化（Layer Normalization，LN）操作，96维1×1卷积层进行通道数升维，GE-

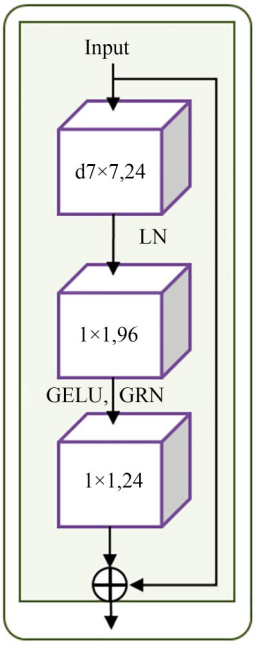


图 2　共有特征提取网络示意图

Fig. 2　Schematic diagram of shared feature extraction network

LU激活函数引入非线性，GRN层对特征进行校准，增强模型的稳定性。最后，24维1×1卷积层负责通道数降维，并通过残差连接，使网络更专注于学习两种异构模态之间的低频共有特征。其中，GRN是通过划分特征图中的相邻区域，对区域内特征响应值进行归一化处理，实现共有信息的传递，即有：

$$b_{x,y}=a_{x,y}/(\sigma_{x,y}+\epsilon), \tag{1}$$

$$\sigma_{x,y}=\left(\alpha+\beta\sum_{k=0}^{C-1}(a_{x,y}^{k})^{2}\right)^{\gamma}, \tag{2}$$

其中：$a_{x,y}^{k}$是输入特征图中位置(x,y)上第k个通道的激活值，特征图尺寸为$H\times W\times C$，H，W为高度和宽度，C表示特征图的通道数，$b_{x,y}$是特征激活值归一化后的结果，$\sigma_{x,y}^{k}$是输入特征图中位置(x,y)上第k个通道的方差，α，β，γ是可学习的参数，用于调节归一化的程度，ϵ是一个很小的数，用于维持数值稳定性。

相比直接串联原始特征或分别单独训练，共有特征提取网络具有更少的参数量、更加紧凑的特征表达能力。因原始HSI图像中独立的光谱通道存在一定程度的冗余性，共有特征提取网络能够对HSI光谱通道信息进行整合，以减少冗余并增强光谱通道间的关联性。

2.3　双分支HSI编码器

在高光谱图像中，每个像素都包含多个光谱波段的信息，波段之间存在复杂的空间和光谱关联。因此，本文设计了基于Transformer的双分支HSI编码器（Two Branch HSI Encoder，TB-HSI），由空间-通道子分支和光谱上下文子分支组成，如图3所示。其中，空间-通道子分支专注建模局部光谱-空间依赖，而光谱上下文子分支聚焦于全局光谱特征挖掘。相比单一分支结构，该设计可以同时捕获局部光谱-空间特征和全局光谱语义信息，全面提高了模型对HSI特征的理解和表达能力。

2.3.1　空间-通道特征提取子分支

空间-通道特征提取子分支利用SW-MHSA和CG-MHSA学习高光谱图像的空间依赖关系和不同光谱通道之间的关联性，以增强模型对高光谱图像空间-通道特征的表达能力。如图3(a)所示，SW-MHSA将输入高光谱图像分割成多个局部图像块，在每个块周围定义一个空间窗口，仅计算窗口内块之间的注意力权重。在多头结构下，不同的头学习不同类型的局部空间依赖模式。SW-MHSA能够更高效建模局部空间信息，增强对空间信息的学习能力。

设输入特征矩阵为$\boldsymbol{X}\in\mathbf{R}^{N\times C}$，其中$N$为空间位置数，$C$为特征维数。对于窗口$w$，提取局部特征子集$\boldsymbol{X}_w$，$\boldsymbol{X}_w=\boldsymbol{X}_{i:i+w}$，根据线性映射得到Query，Key，Value矩阵：

$$\begin{aligned}\boldsymbol{Q}_w&=\boldsymbol{X}_w\boldsymbol{W}_w^{Q}\\ \boldsymbol{K}_w&=\boldsymbol{X}_w\boldsymbol{W}_w^{K}.\\ \boldsymbol{V}_w&=\boldsymbol{X}_w\boldsymbol{W}_w^{V}\end{aligned} \tag{3}$$

这里的W^{Q}，W^{K}，W^{V}表示线性映射的参数矩阵，将输入$\boldsymbol{X}$映射到Query，Key，Value的对角空间中。

对于每个窗口w，计算注意力分数：

$$A_w=Soft\max\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\in\mathbf{R}^{N_w\times N_w}, \tag{4}$$

其中d表示线性映射的参数矩阵的第二维，也就是映射后的特征维度。计算窗口内Value加权和：

$$\boldsymbol{O}_w=\boldsymbol{A}_w\boldsymbol{V}_w\in\mathbf{R}^{N_w\times d}. \tag{5}$$

最后，串联所有窗口输出得到最终的多头自注意力输出。

CG-MHSA将输入特征的通道分成多个组，在每个通道组内计算自注意力，学习同组内通道

之间的依赖关系。设输入特征 $\boldsymbol{X}\in\mathbf{R}^{N\times C\times H\times W}$，其中 N 为 batch size，C 为通道数，H，W 为高度和宽度。将 $\boldsymbol{X}$ 重塑为 $\boldsymbol{X}\in\mathbf{R}^{N\times C'\times M}$，其中 $C'=\dfrac{C}{g}$，$M=H\times W$，g 为组数。在通道组维度上计算注意力分数，串联所有通道组输出得到最终多头自注意力输出。相比全通道的注意力计算，CG-MHSA更高效并可捕捉光谱之间的关联性，增强对光谱信息的建模能力。

2.3.2　光谱上下文特征提取子分支

光谱上下文子分支使用Transformer编码器结构，如图3(b)所示。通过自注意力机制学习光谱维度之间的依赖，并利用编码器部分进一步充分捕捉光谱特征之间的上下文语义信息。

设高光谱图像块为 $\boldsymbol{X}\in\mathbf{R}^{H\times W\times C}$，其中 H，W 为高光谱图像块的高度和宽度，C 为光谱波段数量，提取该光谱特征矩阵中对应中心像素的 C 维特征向量作为光谱上下文子分支的输入，进行线性投影生成Query，Key，Value矩阵。通过多头自注意力计算获得中心像素光谱特征的上下文表示，重复该过程进行多层编码，以学习光谱特征之间的依赖关系，获得中心像素在光谱全局视角下的上下文表示。

