MASTER THESIS XIDIAN UNIVERSITY

西安电子科技大学

硕士学位论文

基于多模态特征融合的遥感图像分类研究

作者姓名 朱晨

学校导师姓名、职称 李阳阳 教授

企业导师姓名、职称 陈彦桥 高工

申请学位类别 电子信息硕士

学校代码 10701　　学　号 21171213943
分 类 号 TP75　　密　级 公开

西安电子科技大学

硕士学位论文

基于多模态特征融合的遥感图像分类研究

作者姓名：朱晨

领　　域：计算机技术

学位类别：电子信息硕士

学校导师姓名、职称：李阳阳 教授

企业导师姓名、职称：陈彦桥 高工

学　　院：人工智能学院

提交日期：2024 年 6 月

Remote Sensing Image Classification Research Based on Multimodal Feature Fusion

A thesis submitted to

XIDIAN UNIVERSITY

in partial fulfillment of the requirements

for the degree of Master

in Electronic Information

By

Zhu Chen

Supervisor: Li Yangyang Title: Professor

Supervisor: Chen Yanqiao Title: Senior Engineer

June 2024

摘要

遥感图像分类是精准识别地物类型的重要方法，助力环境监测、城市规划、资源调查等领域决策。因为单一模态遥感数据的信息量有限，传统遥感图像分类方法难以应对复杂多变的遥感场景。本文在卷积神经网络、Transformer 和自监督学习等深度学习技术的基础上，设计了多模态特征融合的遥感图像分类方法，进一步提取各模态数据特征，实现模态间特征交互融合，研究了样本受限情况下对比学习方法完成分类任务。具体研究工作如下：

（1）针对单模态数据信息量受限、卷积神经网络对光谱特征提取不充分难题，结合高光谱数据和激光雷达数据设计了一种基于改进卷积神经网络的多模态遥感图像分类方法。首先设计了双分支空谱残差网络对高光谱数据的空间和光谱特征进行提取；其次在分类任务中增加了激光雷达模态数据，并进行特征提取；最后在特征融合阶段利用孪生神经网络和可学习参数，使得特征融合时可以自适应动态调整。由于增加了激光雷达数据的高程信息和空间特征信息，光谱分支进一步提取了高光谱数据的光谱特征，孪生神经网络和可学习参数使得特征融合更合理，所提出的方法具有更高的分类性能。在 Houston2013 和 Trento 数据集上进行了实验，结果表明：在两个数据集上，所提出的方法的分类准确率对比近几年的五个优秀分类方法中最好的提升了 1.01%。

（2）针对多模态遥感数据的异构特征间难以保证一致性、模态间特征交互不充分的问题，设计了一种双向交互融合的特征融合网络。首先为了更好的提取高维度、信息复杂的高光谱图像特征引入 SpectralFormer；其次在预训练阶段加入一致性损失，通过反向传播一致性信息保证两个模态间特征的一致性；最后在特征融合阶段加入交叉注意力，促使两个模态特征充分交互，提高分类的准确率，增加模型的鲁棒性。在 Houston2013 和 Trento 数据集上进行了实验，结果表明：所提出的方法在各类别中有更好的平衡性，在 Houston2013 数据集上有 6 个类别的分类准确率达到最高，分类准确率较除研究内容一外最好的方法提升了 1.79%。

（3）针对监督方法过于依赖样本数量和质量，而多模态遥感数据样本受限、标注成本高的问题，设计了一种基于自监督对比学习的多模态遥感图像分类方法，利用对比学习进行预训练从无标签样本中学习特征表示。首先利用数据增强构建大量正负样本对帮助提高模型泛化性，使模型更容易区分不同类别样本；其次设置模态内对比学习和模态间对比学习，提升模型模态内特征不变性的同时捕获多模态数据间的而对应关系。最后延续使用上一个研究内容中效果较好的一致性损失和交叉注意力模块辅助特征融合。在 Houston2013 和 Trento 数据集上进行了实验，结果表明：所提出的方

法的准确率能够在每类仅有 10 个微调样本的时候远超监督学习的方法，在分类准确率上领先七个对比实验中效果最好的方法 2.63%。

本文得到如下基金资助：航空科学基金（No. 2022Z071081013），科技委技术领域基金（No. F022360003）。

关 键 词：遥感图像分类，多模态，Transformer，自注意力，卷积神经网络，对比学习

ABSTRACT

Remote sensing image classification is an important method for accurately identifying feature types, which helps decision-making in the fields of environmental monitoring, urban planning, and resource investigation. Because of the limited amount of information in single-modal remote sensing data, traditional remote sensing image classification methods are difficult to cope with complex and changing remote sensing scenes. In this thesis, on the basis of deep learning techniques such as convolutional neural network (CNN), Transformer and self-supervised learning, we design a remote sensing image classification method for multimodal feature fusion, further extract the features of each modal data, realize the interactive fusion of inter-modal features, and study the comparison of the learning methods to complete the classification task in the case of sample limitation. The specific research works are as follows:

(1) Aiming at the problem of limited information of unimodal data and insufficient extraction of spectral features by CNN, a multimodal remote sensing image classification method based on improved CNN is designed by combining hyperspectral data and LiDAR data. Firstly, a two-branch null spectral residual network is designed for spatial and spectral feature extraction of hyperspectral data respectively. Secondly, LiDAR modal data is added to the classification task and feature extraction is performed. Finally, Siamese neural network (SNN) and learnable factors are used in the feature fusion stage, so that the feature fusion can be adjusted adaptively and dynamically. The proposed method has higher classification performance due to the addition of elevation information and spatial feature information of the LiDAR data, the spectral branch further extracts the spectral features of the hyperspectral data, and the SNN and learnable factors make the feature fusion more reasonable. Experiments are carried out on Houston2013 and Trento datasets, and the results show that on both datasets, the classification accuracy of the proposed method improves by 1.01% over the best of the five excellent classification methods in recent years.

(2) Aiming at the problems of difficulty in ensuring consistency between heterogeneous features and insufficient inter-modal feature interaction in multimodal remote sensing data, a two-way interactive fusion feature fusion network is designed. Firstly, Spectral-Former is introduced in order to better extract high dimensional and complex hyperspectral image

features. Secondly, consistency loss is added in the pre-training stage to ensure the consistency of the features between the two modalities by back propagating the consistency information. Finally, cross-attention is added in the feature fusion stage to promote the full interaction of the two modal features to improve the accuracy of the classification and increase the robustness of the model. Experiments are conducted on Houston2013 and Trento datasets, and the results show that the proposed method has a better balance among the categories and achieves the highest classification accuracy in six categories on the Houston2013 dataset, with an improvement of 1.79% in classification accuracy over the best method except for research element one.

(3) Aiming at the problem that supervised methods rely too much on the number and quality of samples, while the samples of multimodal remote sensing data are limited and the labelling cost is high, a multimodal remote sensing image classification method based on self-supervised contrast learning is designed to learn the feature representations from unlabeled samples by using contrast learning for pre-training. Firstly, a large number of positive and negative sample pairs are constructed using data augmentation to help improve the model generalization and make the model easier to distinguish between different categories of samples. Secondly, intra-modal contrast learning and inter-modal contrast learning are set up to improve the feature invariance of the model within the modality while capturing the correspondence between the multi-modal data. Finally, we continue to use the loss of coherence and cross-attention modules, which are more effective in the previous study, to assist feature fusion. Experiments are conducted on the Houston2013 and Trento datasets, and the results show that the accuracy of the proposed method is able to far outperform the supervised learning method when there are only 10 fine-tuned samples per class, and is 2.63% ahead of the best of the seven contrastive trials in terms of classification accuracy.

This research is supported by the Aviation Science Fund (No. 2022Z071081013) and the Technology Field Fund of the Science and Technology Commission (No. F022360003).