2.4　结合自注意力机制的LiDAR编码器

LiDAR数据有丰富的建筑物边界、植被形状等高程信息，充分学习LiDAR数据的高程特征，能够极大提升协同分类性能。因此，本文设计基于频域自注意力机制的LiDAR编码器(Spectrum LiDAR Encoder，Spectrum-LiDAR)，该编码器使用Transformer编码器结构，采用基于傅里叶变换的自注意力机制，学习LiDAR的全局依赖关系，聚焦高程信息。

如图4所示，设LiDAR数据经过共有特征提取网络获得空间域特征为 $z(x,y)$，进行二维离散傅里叶变换得到其频域表达 $Z(u,v)$，即：

$$Z(u,v)=F[z(x,y)]=\sum_{x}\sum_{y}z(x,y)\mathrm{e}^{-2\pi\mathrm{j}(ux+vy)},\tag{6}$$

其中 u 和 v 是频率域的变量。

随后，定义频域滤波器 $W_c(u,v)$，与 $Z(u,v)$

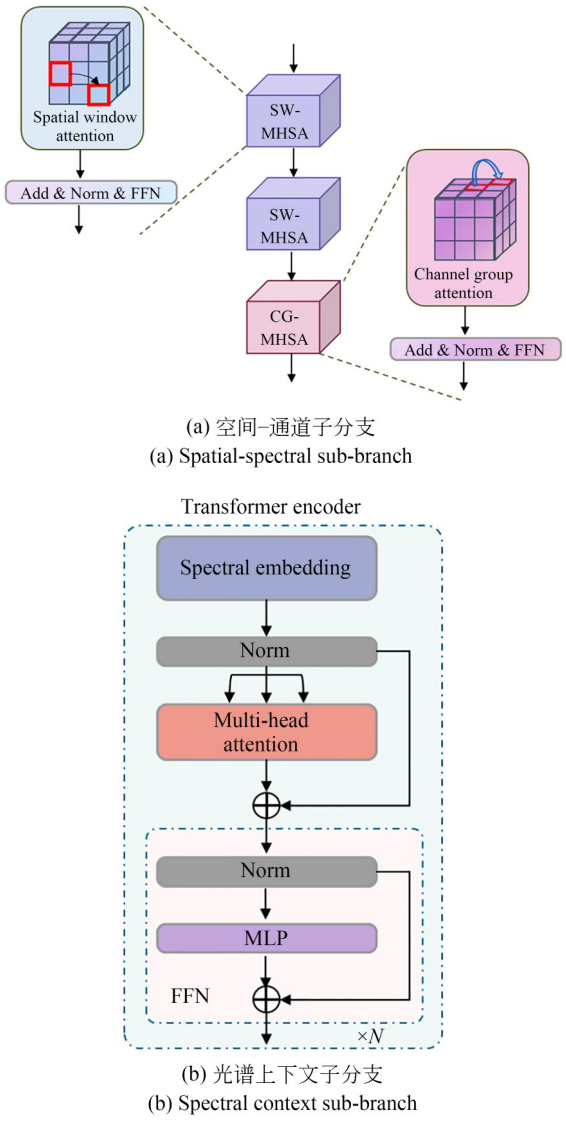


图3　HSI编码器示意图

Fig. 3　Schematic diagram of HSI encoder

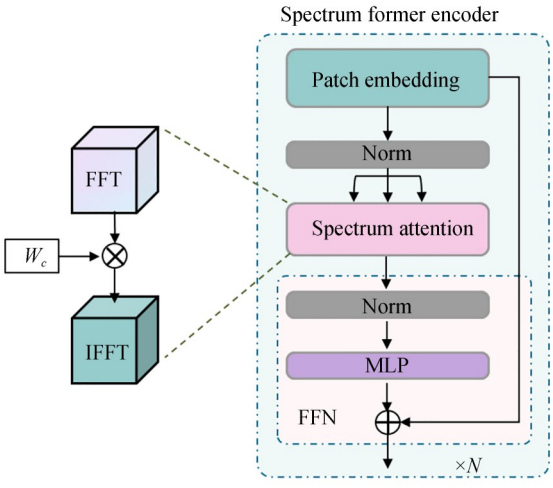


图4　LiDAR编码器示意图

Fig. 4　Schematic diagram of LiDAR encoder

进行逐点乘法，得到加权后的频域函数：

$$Z'(u,v)=Z(u,v)W_c(u,v). \tag{7}$$

最后，对$Z'(u,v)$进行反傅里叶变换，以取得空间域的输出函数：

$$F^{-1}[Z'(u,v)]=\sum_u\sum_v Z'(u,v)\mathrm{e}^{2\pi\mathrm{j}(ux+vy)}$$
$$z'(x,y)=F^{-1}[Z'(u,v)] \tag{8}$$

$z'(x,y)$反映了LiDAR数据在不同频率下的特征分布，能够捕获到不同频率下丰富的高程信息。

2.5 集成对比学习的损失函数

为实现更加有效的异构多模态特征对齐与模态协同性能，本文构建了包含对比学习损失和分类损失的联合损失函数。对比学习损失通过拉近同类异构特征之间的距离，着重跨模态数据中的共性信息，为异构数据协同分类提供更统一可靠的特征表示，以提升模型分类性能。

对比学习损失的具体计算方法如下：首先，数据集$D=(x_i^{\mathrm{HSI}},x_i^{\mathrm{LiDAR}})_{i=1}^N$，其中$x_i^{\mathrm{HSI}}$表示第$i$个样本的HSI数据，$x_i^{\mathrm{LiDAR}}$表示第$i$个样本的LiDAR数据。

对比损失函数由两部分构成：HSI对LiDAR的特征对比损失函数，以及LiDAR对HSI的特征对比损失函数。第i个样本的对比损失函数如下：

$$l_{\mathrm{contrast}}=l_i^{(\mathrm{HSI}\to\mathrm{LiDAR})}+l_i^{(\mathrm{LiDAR}\to\mathrm{HSI})}. \tag{9}$$

HSI对LiDAR，LiDAR对HSI的对比损失函数的计算公式如下：

$$\begin{aligned} l_i^{(\mathrm{HSI}\to\mathrm{LiDAR})}&=-\log\left(\frac{\mathrm{e}^{\frac{s_{i,i}^{h2l}}{\tau}}}{\sum_{j=1}^N\mathrm{e}^{\frac{s_{i,j}^{h2l}}{\tau}}}\right)\\ l_i^{(\mathrm{LiDAR}\to\mathrm{HSI})}&=-\log\left(\frac{\mathrm{e}^{\frac{s_{i,i}^{l2h}}{\tau}}}{\sum_{j=1}^N\mathrm{e}^{\frac{s_{i,j}^{l2h}}{\tau}}}\right)\end{aligned}, \tag{10}$$

其中：

$$s_{i,j}^{h2l}=f_{\mathrm{HSI}}(x_i^{\mathrm{HSI}})^{\mathrm{T}}\cdot f_{\mathrm{LiDAR}}(x_j^{\mathrm{LiDAR}}), \tag{11}$$

$$s_{i,j}^{l2h}=f_{\mathrm{LiDAR}}(x_i^{\mathrm{LiDAR}})^{T}\cdot f_{\mathrm{HSI}}(x_j^{\mathrm{HSI}}), \tag{12}$$

其中：$f_{\mathrm{HSI}}(\cdot)$和$f_{\mathrm{LiDAR}}(\cdot)$分别是HSI和LiDAR模态的特征提取函数；$s_{i,j}^{h2l}$，$s_{i,j}^{l2h}$表示样本对中HSI与LiDAR特征之间的相似性；$\tau\in\mathbf{R}$，表示温度参数。

总的对比损失函数通过对所有样本对的对比损失求平均得到：

$$L_{\mathrm{contrast}}=\frac{1}{N}\sum_{i=1}^N l_{\mathrm{contrast}}. \tag{13}$$

通过最小化该损失函数，可以学习到语义上对齐的HSI和LiDAR表征，从而提升两者特征的联合表示能力。

分类损失采用交叉熵损失的形式，用于度量预测类别分布与真实类别分布之间的距离。

$$L_{\mathrm{class}}=-\sum_{i=1}^N y_i\log(p_i), \tag{14}$$

其中：y_i是样本i的编码类别标签，p_i是模型预测的类别分布概率，分类损失能够优化神经网络的分类性能。

最终的损失函数为对比学习损失和分类损失的加权结合：

$$L_{\mathrm{total}}=\lambda_1 L_{\mathrm{contrast}}+\lambda_2 L_{\mathrm{class}}. \tag{15}$$

通过联合训练两种损失函数，模型既学习了判别性的特征表示，又获得了准确的地物分类结果。

3 实验结果与分析

3.1 实验数据集

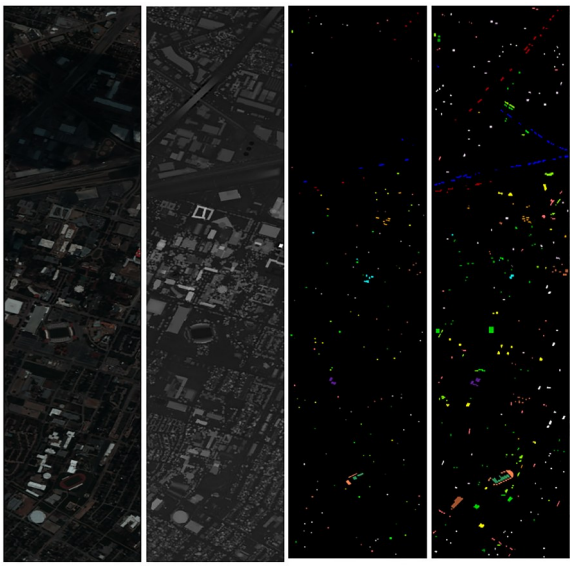
Houston2013数据集由美国国家科学基金会资助的空中激光雷达制图中心（NCALM）在2013年获取，覆盖休斯顿大学校园及周边城市区域。高光谱和LiDAR DSM数据都包含349×1 905个像素，具有相同的空间分辨率（2.5 m）。高光谱图像包含144个光谱波段，波段为380～1 050 nm，包含15个类别。表1列出了不同类别的样本数量及对应的颜色，图5给出了Houston2013数据集的可视化结果，可在IEEE GRSS网站（http://dase.grss-ieee.org/）上获得。

Trento数据集中高光谱图像由AISA Eagle传感器获取，LiDAR DSM利用Optech ALTM 3100EA传感器的第一和最后一个点云脉冲生成，两者均为600×166像素，空间分辨率均为1 m。高光谱图像包含63个波段，覆盖402.89～989.09 nm，包含6个类别。表2列出了不同类别的样本数量以及对应的颜色，图6给出了Trento数据集的伪彩色图和真值图。

表 1　Houston2013数据集土地类别详情

Tab. 1　Land class details in Houston2013 dataset

Class name	Train num	Test num	Color
Healthy grass	198	1 053	
Stressed grass	190	1 064	
Synthetic grass	192	505	
Trees	188	1 056	
Soil	186	1 056	
Water	182	143	
Residential	196	1 072	
Commercial	191	1 053	
Road	193	1 059	
Highway	191	1 036	
Railway	181	1 054	
Parking Lot1	192	1 041	
Parking Lot2	184	285	
Tennis court	181	247	
Running track	187	473	
Total	2 832	12 197	



(a) HSI的伪彩色图像 (a) Pseudo color map of HSI　(b) LiDAR的DSM (b) DSM of LiDAR　(c) 训练样本分布 (c) Training samples map　(d) 测试样本分布 (d) Testing samples map

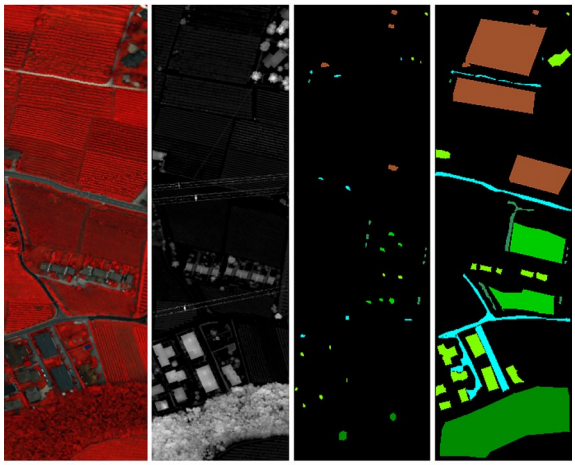
图 5　Houston2013 数据集的伪彩色图和真值图

Fig. 5　Pseudo color map and ground-truth map of Houston2013 dataset

表 2　Trento数据集土地类别详情

Tab. 2　Land class details in Trento dataset

Class name	Train num	Test num	Color
Apples	129	3 905	
Buildings	15	2 778	
Ground	105	374	
Woods	154	8 969	
Wineyard	184	10 317	
Roads	122	3 052	
Total	819	29 395	