Keywords: Remote Sensing Image Classification, Multimodal, Transformer, Self-Attention, Convolutional Neural Network, Contrastive Learning

插图索引

表格索引

符号对照表

符号	符号名称
f	特征
E_W	两组特征向量间距离
N	样本总数
k	总类别数
m	测试样本总类
Q	查询向量
K	键向量
V	值向量
d_k	维数
$\sqrt{d_k}$	比例因子
A_i	特征
x	输入数据
x^+	与 x 相似的正样本
x^-	与 x 不相似的负样本
$f(\cdot)$	编码器
$score()$	度量函数
$sim(\cdot)$	余弦相似度
τ	温度系数
$l(\cdot)$	损失函数
F_i	特征映射后的变量
$\Gamma_{[k\neq i]} \in \{0,1\}$	指标函数
$G_W(\cdot)$	特征向量

缩略语对照表

缩略语	英文全称	中文对照
AA	Average Accuracy	平均准确率
BN	Batch Normalization	批归一化
CAF	Cross-layer Adaptive Fusion	跨层自适应融合
CL	Contrastive Learning	对比学习
CAM	Cross-Attention Module	交叉注意力模块
CNN	Convolutional Neural Network	卷积神经网络
DBSSRN	Dual-branch Spatial-spectral residual network	双分支空谱残差网络
DSM	DigitalSurface Model	数字表面模型
EAPs	Extended Attribute Profiles	扩展属性特征
GAN	Generative Adversarial Network	生成对抗网络
GSE	Groupwise Spectral Embedding	分组谱嵌入
HSI	Hyperspectral Image	高光谱图像
KC	Kappa Coefficient	Kappa 系数
LSTM	Long Short-Term Memory	长短期记忆网络
LiDAR	Light Detection and Ranging	激光雷达
LFM	Learnable Factor Module	可学习参数模块
MSE	Mean Squared Error	均方误差
NLP	Natural Language Processing	自然语言处理
OA	Overall Accuracy	总体准确率
RNN	Recurrent Neural Network	循环神经网络
ResNet	Residual Network	残差网络
SAR	Synthetic Aperture Radar	合成孔径雷达
SNN	Siamese Neural Network	孪生神经网络
SNNM	Siamese Neural Network Module	孪生神经网络模块
SSL	Self-supervised Learning	自监督学习
ViT	Vision Transformer	视觉 Transformer

目录

第一章　绪论

1.1 课题研究背景及意义

遥感科学是一门交叉学科，20 世纪初，航空遥感开始发展，1972 年美国发射第一颗陆地卫星，卫星遥感开始正式步入我们的视野。遥感是一种应用探测仪器，使用时并不与探测目标相接触，而是通过从远处把目标的电磁波特性记录下来，通过分析，揭示出物体的特征性质及其变化的综合性探测技术。随着遥感技术、传感器、成像技术的不断成熟与发展，现在很容易获得大量具有复杂异质性的地球观测数据，这使得研究人员有机会以新的方式解决当前地球科学应用中的问题。遥感图像逐渐从全色遥感、多光谱遥感图像，发展到高光谱遥感图像，空间、时间、以及光谱分辨率逐渐得到提升，提高了遥感技术的对地观测能力，进而在地球资源勘探、环境监测、住宅规划、土地资源管理等领域获得广泛应用与深入发展[1]。遥感图像分类作为遥感技术应用的关键环节，旨在通过对遥感图像中各类地物的光谱信息和空间信息进行分析，选择特征，将图像中各个像元按照某种规则或算法划分成不同的类别，从而实现对实际地物的对应信息的获取[2]。单模态遥感数据在分类时无法全面反映地物的复杂性和多样性、环境受限比较明显、分类精度和鲁棒性不足，而遥感技术的最新进展提高了多传感器数据的可用性，允许对同一地理区域进行多种表示。根据传感器的特性，捕获的数据可以是同一观测区域提供的不同特性信息，将这些数据与不同的模态集成提供了独特的补充信息，使我们能够进一步获得完整的特征表示[3]。因此，遥感图像分类逐渐从单模态向多模态发展。

从 20 世纪 70 年代至今，多模态技术的研究可分成四个发展时期：人类行为多模态研究、多模态计算机处理研究、多模态互动研究和多模态深度学习研究。多模态技术的早期探索始于计算机科学与人工智能领域，目的是模仿人类处理多种感官输入的能力。后来，研究者们开始尝试将文本、图像和声音信息结合起来，实现更加丰富的数据解释和用户交互。传统的方法包括基于规则和特征工程的系统，但随着深度学习崛起[4][5]，研究重点转向了神经网络的融合和端对端学习模型。并且随着深度学习和大数据的不断发展，多模态技术通过融合例如图像、视频、文本、语音、触觉等不同模态的数据，使得模型能够捕捉更丰富的信息。模型如 Attention 机制的引入，以及 Transformer 架构的出现促进了不同模态之间更高效的信息整合。多模态技术也逐渐应用到遥感领域中，旨在克服单一遥感数据源在物体检测、识别和分类方面的困难。深度学习相关技术在多模态遥感图像分类任务中的像素级、特征级、决策级等数据融合方法上的应用，起到了举足轻重的作用，大大解决了单一模态数据的缺陷，提高了

遥感图像分类的准确率。

目前的遥感图像分类任务存在的不足之处：

（1）单一模态数据含有的特征信息量有限，都有各自的缺陷，但是模态间的数据往往拥有互补信息，如果可以通过多模态数据融合，可以很大程度的丰富特征信息，具体可参考以上描述，本文也是基于多模态特征融合开展的一系列研究。

（2）现在绝大部分多模态遥感图像分类方法都是用卷积神经网络（Convolutional Neural Network，CNN）对两种模态数据进行分类，没有针对特定数据集设计的网络去进行特征提取，导致特征提取不充分。另外 CNN 对于较为复杂的光谱和空间特征的提取能力比较有限，且难以关注到图像中的长距离依赖关系和全局特征。

（3）不能做到充分关注两个模态间的互补特征信息，虽然理论上是通过模态间的特征信息互补来提高分类准确率，但还是会存在很难充分利用到模态间特征信息的情况，如果能够充分关注到互补的特征信息和维持异构特征一致性，则可以进一步提升分类的效果。

（4）现存的多模态遥感图像数据不足，其样本量少，且标注成本高，使用监督学习的方法很容易出现过拟合情况，所以也逐渐出现少样本和小样本的遥感图像分类方法研究，另一个角度，如果使用自监督学习方法，也可以有效解决该缺陷。

本文便是基于以上的背景和意义进行的基于多模态特征融合的遥感图像分类研究。

1.2 国内外研究现状

整个遥感图像分类过程基本分为三种：监督学习、半监督学习和无监督学习方法。监督学习技术进一步分为分布式学习和统计学习[6][7][8]。分布式学习有多种类型，如逻辑回归、决策树、支持向量机、集成方法等，而统计学习技术又分为参数方法和非参数方法。这些方法的关键在于如何设计有效的特征提取算法，以从遥感图像中提取出对分类有用的信息。无监督学习技术包括 K 均值聚类、谱聚类、模糊 C 均值和强化学习等，无监督分类方法的主要挑战在于如何确定合适的聚类数量和聚类中心，以及如何处理不同类别之间的边界问题。现在基于深度学习的分类方法还可以进一步分为三类：生成方法、混合方法和判别方法。生成方法中有如深度信念网络、网络自动编码器和深度玻尔兹曼机等方法，而混合方法中有如深度神经网络和灰狼优化等方法。在判别方法中，例如 CNN[9]、循环神经网络（Recurrent Neural Networks，RNN）[10]、长短期记忆网络（Long Short-Term Memory，LSTM）[11]、生成对抗网络（Generative Adversarial Networks，GAN）[12]、人工神经网络等，都是常用的方法。此外近些年还出现迁移学习（Transfer Learning）[13]、自监督学习（Self-supervised Learning，SSL）

[14]等基于深度学习的有效分类方法。

传统方法通常需要大量的手工特征工程，即需要专家知识和经验来设计和提取有效的特征。这是一个既耗时又复杂的过程，且特征的选择和设计对分类性能有着至关重要的影响。传统方法在处理复杂和多样化的遥感数据时，其泛化能力可能受到限制。由于遥感图像的多样性和复杂性，手工设计的特征可能无法完全覆盖所有重要的信息，导致分类性能下降。此外，传统方法的适应性低，在面对新的数据类型或场景时，可能需要重新设计和调整特征提取算法及分类器，以适应新的情况。显然仅依靠人工的设计和分析是远远不够的，更多的是希望在计算机强大数据处理能力的帮助下，自动分析和处理这些大量的数据[5]。深度学习在遥感图像分类任务中的兴起和发展，通过自动特征学习、提高泛化能力、计算效率和适应性等方面的优势，克服了传统方法的主要缺陷，为遥感图像分类带来了显著的进步和突破。

基于以上背景知识及本文研究内容涉及的方法，下面着重介绍一下多模态遥感图像分类、孪生神经网络、Transformer、SSL 与对比学习的国内外研究现状。

1.2.1 多模态遥感图像分类国内外研究现状

近年来，由于不断发展的遥感技术和先进的传感器技术，已经可以通过多种遥感数据对地球进行观测，其中包含对同一地区地物特征不同模态的遥感图像数据获取，如包含丰富光谱反射信息的高光谱图像（Hyperspectral image，HSI）[15][16]，提供地面高程信息的激光雷达（Light Detection and Ranging，LiDAR）数据[17][18]，以及具有全天候提供地表结构信息的优势的合成孔径雷达（Synthetic Aperture Radar，SAR）数据[19]等。每种模态的数据都有一定的缺陷，例如 SAR 图像的分辨率相对较低，且图像解译困难。因为采用侧视相干成像方式，导致图像噪声污染较严重，系统设计复杂，处理的数据量庞大。HSI 数据的获取和处理相对复杂，数据量庞大，处理时间较长。同时，高光谱技术受到大气散射和仪器噪声等因素的干扰，可能会影响数据的质量。LiDAR 数据的获取和处理成本较高，且受到天气条件（如雨雪、大雾等）的影响较大。此外，对于某些特殊地表（如水面、光滑表面等），LiDAR 的反射信号可能较弱，导致数据质量下降。在互补性方面，SAR、HSI 和 LiDAR 图像可以相互补充，提高遥感信息提取的准确性和完整性。例如，HSI 数据可以准确地区分水和草但是，使用相同材料建造的道路和屋顶无法区分，但是，从 LiDAR 数据获得的补充高程信息对于分类目的非常有益[18]。