(a) HSI的伪彩色图像 (a) Pseudo color map of HSI　(b) LiDAR的DSM (b) DSM of LiDAR　(c) 训练样本分布 (c) Training samples map　(d) 测试样本分布 (d) Testing samples map

图 6　Trento 数据集的伪彩色图和真值图

Fig. 6　Pseudo color map and ground-truth map of Trento dataset

3.2　实验平台及参数

实验基于Ubuntu 18.04系统，使用配备Tesla P100 GPU与Intel(R) Xeon(R) CPU E5-2640 v4 @ 2.40 GHZ处理器的计算服务器，Python3.7语言及PyTorch 1.10深度学习框架构建实验环境，模型训练使用的batch size为64，epoch为200，随机划分训练集和验证集，训练集和验证集的划分比例为8∶2，采用AdamW优化器、cosine学习率调度策略，初始学习率设置为5×10^{-4}，权重衰减系数为1×10^{-1}。CG-MHSA中组数g设置为1，对比学习损失中超参数τ设置为0.07，最终联合损失中的比重超参数，本文设置为$\lambda_1=0.5$，$\lambda_2=1.0$。

3.3　实验对比及分析

3.3.1　*t*-SNE 分析

根据图 7 和图 8 所示的 t-SNE(t-Distributed Stochastic Neighbor Embedding)可视化结果，在 Houston 2013 和 Trento 两个数据集上仅利用 HSI 图像进行分类，不同类别的数据点分布存在明显的重叠现象。这表明仅依靠光谱信息进行分类的效果受限。另一方面，仅利用 LiDAR 数据进行分类时，数据点的分布比较散乱，这表明仅依靠空间结构信息进行分类的性能也较差，且明显不及仅使用 HSI 图像进行分类的效果。

相较而言，同时利用 HSI 图像和 LiDAR 数据进行联合分类时，不同类别的数据点能够获得更好的聚类和区分。由此表明，高光谱和 LiDAR 协同分类模型能够更有效地利用两种数据的互补信息，提高对不同地物类别的判别能力，从而获得优于单数据源的分类性能。

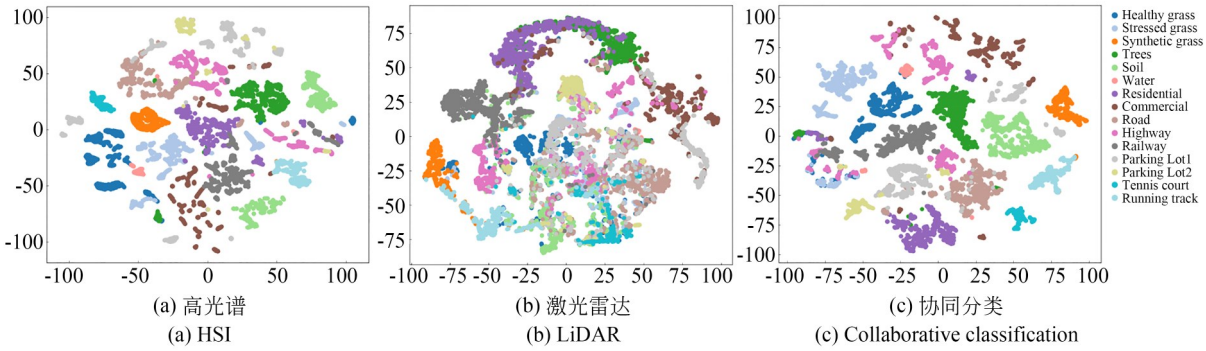


图 7　Houston2013 数据集的特征可视化

Fig. 7　Feature visualizations of Houston2013 dataset

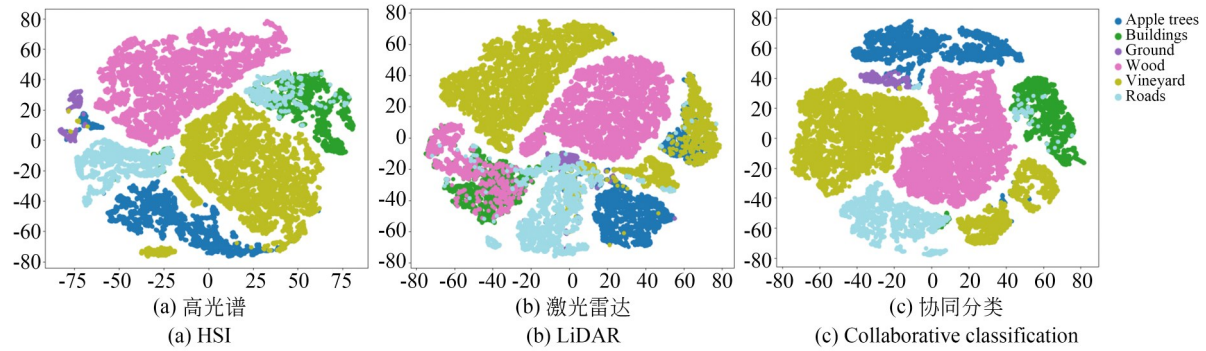


图 8　Trento 数据集的特征可视化

Fig. 8　Feature visualizations of Trento dataset

3.3.2　不同分类方法的对比

为验证 CLCT-Net 模型的联合分类的有效性，将它与 Two-Branch CNN[16]，EndNet[17]，MDL-Middle[18]和 MAHiDFNet[19]进行比较。同时，本文还比较了双分支 HSI 编码器(TB-HSI)、基于频域信息的 LiDAR 编码器(Spectrum-LiDAR)两种单传感器分类模型。

实验评价指标为整体精度(Overall Accuracy，OA)、平均精度(Average Accuracy，AA)和 Kappa 系数。OA 表示模型在所有测试样本上的正确预测样本与总样本数之间的比例。AA 是每个类别中正确预测数与该类别总数之间的比例，取各类别精度的平均值。Kappa 系数用于评估分类准确性，验证遥感分类结果图与地面真实图之间的一致性。

表 3 和表 4 给出了不同算法在 Houston2013 和 Trento 数据集上测试 15 次得到的平均分类结果。由表 3 可知，双传感器协同分类模型的分类精度明显优于单传感器分类方法，这一结论与 t-SNE 的分析结果一致。与 Two-Branch CNN，

表 3　不同方法在 Houston2013 数据集上的分类精度对比

Tab. 3　Comparison of classification accuracy of different methods on Houston2013 dataset　(%)

Class	Two-Branch	EndNet	MDL-Middle	MAHiDFNet	Spectrum-LiDAR	TB-HSI	CLCT-Net
C1	83.10	81.58	83.10	98.53	49.17	100.0	87.07
C2	84.10	83.65	85.06	92.87	35.93	98.04	**98.05**
C3	100.00	100.00	99.60	91.11	72.12	30.64	96.67
C4	93.09	93.09	91.57	98.10	65.08	99.28	94.78
C5	100.00	99.91	98.86	98.38	59.63	99.62	99.34
C6	99.30	95.10	100.00	98.58	24.54	93.38	77.14
C7	92.82	82.65	97.64	99.15	75.61	85.66	89.50
C8	82.34	81.29	88.13	80.94	75.23	92.76	83.31
C9	84.70	88.29	85.93	98.04	74.74	94.46	**94.33**
C10	65.44	89.00	74.42	72.81	62.37	88.41	92.84
C11	88.24	83.78	84.54	72.71	85.37	96.16	**95.79**
C12	89.53	90.39	95.39	76.80	50.13	84.15	86.26
C13	92.28	82.46	87.37	95.80	41.26	96.30	87.37
C14	96.76	100.00	95.14	99.53	28.46	100.00	**100.00**
C15	99.79	98.10	100.00	100.53	70.24	90.27	94.22
OA	87.98	88.52	89.55	89.58	60.81	85.22	**92.01**
AA	90.11	89.95	91.05	91.36	57.99	89.96	**91.78**
$K\times100$	86.98	87.59	87.59	88.74	57.67	84.02	**91.33**

表 4　不同方法在 Trento 数据集上的分类精度对比

Tab. 4　Comparison of classification accuracy of different methods on Trento dataset　(%)

Class	Two-Branch	EndNet	MDL-Middle	MAHiDFNet	Spectrum-LiDAR	TB-HSI	CLCT-Net
C1	99.78	88.19	99.50	99.91	74.00	99.19	99.30
C2	97.93	98.49	97.55	88.92	62.45	81.24	97.49
C3	99.93	95.19	99.10	97.53	26.00	63.46	96.45
C4	99.46	99.30	99.90	99.98	99.54	99.93	99.29
C5	98.96	91.96	99.71	99.90	98.45	97.35	99.70
C6	91.68	90.14	92.25	99.78	88.94	94.52	96.28
OA	98.36	94.17	98.73	98.59	84.94	95.42	**98.90**
AA	97.96	93.88	98.00	97.55	74.90	89.28	**98.10**
$K\times100$	97.83	92.22	98.32	98.12	80.56	93.89	**98.54**

EndNet，MDL-Middle和MAHiDFNet相比，本文提出的方法在OA，AA和Kappa系数方面都有明显改善，尤其对Stressed grass，Road，Railway和Tennis court有显著提升。其中，Stressed grass的分类精度达到了98.05%，Tennis court的分类精度为100.00%。

根据表4，在Trento数据集上，Spectrum-LiDAR分类模型的OA为84.94%，AA为74.90%，Kappa为80.56%。TB-HSI分类模型这三个指标分别为95.42%，89.28%和93.89%。联合使用双传感器进行分类时，OA提高到98.90%，AA提高到98.10%，Kappa提高到98.54%。本文方法在Roads的分类性能方面也有明显提升，达到了96.28%。

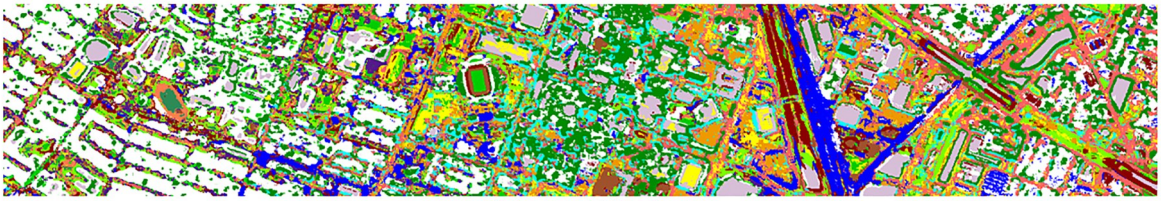
为了直观验证所提出的CLCT-Net模型的效果，在Houston 2013和Trento两个数据集上进行了分类结果的可视化对比，如图9和图10所示。本文提出的CLCT-Net能够更准确地描绘出Highway区域以及Apples区域的边缘，呈现更清晰且平滑的轮廓，其他方法获得的地物