与单模态图像的分类相比，多模态遥感图像通过充分利用互补优势，提供更全面的信息来表示地面特征和属性，进而提升分类的准确率。Huang 等人[20]，通过 SAR 和 HSI 数据的融合，减少了恶劣天气对 HSI 数据的影响以及严重的散斑噪声对 SAR 数据的影响。其实在早期就有一些传统方法应用于HSI和LiDAR数据的联合分类了，

Pedergnana 等人[21]通过计算 HSI 和 LiDAR 数据的扩展属性特征（Extended Attribute Profiles，EAPs）来进行分类，相比传统特征，EAPs 可以显著提高分类精度和鲁棒性，特别是在复杂的地表覆盖和多变的环境条件下，EAPs 表现出了更好的适应性。此外，作者还讨论了不同分类算法在不同数据集上的性能差异，为实际应用提供了参考。Khodadadzadeh 等人[22]使用形态学方法提取 HSI 和 LiDAR 数据的多种特征进行分类，即从 LiDAR 数据中提取几何表示，结合 HSI 数据特征进行土地覆盖分类，这种方式无需引入任何正则化或权重参数，因此能够以一种协同和灵活的方式高效的运用及整合不同模态数据的特征信息。

然而，传统的基于手动特征的方法表达能力有限，而且这些高维特征可能会导致维度灾难，使其性能不能令人满意。随着深度学习和 CNN 的快速发展，科研人员们已经提出了系列基于 CNN 的多模态遥感图像分类方法。Hong 等人[23]提出了 CCR-Net，即在 CNN 中引入跨通道重建模块，通过充分利用不同模态数据的优势和 CNN 强大的特征提取能力，显著提高分类的准确性和鲁棒性，证明了利用 CNN 对多模态遥感图像进行分类是一种有效且可行的方法。Hang 等人[24]使用两个耦合的 CNN（Coupled CNNs，Co-CNNs）将 HSI 与 LiDAR 的数据进行融合。其中，一个 CNN 会从 HSI 数据中提取出光谱的空间属性，另一个从 LiDAR 数据中提取出高程特征信息。两个 CNN 均使用三个卷积层，并且，后两个卷积层会采取参数共享的方式进行连接。在融合过程中，作者采取了特性层面与决策层面的融合策略，以最大限度地融合了不同模态数据的特征，Co-CNNs 的提出不仅提高了分类性能，还为遥感数据分析和处理提供了新的视角和思路。2022 年 Wang 等人[25]提出了一种多尺度交互式网络（MIFNet）对 HSI 和 SAR 图像进行分类。作者设计了多尺度交互式信息提取模块（Multiscale Interactive Information Extraction，MIIE）来提取有意义的多尺度信息。与传统的多尺度模型相比，它不仅可以获得更丰富的尺度信息，而且减少了模型参数，降低了网络复杂度。研究人员还开发了全局依赖融合模块（Global Dependence Fusion Module，GDFM）来融合多源数据的特征，从全局角度实现多源数据之间的交叉关注并捕获长程依赖。此外还有深度分层视觉 Transformer（Deep Hierarchical Vision Transformer，DHViT）[26]、联合 CNN 和形态特征学习[27]等的陆续出现和发展。

1.2.2 孪生神经网络国内外研究现状

孪生神经网络（Siamese neural network，SNN），也被称为连体神经网络，是通过共享权重来实现两个网络的连接。在 1993 年，Jane Bromley 等人[28]首次提出了 SNN 的概念。SNN 因具有很好的可拓展性和简洁性，也被广泛应用于文本、语音和图像处理等领域[29]。

SNN 最初被应用于签名认证，即验证两个签名是否一致，其工作原理是通过比

较两张签名图像的相似度来判定它们是否属于同一个人的签名。随后，2010 年 Hinton 等人[29]使用孪生结构进行人脸识别和验证工作，将两张面孔视为一个扩展输入向量，采用连体结构，计算出某个特征向量。给定一对面孔，使用一个固定的对称函数将两个相应的特征向量组合成一个单一的表征，计算出这两个人脸是同一个人的概率。之后 SNN 在图像检索、匹配和跟踪等图像处理领域陆续得到广泛的推广和应用。2016 年，Melekhov 等人[31]在 SNN 中使用欧式距离对图像的卷积描述子匹配，使用 Siamese 结构来训练深度卷积网络，用于从图像块中提取描述符，并尝试用 CNN 代替尺度不变特征变换算法。

在 SNN 的发展历程中，其网络结构经历了不断的优化和改进。最初的 SNN 结构由两个结构完全相同的 CNN 构成，通过比较两个 CNN 的输出结果来计算输入图像之间的相似度或距离。近年来，SNN 在结合其他技术方面也取得了显著的进展。在图像检索任务中，研究者通过将改进的 CNN 模型与 SNN 相结合，实现了多模态检索。在文本匹配任务中，通过利用注意力机制、句法分析等技术，成功提升了 SNN 的性能。随着研究的深入，SNN 不但逐渐在图像领域之外的如目标跟踪等领域受到关注，其结构也逐渐引入了更多的注意力机制、残差连接等模块，进一步提升了网络的性能。2016 年，Paul 等人[32]提出了一种新的方法来衡量不同长度的字符序列之间的相似性，他们结合了字符级的双向 LSTM 和 Siamese 架构，以此为基础构建了一个网络系统，用以评估文本间的相关性。这个网络只依赖于关于字符串对之间相似性的知识，并将其映射至一个固定大小的嵌入式空间内。2017 年，Guo 等人[33]提出了动态 SNN，网络包含两部分：一个是用来处理模板图片的子网络，另一个是负责搜索图像的子网络。它能够高效地实时更新目标的外观变化及消除之前的背景干扰，同时充分运用丰富的时间-空间信息的运动对象。最后，这些数据被逐元素多层融合在一起，以便实现基于多层次深度特征的多级自动组合网络输出的过程。2021 年，An 等人[34]使用 SNN 来检测和跟踪视觉对象。SNN 是基于对象检测网络和对象跟踪网络的关联而构建的，用于对运动对象进行分类。近几年，SNN 还被用来训练框架应用到医学领域，如 Liu 等人[35]利用在成对的横向半球间区域上训练的 SNN 的深度学习框架来利用全脑体积不对称的辨别力，最终实现检测与阿尔茨海默病和轻度认知障碍相关的大脑不对称性。

SNN 在各图像、跟踪算法、医学等领域都具有独特的优势，且在持续的优化和新技术结合中，相信 SNN 模型会在更多领域创造更多的可能性。

1.2.3 Transformer 国内外研究现状

Transformer 是谷歌于 2017 年提出的[36]，旨在克服传统 RNN 难以实现并行训练的挑战，这一突破性的模型在深度学习行业产生了深远影响，对自然语言处理领域（Natural Language Processing，NLP）的冲击尤为明显，具有里程碑式的意义。其核

心部分就是注意力机制，它能够迅速捕获到稀疏数据的关键特征信息，随后又进一步改进为自注意力机制、交叉注意力机制和多头注意力机制等多种形式，这些已经被广泛运用在图像处理、语音识别等领域。

随着进一步发展，2018 年，Niki 等人发布的 Image Transformer[37]最早将 Transformer 迁移到计算机视觉领域，通过限制自注意机制来关注局部邻域，显著增加了模型在实践中可以处理的图像的大小，自此打开了 Transformer 在图像处理领域应用的大门。典型的应用和算法有 iGPT（image GPT）[38]和 ViT（Vision Transformer）[39]，2020 年，Dosovitskiy 等[39]提出了 ViT 模型，它采用完全的自注意力机制作为其核心技术进行图像分类，并能在大规模的数据集上实现良好的性能表现。如果能够提供足够多的训练数据，ViT 可以超越传统的 CNN 模型，突破 Transformer 缺少归纳偏置的限制，可以在下游任务中获得较好的迁移效果。针对图像分类这一计算机视觉的基础和核心任务，一些经典模型取得了显著的成就。DeiT（data-efficient image transformers）[40]采用了知识蒸馏策略，使得视觉 Transformer 能够学习归纳偏差，在 ImageNet 数据集上取得了 83.1%的 Top-1 准确度；Swin Transformer[41]将自注意力的计算范围限制在不重叠的局部窗口内，并通过移位窗口操作实现了局部窗口间的交互，在 ImageNet22k 数据集上取得了 86.4%的 Top-1 准确度；DINO（DETR withimproved denoising anchor boxes）[42]结合了 SSL 和 Transformer，使得可学习的特征更具解释性，在 ImageNet 数据集上达到了 80.1%的 Top-1 准确度。上述工作表明，尽管 Transformer 的提出时间不长，但在计算机视觉领域已经取得了明显的进步，奠定了其在该领域的主导地位。