(a) Two-Branch

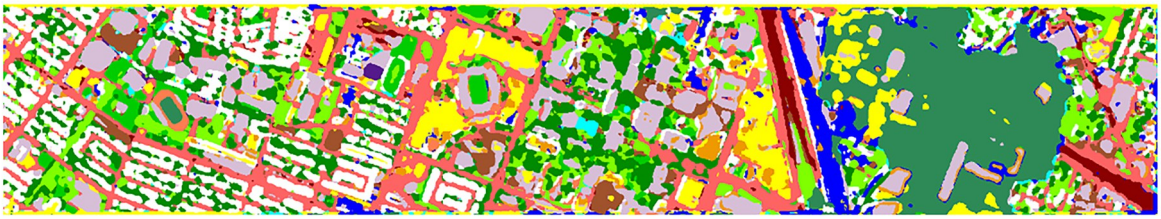
(b) MDL-Middle

(c) EndNet

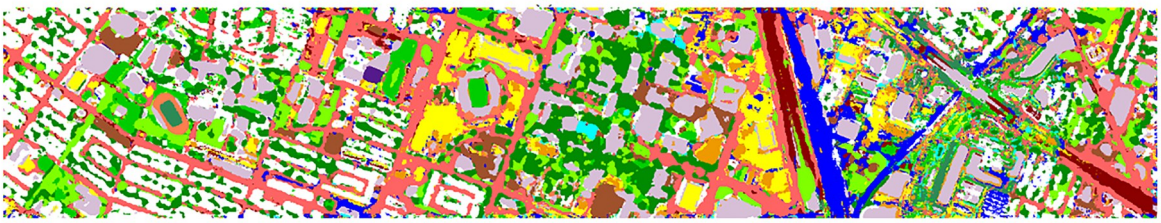
(d) MAHiDFNet



(e) Spectrum-LiDAR



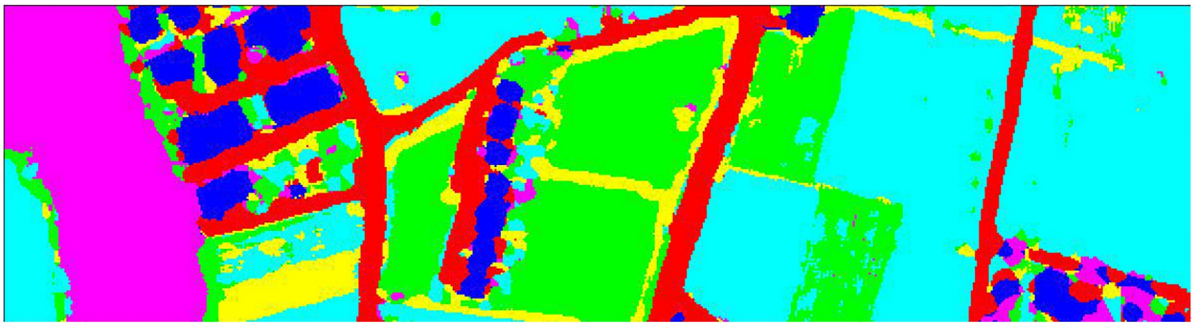
(f) TB-HSI



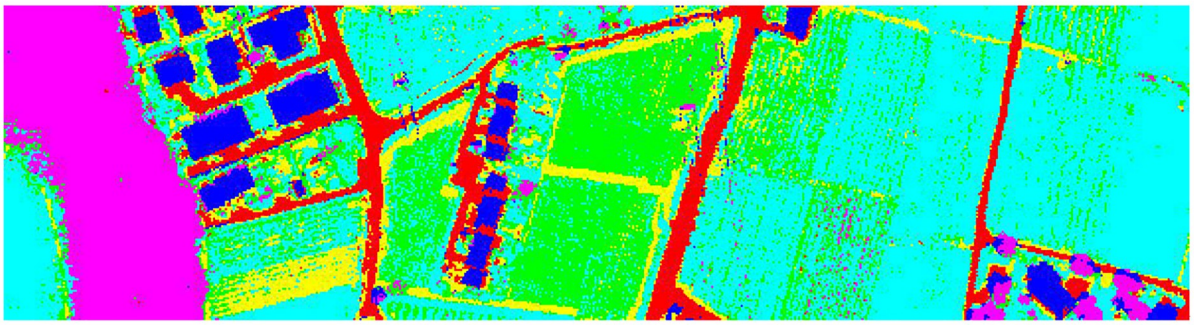
(g) CLCT-Net

图 9　不同方法在 Houston2013 数据集上的分类结果

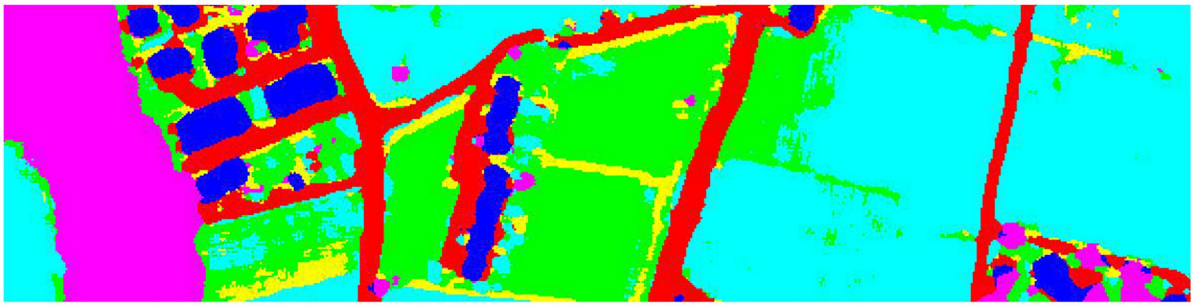
Fig. 9　Classification results of different methods on Houston2013 dataset



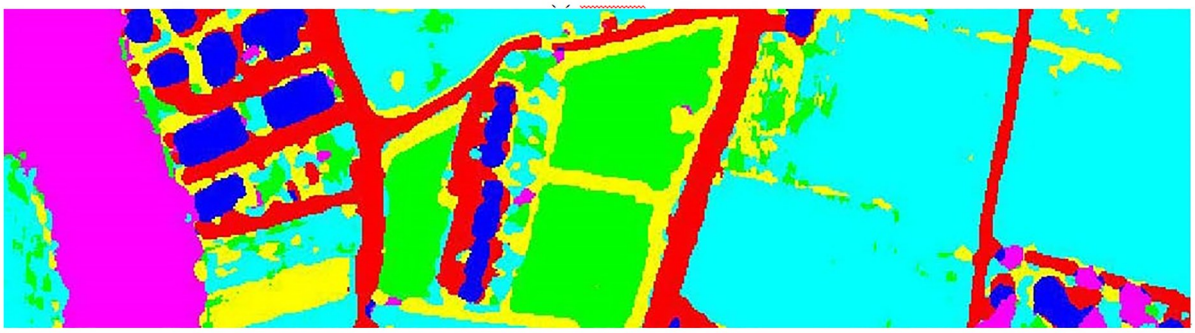
(a) Two-Branch



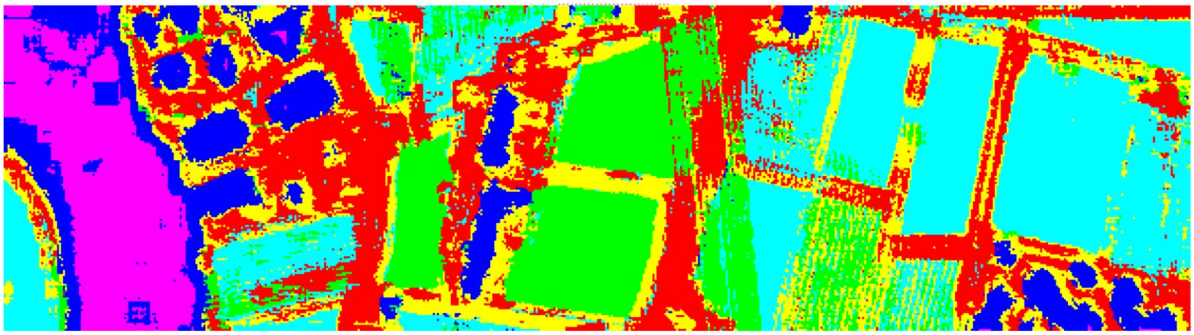
(b) MDL-Middle



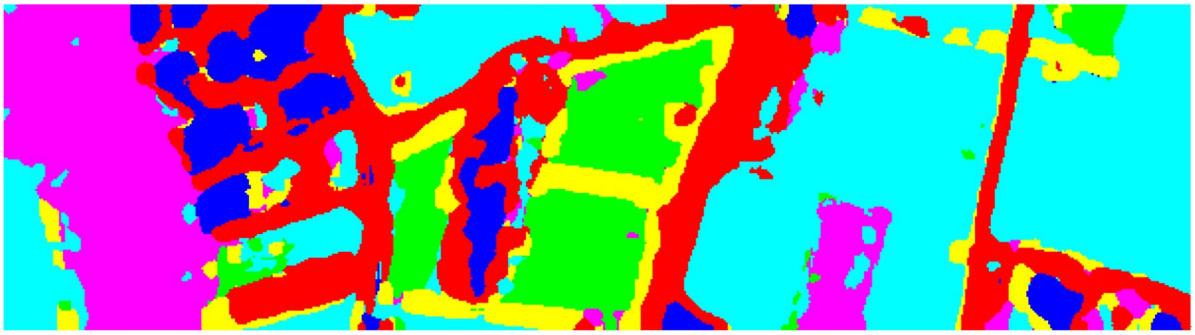
(c) EndNet



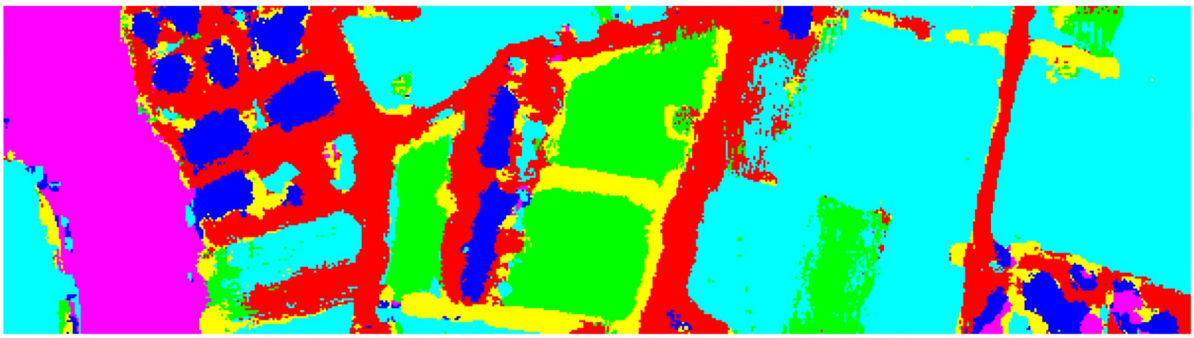
(d) MAHiDFNet



(e) Spectrum-LiDAR



(f) TB-HSI



(g) CLCT-Net

图 10　不同方法在 Trento 数据集上的分类结果

Fig. 10　Classification results of different methods on Trento dataset

边界存在明显的锯齿状边界，不够平滑。这表明 CLCT-Net 模型在细粒度特征表示和提取能力方面更为强大，能够捕捉复杂场景的微小细节，进行更精细和连贯的语义理解，在复杂边界描绘方面的表现更加出色。

3.3.3　计算复杂性分析

本文采用浮点运算数（FLOPs）和参数量（♯param）两个指标评估不同模型的计算复杂性，如表 5 所示。其中，FLOPs 表示模型处理单幅图像并完成一次前向传播所需的浮点数运算量，反映了模型的时间复杂性。♯param 表示模型的参数总量，决定了模型本身的大小，并直接影响模型在推理时所需的内存占用，反映了模型的空间复杂性。

表 5　不同分类模型的 FLOPs 和参数数量

Tab. 5　FLOPs and parameters of different classification models

Method	♯param. /M	FLOPs/M
Two-Branch	12	25
EndNet	0.07	0.49
MDL-Middle	0.25	4.7
MAHiDFNet	77	155
CLCT-Net	5.1	384

由于未考虑空间邻域信息，EndNet 模型的时间和空间复杂度相对较低。仅使用单个像素作为输入可以降低模型复杂度，忽略邻域依赖关系也会导致特征表达能力的局限，降低模型的分类准确率。对比 Two-Branch，MAHiDFNet 模型，本文提出的模型具有更为紧凑和高效的模型结构，可以在模型空间复杂度较低的情况下保持较好的性能。CLCT-Net 采用多个基于 Transformer 的编码器分支，能够更全面地提取特征。然而，由于多头自注意力机制的特性，Transformer 常需大量计算资源，这使得模型的浮点数运算量不可避免地增加。考虑到效果和复杂度综合因素，CLCT-Net 模型虽然需要较多浮点数运算，但占用的内存空间较少。这种权衡使分类准确率显著提升，达到了性能和复杂度的最佳平衡。

4　结　论

本文提出了一种基于 CNN-Transformer 的

端到端联合分类网络CLCT-Net。该网络应用共有特征提取网络模块，通过提取不同模态间的共性特征实现异构传感数据在语义级别的深层对应。其次，设计了双分支HSI编码器和频域自注意力LiDAR编码器，结合各模态特性分别学习丰富有效的特征表示。最后，引入集成对比学习策略，进一步提升了模型协同跨模态数据的地物分类能力。实验在Houston 2013和Trento数据集上进行，CLCT-Net的OA值分别为92.01%和98.90%，AA值分别为91.78%和90.10%，Kappa值分别为91.33%和98.54%，优于其他分类方法。实验结果表明，基于CNN-Transformer的框架进行异构数据联合表达和建模是地物分类任务的有效途径。