随着 ViT 在计算机视觉领域的成功实验，遥感界在许多任务中使用基于 Transformer 的框架也出现了显著增长。下面着重介绍 Transformer 在遥感图像分类领域的研究现状，以 HSI 数据分类为例，分为只基于 Transformer 的方法、基于 CNN-Transformer 的混合方法和基于多模态融合 Transformer 的方法[43]。只基于 Transformer 的方法：在现有的工作中，He 等人[44]引入了 Transformer 的双向编码器表示，称为 HSI-BERT，具有全局感受野，无论其空间距离如何，都可以捕获像素之间的全局依赖性。所提出的架构非常灵活，可以从需要执行预训练的不同区域进行推广，实现灵活且动态的输入区域。2021 年，Hong 等人[45]提出了基于 Transformer 的骨干网络（SpectralFormer），该方法能够接受以像素级别或者区块为单位的光谱数据作为输入，可以从附近的高光谱波段捕获光谱局部序列知识。SpectralFormer 利用跨层跳跃连接，通过学习层级的软残差，将信息从浅层循环到深层，从而产生分组谱嵌入。Zhong 等人[46]提出了 SSTN（Spectral-Spatial Transformer Network），该网络模型包括空间注意力和光谱注意力，前者的任务是在所有的输入元素当中寻找最适合的空间关系表达方式，后面的则负责对相应的遮挡图像里的每一个具体坐标点上的相关属性值做出综合

性的评估分析工作。通过联合两个模块解决了卷积核的固定几何结构阻碍了远距离特征之间的远程交互问题。Liu 等人[47]还在空间和光谱维度上探索了 Transformer，使用 Transformer 层代替卷积层，提出了 DSS-TRM（Deep Spatial-Spectral Transformer）来实现端到端的 HSI 分类，此方法中包含光谱自注意力和空间自注意力，分别用来捕获光谱和空间维度的特征，然后将两部分产生的特征融合后输入分类器进行分类。

基于 CNN-Transformer 的混合方法：把 CNN 和 Transformer 的优势整合到一起，有助于更有效地获取细节的信息并处理遥感影像的高级类别识别问题中的远距离关联关系。Zhao 等人[48]引入了 CTN（Convolutional Transformer Network）结构，该模型通过中心位置编码将像素位置和光谱特征结合起来，进而生成数据的空间位置特征，最后借助 CNN-Transformer 提取到局部特征和全局特征。Yang 等人[49]提出了一种 HSI Transformer（HiT）分类方法，将 CNN 嵌入到 Transformer 架构中以进一步集成局部空间上下文信息。所提出的方法包括两个主要模块，其中一个模块是光谱自适应 3D 卷积投影，可以通过 HSI 的光谱自适应 3D 卷积层生成空间光谱局部信息。另一个模块是卷积置换器（Conv-Permutator），采用深度卷积来沿光谱、高度和宽度维度分别捕获空间光谱表示。Jia 等人[50]引入了一种多尺度 CNN-Transformer，该网络模型可以关注并提取到数据的空间和光谱特征，关注到更细节的信息。此外，作者还定义了一个自监督的 pretext，pretext 通过在编码器部分屏蔽掉中心像素所关联的 token，然后把剩下的 token 传送给解码器，来重新构建与中心像素相关联的光谱特征信息。pretext 可以更好地对中心特征和领域特征之间的关系进行建模，获得更稳定的训练结果。Sun 等人[51]提出了一种名为 SSFTT 的模型，它旨在获取光谱与空间特征和并融合生层次的语义信息，SSFTT 包含一个特征提取模块，该模块通过采用 3D 和 2D 卷积层来提取浅层光谱和空间特征。此外，SSFTT 中通过高斯加权特征标记器进行特征变换，然后将生成的特征输入到 Transformer 编码器进行特征表示，最后采用线性层来识别样本标签。

基于多模态融合 Transformer 的方法：最近有一些基于 Transformer 的工作也探索了融合不同模态的遥感数据（例如 HSI、SAR 和 LiDAR）来进行遥感图像分类。Roy 等人[52]提出了一种多模态融合 Transformer（MFT），MFT 中的数据融合方案用于从多模态数据以及标准高光谱图像块 token 导出 Transformer 中的类 token。此外，MFT 中的注意力机制将来自高光谱和其他模态 token 的信息融合到一个新的综合特征 token 中。Xue 等人[53]提出了一种深度视觉 Transformer（Deep Hierarchical Vision Transformer，DHViT）架构，用于 HSI 和 LiDAR 数据融合分类，利用光谱序列 Transformer 在光谱维度从 HSI 数据中提取特征，并利用空间分层 Transformer 从 HSI 和 LiDAR 数据中提取分层空间特征。

1.2.4 自监督学习国内外研究现状

SSL 是监督学习和无监督学习之间的桥梁，SSL 模型可以从未标记的样本中学习数据表示。目前 SSL 应用到 NLP、医学、计算机视觉等领域中都有很好的效果[54]。

自 2018 年 Google 推出 NLP 模型以来，SSL 已成为研究人员的研究热点。SSL 在 NLP 领域最丰硕的成果是 BERT 和 T5。学者们在这一领域已经进行了大量的研究。Zhou 等人[55]在 NLP 中使用 SSL 正则化技术进行文本分类，他们将文本分类定义为 NLP 中的一个关键概念，并提供了使用文本编码器作为输入进行编码的训练文本，在编码文本中，他们定义了两个任务：预测和自监督，这两个任务使用相同的编码文本，他们在 17 个文本分类数据集上测试了模型，试图最大限度地减少分类和正则化损失。Chen 等人[56]使用混合 SSL 方法来规范文本分类任务的训练，他们将对抗性训练引入 SSL，首次提出通用的鲁棒预训练模型，介绍了一个用于视觉表示对比学习的简单框架，简化了对比自监督学习算法，从而不需要专门的架构或存储库。

在计算机视觉领域中：SSL 已广泛应用于目标检测、图像分类、视觉问答等领域。文献[57]是第一个应用于遥感图像分类的 SSL 方法，文中提出了多层特征匹配生成对抗网络（MARTA GAN），MARTA GAN 由生成模型 G 和判别模型 D 组成，将 D 视为特征提取器，为了适应遥感数据的复杂特性，他们使用融合层来合并中层和全局特征，G 可以产生大量与训练数据相似的图像。MARTA GAN 的核心概念是从不同网络层提取多级特征，并通过串联将它们聚合在一起。Stojnić 等人[58]提出了 SSL 中生成模型应用于遥感的另一个早期用途，作者评估了裂脑自动编码器在自监督图像表示中的使用，在学习重建输入图像的过程中，自动编码器发现有关数据分布的相关信息。实验结果证明，即使使用少量未标记的训练图像，通过微调自动编码器学习的权重，也可以在 AID 数据集上有较好的结果。Gidaris 等人[59]提出了一种基于 ConvNets 的方法来识别图像中的 2D 旋转。Shu 等人[60]提出了一种基于 SSL、无需手动标记的图像分类算法 OCFC，图像预处理后，通过三层卷积受限玻尔兹曼机提取特征，然后通过模糊 C 均值算法对提取的特征簇进行伪标签标记，最后利用 CNN 模型对其他图像类别进行分类和预测，SSL 模型可以任意迁移到浅层模型或深层模型。Li 等人[61]针对显微镜图像提出了一种多任务 SSL 框架 ColorMe，该框架深度挖掘了原始数据中蕴含的丰富信息，并摆脱了对训练数据的要求。Zhao 等人[62]为了提高模型的特征提取能力和泛化能力，提出通过使用具有混合损失函数的多任务学习模型，该模型结合了自监督和监督训练策略。Jung 和 Jeon[63]提出了另一种利用三重态损失的方法，他们修改了原始三重态损失以更好地适应遥感图像，他们的主要贡献是将三元组目标重新表述为二元分类问题，而不是度量学习问题。Scheibenreif 等人[64]利用对比自监督方法来执行多模态光学-SAR 数据融合，用于土地覆盖场景分类任务。他们提出了无增强对比 SSL 框架 Dual-SimCLR，只需 10%的标记数据即可进行微调，与整个数据集

上的监督训练相比，其准确性更高。这项工作展示了在遥感应用的自监督范例中使用Transformer 主干网取代 CNN 主干网的巨大潜力。随着计算机视觉和机器学习领域自监督方法的迅猛发展，除了分类任务，其他遥感任务，如分割、目标检测、超分辨率和变化检测，也在竞相达到最先进的水平。

自监督学习方法主要可以分为三种：生成式、对比式和生成-对比式（对抗性），使用较多的为生成式和对比式[65]。过去，通过视觉相似性形成的自然聚类学习生成无监督学习表征[66]。Zhuang 等人[67]通过对比学习来实现自监督学习的目标，近邻与背景中的相邻数据结合起来形成一个集合。背景邻居是嵌入空间中与查询图像接近的点的随机样本。作为无监督聚类算法的一部分，使用一组正样本及其背景邻居将数据馈送到聚类算法中，该算法使用来自这些样本的数据。对比学习能够将场景或上下文的不同视角映射到表征空间的同一区域[68]。根据最终目标的不同，可以产生许多不同的相似性和不相似性概念，这就是对比方法如此有效的原因[69]。He 等人[70]提出了基于对比自监督学习的方法证明了自监督图像表示学习的有效性，该方法在各种下游任务上缩小了无监督和有监督表示学习之间的差距。它们可以在没有标签的情况下学习通用的视觉表示，并且在线性分类和转移到其他任务或数据集方面表现良好。Ermolov 等人[71]提出了最小化增强实例间的 MSE 距离，以保证在许多图像处理管道中可以批量应用 whitening 操作。Drouyer 等人[72]探索了一些最相关的对比学习技术在建筑屋顶航拍图像的大型未标记数据集上的分类任务应用，望解使用更小的标记数据集进行屋顶类型分类任务。

1.3 本文主要工作

本文基于 CNN、Transformer 和对比学习，研究了基于多模态特征融合的遥感图像分类技术。以遥感图像分类任务为研究对象，基于目前遥感图像分类任务中的单一模态特征信息有限且 CNN 对光谱信息提取不充分、模态间特征信息交互不充分、数据标注困难样且本量小三个问题，分别提出了基于双分支空谱残差网络的多模态遥感图像分类、基于双向交互融合的多模态遥感图像分类、基于对比学习的多模态遥感图像分类三种方法。本文主要研究思路如图 1.1 所示，具体研究内容为如下三部分：

（1）提出了基于双分支空谱残差网络的多模态遥感图像分类算法。针对遥感图像数据单一模态的分辨率和提供的特征信息有限、缺乏空间结构特征和高程特征信息，CNN 对 HSI 数据的光谱特征提取不充分等问题，引入 LiDAR 数据进行多模态数据融合，并设计针对 HSI 数据空间特征和光谱特征提取的双分支空谱残差网络，同时引入 SNN 和可学习参数进一步改善特征融合阶段的效果，提升了模型的特征提取能力和特征丰富程度。

（2）提出了基于双向交互融合的多模态遥感图像分类算法。针对异构特征间难以保持一致性、特征未充分交互、CNN 对高维数据 HSI 提取不充分且难以关注长距离依赖信息的问题，使用 Transformer 针对 HSI 数据进行特征提取，并引入一致性损失来保障异构特征之间的双向交互和特征一致性，在特征融合阶段引入交叉注意力机制，进一步增强两个模态间的特征的交互融合，从而提升模型的分类效果。

（3）提出了基于对比学习的多模态遥感图像分类算法。针对遥感图像分类数据样本量稀缺且标注困难的问题，在多模态遥感图像分类任务中引入对比学习方法，在单一模态内和两模态间增加对比学习，利用未标注数据学习特征表示的同时增加了模态间特征的语义一致，有效缓解了样本量稀缺及数据标注问题，在少样本条件下表现可观。

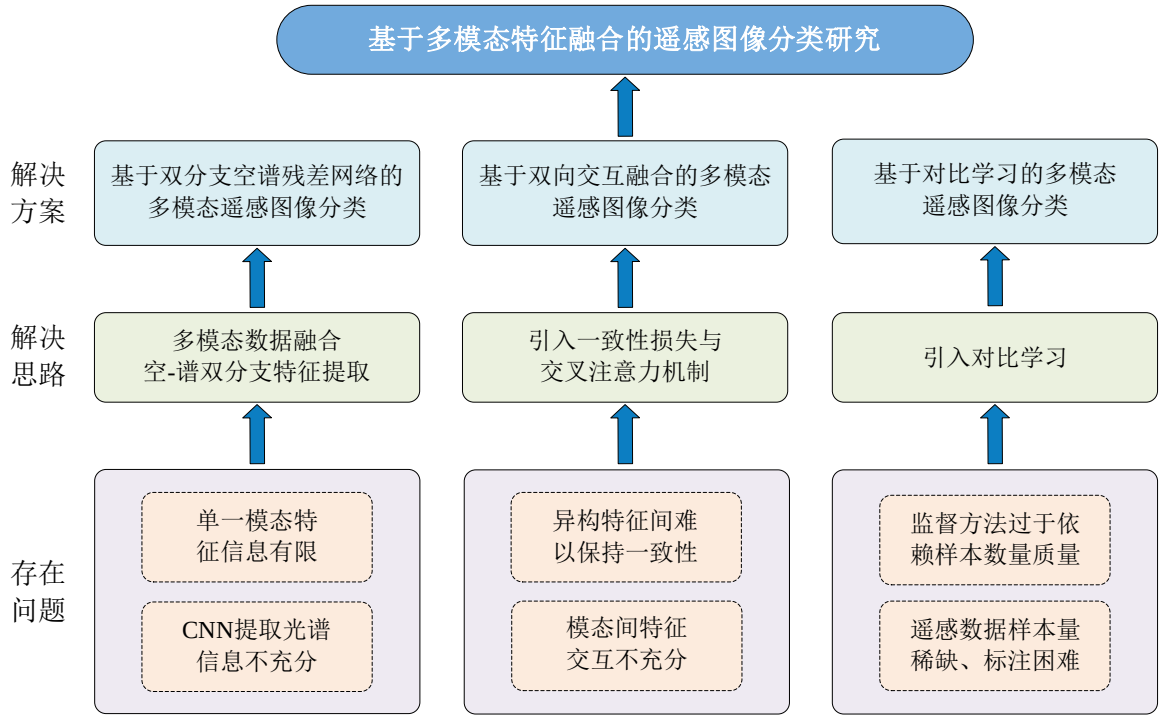


图1.1 论文架构流程图

论文的第一个创新点主要针对的是特征提取部分的缺陷进行改进和创新，第二个创新点的工作集中于特征融合阶段，前两个工作所使用的方法为监督学习的方法，由于监督学习固有的缺陷：存在对数据集数量及质量的依赖性，同时遥感图像分类任务的数据稀缺且标注困难，因此论文的第三个点使用的为自监督对比学习的方法。

1.4 本文结构安排

本文主要对基于多模态特征融合的遥感图像分类方法进行研究，论文共分五章，内容具体安排如下：

第一章为绪论，首先简单介绍了本文的研究背景及意义，叙述本文所涉及到的多模态遥感图像分类、SNN、Transformer 及自监督和对比学习技术的研究现状，最后，对论文研究工作中的主要贡献及创新之处进行介绍，阐述论文的研究内容及组织架构。

第二章为基于双分支空谱残差网络的多模态遥感图像分类方法的研究。章节首先介绍了多模态遥感图像分类和 SNN 的理论基础，随后详细描述了本章算法框架、基于双分支空谱残差网络（Dual-branch spatial-spectral residual network，DBSSRN）的特征提取网络以及基于孪生神经网络模块（Siamese neural network module，SNNM）和可学习参数模块（Learnable factor module，LFM）的特征融合部分。在实验阶段介绍了本篇论文使用的两种数据集，并通过对比实验、消融实验和可视化结果图证明了本章算法在多模态遥感图像分类任务上有着很好的表现。

第三章为基于双向交互融合的多模态遥感图像分类方法的研究。章节首先介绍 Transformer 和注意力机制的理论基础，随后详细介绍了本章算法框架、双支路 CNN 与 Transformer 的特征提取网络、交叉注意力特征融合模块和一致性损失模块。最后利用对比实验、消融实验及可视化结果图分析说明了本章所提出方法的有效性。

第四章为基于对比学习的多模态遥感图像分类方法的研究。章节首先介绍了自监督学习及对比学习的理论基础，随后介绍了本章算法框架，紧接着介绍了本章的数据增强方案、使用的特征提取器、模态内的对比学习和模态间的对比学习四个部分。最后通过对比实验、消融实验和可视化分类结果图说明了该方法的有效性。

第五章为全文总结，对论文的全部工作进行了综述，并对论文中存在的不足进行了分析，最后对今后的研究进行了展望。

第二章　基于双分支空谱残差网络的多模态遥感图像分类

2.1 引言

常见的传感器拍摄的单一模态数据都有其缺陷，如 HSI 模态数据无法区分使用相同材料建造的道路和屋顶，关于物体形状、纹理或空间结构的信息比较少，但是，LiDAR 通过测量激光脉冲的往返行程时间来确定传感器与测绘地形之间的距离，从而获取地形的高程信息，这对于区分不同地形地貌（如山地、平原、水体等）非常有用，可以很好的补充 HSI 中不足的空间结构信息和高程信息。因此结合 HSI 和 LiDAR 的特征信息对于分类目的非常有益。通过多模态遥感数据融合，可以克服单一模态遥感数据存在的局限性和不足，综合不同遥感传感器数据的地物特征信息，克服光谱信息局限性、时间和天气条件限制、地物识别困难等问题，目前常用的多模态遥感图像分类方法中以基于特征融合的方法效果最好，在特征提取和特征融合阶段利用深度学习方法完成。此外，CNN 因为其感受野和卷积操作，而对空间结构提取比较充分，但 HSI 数据含有复杂的光谱信息，CNN 对 HSI 数据的光谱特征提取不充分，会影响整体的分类效果。

为解决上述问题，本章主要进行以下几项工作：

（1）针对 HSI 单一模态数据的分辨率和提供的特征信息有限、缺乏空间结构特征和高程特征信息等问题，采取多模态特征融合完成遥感图像分类任务，引入空间结构特征和高程特征信息比较丰富的 LiDAR 数据进行双支路特征提取和融合，更好的完成分类任务。