参考文献：

[1] 黄鸿，张臻，李政英. 面向高光谱影像分类的深度流形重构置信网络[J]. 光学 精密工程，2021，29(8)：1985-1998.
HUANG H，ZHANG ZH，LI ZH Y. Deep manifold reconstruction belief network for hyperspectral remote sensing image classification[J]. *Opt. Precision Eng.*，2021，29(8)：1985-1998.(in Chinese)

[2] 张兵. 高光谱图像处理与信息提取前沿[J]. 遥感学报，2016，20(5)：1062-1090.
ZHANG B. Advancement of hyperspectral image processing and information extraction[J]. *Journal of Remote Sensing*，2016，20(5)：1062-1090.(in Chinese)

[3] MA A D，FILIPPI A M，WANG Z Y，*et al*. Fast sequential feature extraction for recurrent neural network-based hyperspectral image classification[J]. *IEEE Transactions on Geoscience and Remote Sensing*，2021，59(7)：5920-5937.

[4] LIU Q C，XIAO L，YANG J X，*et al*. Content-guided convolutional neural network for hyperspectral image classification[J]. *IEEE Transactions on Geoscience and Remote Sensing*，2020，58(9)：6124-6137.

[5] ZAKARIA Z B，ISLAM M R. Hybrid 3DNet：Hyperspectral Image Classification with Spectral-spatial Dimension Reduction using 3D CNN[J]. *International Journal of Computer Applications*，2022，184(23)：6-11.

[6] 冯裴裴，杨风暴，卫红，等. 不确定性理论在激光雷达数据地物分类方面的应用研究进展[J]. 光电技术应用，2015，30(5)：1.
FENG P P，YANG F B，WEI H，*et al*. Research progress on uncertain theory for feature classification with LIDAR data[J]. *Electro-Optic Technology Application*，2015，30(5)：1.(in Chinese)

[7] 白慧，杨风暴. 基于高辨识复合衍生特征的LiDAR数据分类方法研究[J]. 激光与光电子学进展，2021，58(5)：0528001.
BAI H，YANG F B. LiDAR data classification method based on high recognition compound derivative feature[J]. *Laser & Optoelectronics Progress*，2021，58(5)：0528001.(in Chinese)

[8] LIU Z，MAO H Z，WU C Y，*et al*. A ConvNet for the 2020s[C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. *June* 18-24，2022. *New Orleans*，*LA*，*USA*. IEEE，2022：11976-11986.

[9] WOO S，DEBNATH S，HU R H，*et al*. ConvNeXt V2：co-designing and scaling ConvNets with masked autoencoders[C]. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. *June* 17-24，2023. *Vancouver*，*BC*，*Canada*. IEEE，2023：16133-16142.

[10] DOSOVITSKIY A，BEYER L，KOLESNIKOV A，*et al*. An image is worth 16x16 words：transformers for image recognition at scale[J]. *ArXiv e-Prints*，2020：arXiv：2010.11929.

[11] DING M Y，XIAO B，CODELLA N，*et al*. *DaViT*：*Dual Attention Vision Transformers*[M]. Lecture Notes in Computer Science. Cham：Springer Nature Switzerland，2022：74-92.

[12] PATRO B N，NAMBOODIRI V P，SRINIVAS AGNEESWARAN V. SpectFormer：frequency and attention is what you need in a vision transformer[J]. *ArXiv e-Prints*，2023：arXiv：2304.06446.

[13] PEDERGNANA M，MARPU P R，DALLA MURA M，*et al*. Classification of remote sensing optical and LiDAR data using extended attribute profiles[J]. *IEEE Journal of Selected Topics in Signal Processing*，2012，6(7)：856-865.

[14] ZHANG M M，LI W，TAO R，*et al*. Information fusion for classification of hyperspectral and LiDAR data using IP-CNN[J]. *IEEE Transactions on Geoscience and Remote Sensing*，2022，60：1-12.

[15] XUE Z X，TAN X，YU X C，*et al*. Deep hierar-

chical vision transformer for hyperspectral and LiDAR data classification[J]. *IEEE Transactions on Image Processing*, 2022, 31: 3095-3110.

[16] XU X D, LI W, RAN Q, *et al*. Multisource remote sensing data classification based on convolutional neural network[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2018, 56(2): 937-949.

[17] HONG D F, GAO L R, HANG R L, *et al*. Deep encoder-decoder networks for classification of hyperspectral and LiDAR data[J]. *IEEE Geoscience and Remote Sensing Letters*, 2022, 19: 1-5.

[18] HONG D F, GAO L R, YOKOYA N, *et al*. More diverse means better: multimodal deep learning meets remote-sensing imagery classification [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2021, 59(5): 4340-4354.

[19] WANG X H, FENG Y N, SONG R X, *et al*. Multi-attentive hierarchical dense fusion net for fusion classification of hyperspectral and LiDAR data [J]. *Information Fusion*, 2022, 82(C): 1-18.

[20] HE K M, FAN H Q, WU Y X, *et al*. Momentum contrast for unsupervised visual representation learning [C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. *June* 13-19, 2020. *Seattle*, *WA*, *USA*. IEEE, 2020: 9729-9738.

[21] CHEN X L, FAN H Q, GIRSHICK R B, *et al*. Improved baselines with momentum contrastive learning [J]. *arXiv preprint* arXiv: 200304297, 2020.

[22] CHEN T, KORNBLITH S, NOROUZI M, *et al*. A simple framework for contrastive learning of visual representations[C]. *Proceedings of the* 37*th International Conference on Machine Learning*. ACM, 2020: 1597-1607.

作者简介：

吴海滨(1977－)，男，上海人，博士，教授，博士生导师，2002年于哈尔滨工业大学获得硕士学位，2008年于哈尔滨理工大学获得博士学位，主要从事机器视觉、医学虚拟现实、深度学习图像分类的研究。E-mail: woo@hrbust.edu.cn

通讯作者：

王爱丽(1979－)，女，天津人，博士，教授，硕士生导师，2008年于哈尔滨工业大学获得博士学位，主要从事机器视觉、深度学习图像分类的研究。Email: aili925@hrbust.edu.cn