（2）针对 CNN 对 HSI 数据的光谱特征提取不充分的问题，本章提出了双分支空谱残差网络（DBSSRN）分别针对空间特征和光谱特征进行提取，进一步丰富 HSI 数据特征提取阶段的信息，使得分类结果进一步优化。

（3）在特征融合阶段使用 SNNM 和 LFM，辅助在特征融合阶段更好的融合两种模态的特征，而不是直接进行拼接或相加融合，进一步提高分类的准确率。

2.2 相关理论基础

由于本篇论文所提出的方法都是关于解决多模态遥感图像问题的，所以先在本节介绍多模态相关的理论知识，随后介绍本章所使用到的 SNN 相关的理论基础，关于 CNN 的理论基础在此就不赘述了。

2.2.1 多模态遥感图像分类

多模态遥感图像分类是指结合不同传感器的数据涵盖的特征，通过融合使得分类任务的数据信息进一步得到补充，提高分类精度和效率。在多模态遥感图像场景分类任务中，分为两个大的任务环节，即遥感图像特征提取环节和融合环节，因为本章主要针对融合环节做相应创新研究，所以此处主要对融合环节进行相应的理论基础介绍。

多模态遥感图像通常包含来自不同传感器、不同时间、不同分辨率的图像数据。常用的有 HSI 数据、合成孔径雷达（SAR）数据和 LiDAR 数据等，这些数据具有不同的特点和优势，但同时也存在冗余和互补性。

根据融合方式的不同，通常可以划分为以下三种融合方法：像素级融合、特征级融合和决策级融合。其中像素级融合保留了尽可能多的细节信息，相较于其他两种策略精度更高，但是传感器需要处理的数据量大，需要配准的精度要求太高，效率很低，分析能力不强，常见的像素级融合方法有 Brovey 变换融合、亮度-色调-饱和度变换融合、主成分变换融合、小波变换融合等。图 2.1 为像素级融合的示意图。

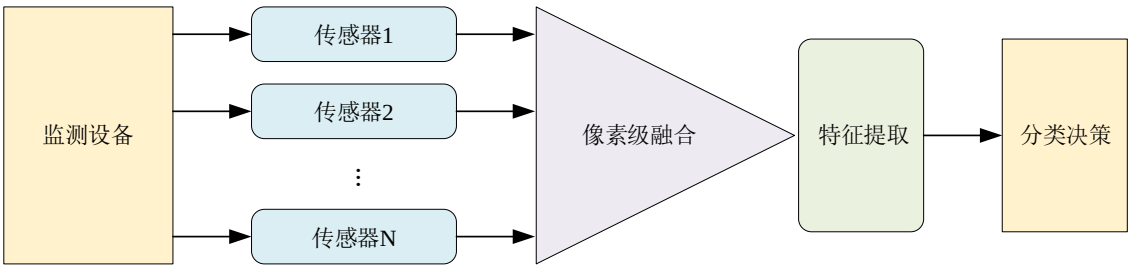


图2.1 像素级融合示意图

特征级融合主要关注的是特征之间的关系而非像素之间的关系，这有助于减少人为误差，如像素重采样等问题。然而，由于它是通过特征来实现而不是直接使用原始图像数据，因此在特征提取的过程中可能会丢失部分原始数据信息，进而可能引发分类精度上的问题。通常的操作流程是：首先将各模态的遥感图像数据进行特征提取，然后根据这些特征信息对各模态数据进行分类、聚集和综合，形成特征向量，再采用一些基于特征级融合的方法融合这些特征向量，最后输入分类器进行分类操作，示意图如图 2.2 所示。

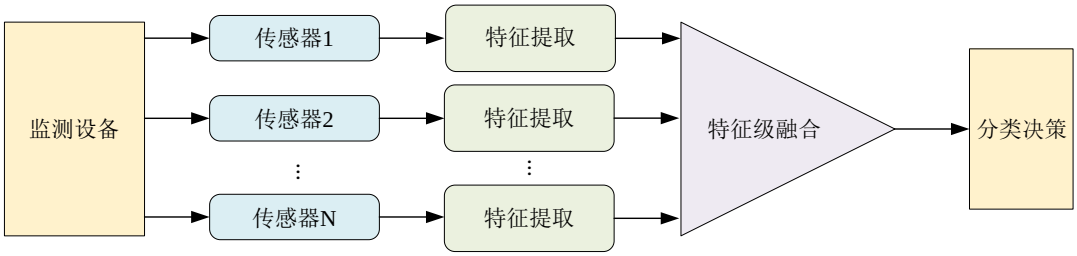


图2.2 特征级融合示意图

决策级融合的方法具有高容错性，其执行速度快、对数据要求不高、具备强大的分析能力。然而，它需要高水平的数据预处理和特征提取技术，因此成本相对较大。通常的操作流程包括：数据预处理、特征提取、定义属性和融合这些属性，最后再确定融合后的属性描述，可以参考图 2.3 来理解。

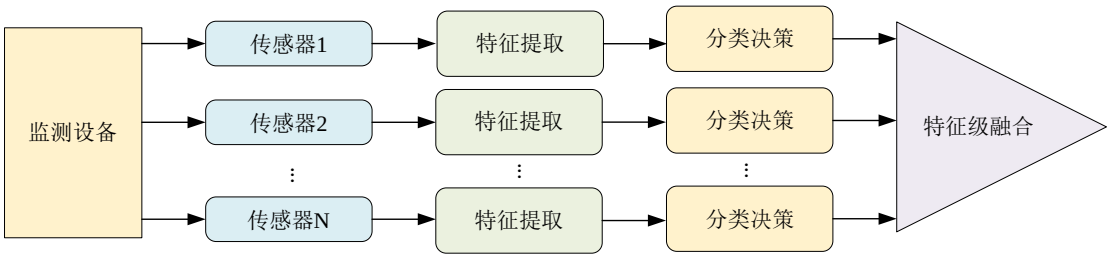


图2.3 决策级融合示意图

2.2.2 孪生神经网络

SNN 是建立在两个或多个共享权重的人工神经网络基础之上形成的一种紧密的耦合结构，网络由多个卷积层、池化层和全连接层组成，可以接受不同大小的输入。SNN 在训练时输入的是一对样本而不是单个样本，并通过高维空间中的嵌入表征来输出它们之间的相似度比较结果，结果呈现为标签，标签为 1 则属于同一类，为 0 则属于不同类，然后使用交叉熵损失函数训练模型。根据网络构成的不同，SNN 可以分为狭义和广义两种类型。

狭义上的 SNN 由两个结构完全相同且权重共享的神经网络组成，它们相互拼接而成。广义上的 SNN 则更为灵活，可以由任意两个神经网络拼接而成，例如 CNN 和 RNN 等。这种架构使得 SNN 在处理各种需要比较数据相似度的任务时表现出色。

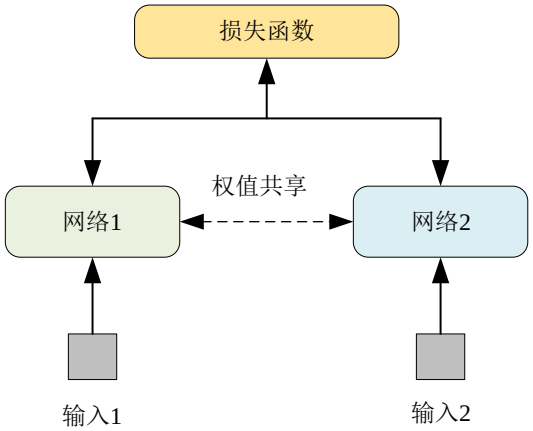


图2.4 SNN 结构示意图

SNN 的结构可以根据具体任务和数据类型进行灵活设计。在图像处理任务中，CNN 通常被用作孪生网络的基本组件，因为它们能够有效地提取图像中的空间特征。

在 NLP 任务中，RNN 或变体如 LSTM 可能更合适，因为它们能够处理序列数据并捕捉时序依赖关系。

此外，为了提取更具鉴别力的特征表示，研究者们还探索了各种先进的网络结构和技巧，如深度残差网络（Residual Network，ResNet）、注意力机制等。这些创新不仅提高了 SNN 的性能，还拓宽了其在不同领域的应用范围。

SNN 包含多种损失函数，如二元交叉熵损失、对比损失、三重态损失，用于在给定特定输入的情况下衡量机器学习模型中预期输出与实际输出之间的差异。训练模型时，目标是通过调整模型的参数来最小化该损失函数[73]。

二元交叉熵损失函数在二分类任务中的作用比较明显，因为该损失函数是用来预测两个样本间的可能结果，而图像分类任务中 SNN 目标恰恰是判断两个图像数据的“相似”和“不相似”。该函数量化了正类的预测概率与实际结果之间的差异，损失函数公式如下：

$$l = -\left[y\log(p) + (1-y)\log(1-p) \right], \tag{2-1}$$

其中，y 表示真实标签，p 表示预测概率。使用二元交叉熵损失训练模型力求通过参数调整来最小化该函数。通过这种最小化，模型可以熟练地进行准确的类别预测。

对比损失通过使用距离作为相似性度量来判断数据样本之间的相似程度。对比损失函数需要成对的负训练样本和正训练样本，且一般在每个类的训练样本数量有限时比较有效。损失函数公式如下：

$$l = \frac{1}{2N}\sum_{n=1}^{N}\left[yd^2 + (1-y)max(0, m-d)^2 \right], \tag{2-2}$$

其中，N 表示样本数量，X_1 和 X_2 是输入数据对，$d = \|X_1 - X_2\|_2$ 是指两个特征样本间的欧氏距离，y 为 0 表示两个样本不相似，为 1 表示两个样本相似，m 是设定的阈值，表示不相似的距离阈值$[0,m]$，超过 m 时，两个样本不相似性可以看做 0。

2.3 基于双分支空谱残差网络的多模态遥感图像分类算法

2.3.1 整体网络架构

为解决单一模态数据进行遥感图像分类任务的缺陷，如高光谱虽然有丰富的光谱信息，但是缺乏比较准确和细节的空间结构信息和高程特征，所以考虑引入 LiDAR 支路进行多模态特征融合，整篇论文也是基于这个研究基础出发和进行实验的，同时

为了解决 CNN 对光谱特征提取不充分的问题，本章提出了基于双分支空谱残差网络的多模态遥感图像分类，网络的整体结构如图 2.5 所示。

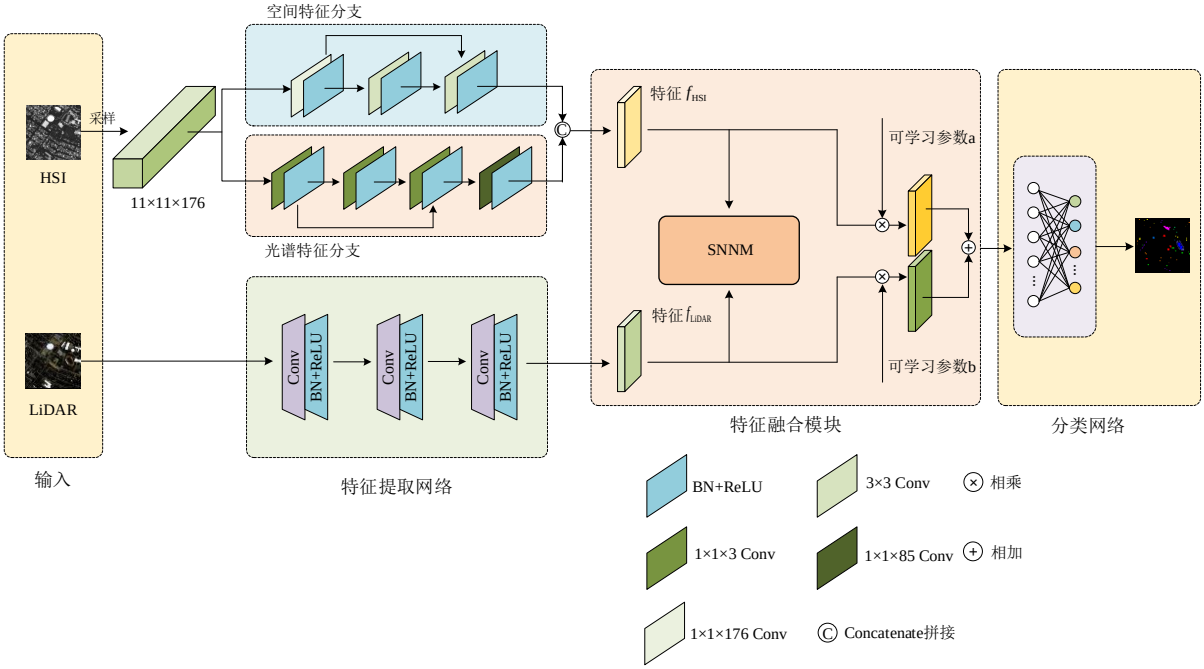


图2.5 基于双分支空谱残差网络的多模态遥感图像分类算法结构图

网络模型主要包含 DBSSRN 和 Resnet18 组成的特征提取网络、SNNM 和 LFM 组成的特征融合部分、分类网络三个部分。其中 DBSSRN 包含对 HSI 的空间特征提取分支和光谱特征提取分支，SNNM 中使用的是经典的 SNN 进行的特征相似度度量，LFM 中使用可学习参数保障了融合阶段最合理的权值，最后使用的分类网络为简单的全连接网络

2.3.2 基于双分支空谱残差网络的特征提取网络

在特征提取阶段，本章使用两个 CNN 分别对 HSI 数据和 LiDAR 数据进行特征提取，其中 HSI 支路使用的是所提出的 DBSSRN，在 LiDAR 支路使用的是经典的 CNN 网络 Resnet18，因为 LiDAR 通道少、数据较为简单，所以此处没有单独针对性的设计对应的 CNN 进行特征提取，具体情况见下文。

（1）HSI 数据特征提取

在 HSI 数据的特征提取阶段，使用 DBSSRN 来进一步提高对 HSI 数据的特征提取效果，DBSSRN 的结构图如图 2.6 所示。DBSSRN 主要包含两个分支结构，即空间分支和光谱分支，分别针对 HSI 数据的空间特征和光谱特征进行针对性的特征提取，之后进行空间-光谱融合。DBSSRN 先从原 HSI 数据中采样出多个三维立方体作为模型的输入，立方体的参数为 11×11×176，分别代表 11 个像素宽、11 个像素高和 176

个光谱带，然后输入到空间分支和光谱分支分别进行对应的特征提取，最后把两分支得到的空谱特征拼接得到最终的 HSI 支路所提取到的特征。

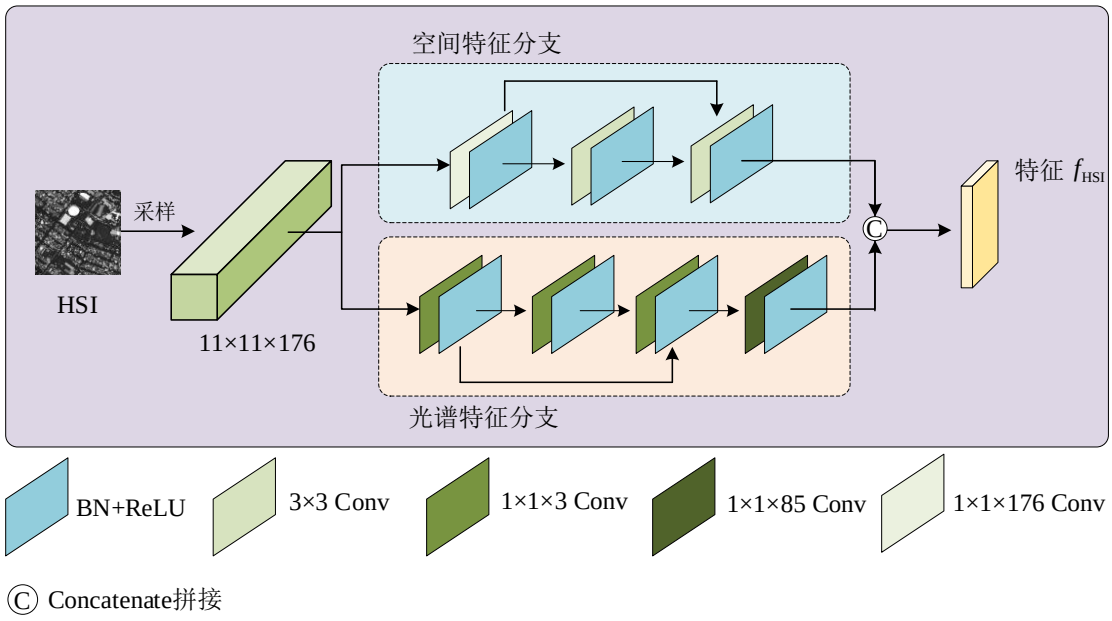


图2.6 DBSSRN 网络结构图

（a）空间分支

空间特征提取分支主要是为了捕获目标像素周围的空间信息而设计的。该分支中采用了三个 2D 卷积层来更好地利用空间特征，每个卷积层的核大小为 3×3，之后都有一个批归一化（Batch Normalization，BN）层和一个 ReLU 激活函数层。此外，每个卷积层的卷积核数量相同，不进行下采样操作，因此空间分支的输出和输入大小相同。在空间特征提取分支中有 1 个残差块，如图 2.7 所示。首先，把输入的特征与 1×1×176 的核进行卷积，将 HSI 的多波段信息映射到只有一个光谱波段的灰度图像上。这个操作旨在让网络专注于检测空间域中的特征的相关性。所得到的 11×11 的块被传递到残差块中，残差块由两个连续的卷积层组成，每个卷积层有 24 个 3×3 的核，以保留和强调关键的空间信息。

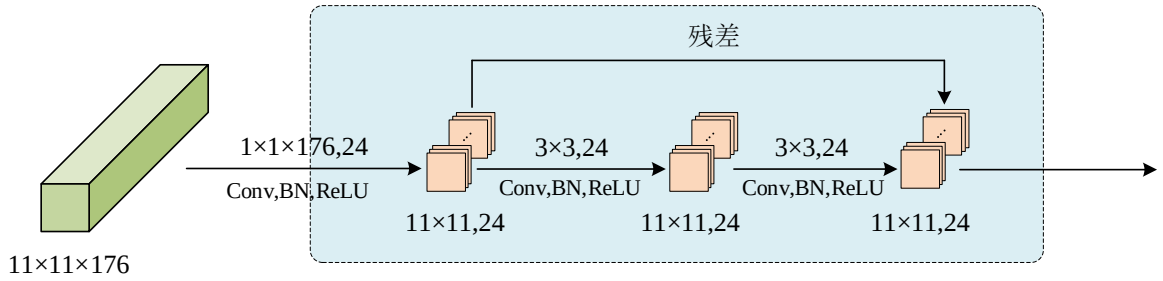


图2.7 空间分支结构图

（b）光谱分支

光谱特征提取分支与空间特征提取分支类似，对于局部光谱序列的特征提取，前3 组 3D 卷积层的核大小设置为 1×1×3，且数量相同，以学习光谱特征，同时保持原有的空间相关性。每个卷积层后面都跟随一个 BN 层和一个 ReLU 激活函数层。在光谱特征提取分支中也有 1 个残差块，如图 2.8 所示。首先在第一个卷积层中，使用 24 个 1×1×3 的核，步长为（1，1，2），主要是为了消除多余的光谱信息并专注于对分类任务更关键的特征，这里的处理将初始数据转换成了 24 个尺寸为 1×1×85 的 3D 立方体。然后，使用 24 个具有 1×1×3 的核的连续卷积层构建了残差块，用来在处理噪声标签时强调关键区域，并且提升光谱抗干扰能力。最后，剩下的立方体与 128 个 1×1×85 核进行卷积操作，生成 128 个大小为 11×11 的 2D 空间块。

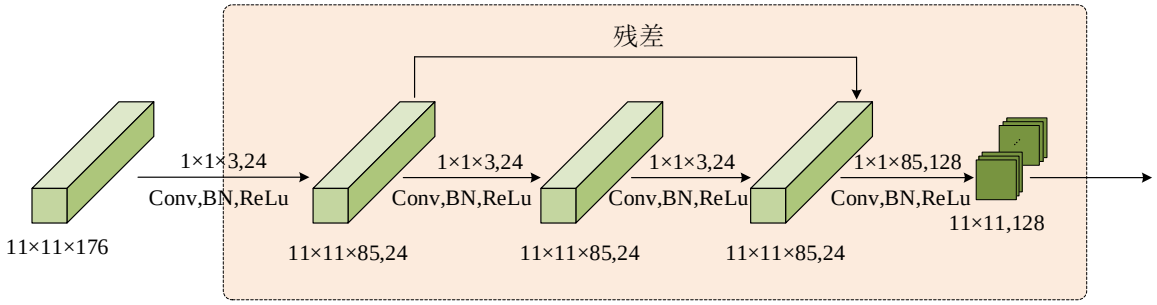


图2.8 光谱分支结构图

（c）空间-光谱融合

基于以上两个分支的工作，从空间特征提取分支和光谱特征提取分支中学习到的特征沿着通道维度进行级联。即将 24 个 11×11 的二维空间块和 128 个 11×11 的二维光谱块直接拼接，得到 152 个 11×11 的 HSI 块，后续工作就是同 CNN 特征提取的 LiDAR 特征进行融合操作。

（2）LiDAR 数据特征提取

CNN 提取 LiDAR 数据特征的网络模型使用的是 Resnet18，其中有 17 个卷积层和 1 个线性映射接层，最终输出一维长度为 256 的特征图。结构如表 2.1 所示。

先是对 LiDAR 数据进行卷积、正则化、最大池化等操作，后续的卷积部分都具备含有二层卷积的残差块，残差块可以把前面数据结果直接送到下一层的输入部分，可以有效避免梯度消失的问题，此外增加网络的层数也不会导致 CNN 模型最终的效果变差。这里本章并没有针对 LiDAR 数据的特点单独设计或者引用对应的 CNN 模型，主要是因为 LiDAR 数据主要包含的是高程信息和结构信息，没有像 HSI 数据的复杂光谱信息，现有的 CNN 对其提取能力和效果都很不错。此处就不进一步介绍使用 Resnet18 网络模型的细节信息了。

表2.1 CNN 支路的 Resnet18 模型结构表

名称	输出尺寸	结构	个数
Conv1	11*11	7*7，64	
Maxpool		3*3	
Conv2_x	11*11	3*3，64 3*3，64	2 个
Conv3_x	11*11	3*3，128 3*3，128	2 个
Conv4_x	11*11	3*3，256 3*3，256	2 个
Conv5_x	11*11	3*3，512 3*3，512	2 个
Linear	256		

2.3.3 基于 SNNM 和 LFM 的特征融合网络

在特征融合阶段，本章结合了 SNNM 和 LFM，避免了对 HSI 和 LiDAR 数据的直接拼接而造成的融合不合理，而是通过 SNNM 先控制两组数据特征的一致性，然后通过可学习参数进行动态融合，通过融合结果和多模态数据集的真实标签进行对比后，反馈到 LFM，进行自适应动态调整。

（1）SNNM

SNNM 中共享权值的 CNN 使用的是 VGG16 网络模型，这里解释一下为什么特征提取阶段已经提取到了 HSI 和 LiDAR 数据的特征，这里还要使用 VGG16 进一步处理，因为在特征融合阶段，共享权值的 VGG16 的作用主要是帮助模型更好地理解和比较来自不同输入的特征，共享权值使两个输入分支中的网络具有相同的结构和参数，有助于确保网络在处理两组输入特征时，可以进一步发现更细微或更深层次的特征，在特征融合阶段，也有助于网络更好地整合来自两个模态输入的特征信息，提高两组特征的可比性和一致性从而提高模型的性能，增加分类的准确率。SNNM 的结构图如图 2.9 所示。

VGG16 的网络模型在此处就不赘述了。从 DBSSRN 中提取到的 HSI 特征 f_{HSI} 和从 Resnet18 中提取到的 LiDAR 特征 f_{LiDAR}，输入到 SNN 中的 VGG16 共享权重网络

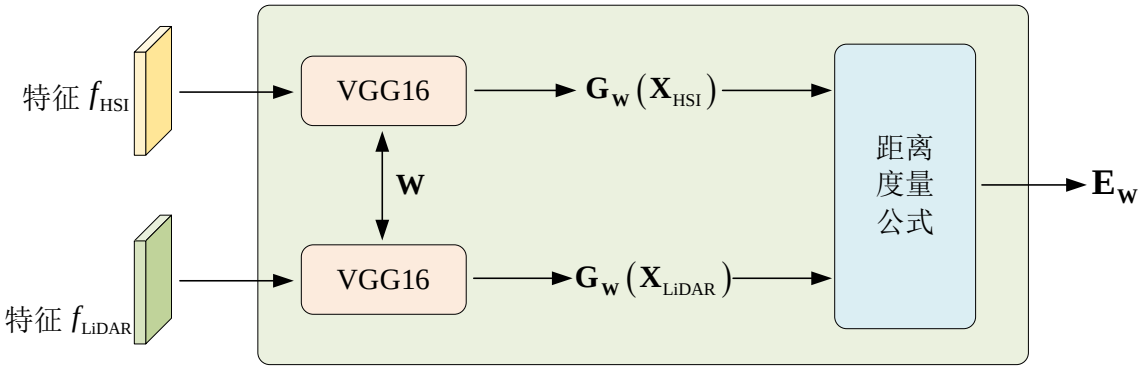


图2.9 SNNM 结构图

模型中，输出各自的特征向量$G_W(X_{HSI})$和$G_W(X_{LiDAR})$，然后使用距离度量公式计算两组特征向量的距离E_W，这里本章使用的距离度量公式为简单的 abs 距离公式：

$$E_W = \left| G_W(X_{HSI}) - G_W(X_{LiDAR}) \right|, \tag{2-3}$$

（2）LFM

可学习参数是指在训练过程中学习的参数值。在深度学习和机器学习中，这些参数是模型的重要组成部分，可学习参数可以通过迭代的方式根据训练数据进行更新，实现优化模型性能的目的。可学习参数通常包括模型的权重和偏置项等，它们在训练过程中会不断调整，以最小化损失函数，从而提高模型的预测能力。

本章在此处设置可学习参数a和b，具体设计位置如图 2.5 所示，现在大部分的特征融合的最后环节使用的是直接拼接的操作，而这样可能会因为模态间的差异化特征而导致融合的特征效果不好，造成最终的分类效果受到影响。此处本章的处理是在 SNNM 干预调整下的 HSI 特征和 LiDAR 特征前分别乘上可学习参数a和b，然后两路特征相加，得到融合后的特征，如以下式所示：

$$f_{HSI+LiDAR} = af_{HSI} + bf_{LiDAR}, \tag{2-4}$$

其中，f_{HSI}是 DBSSRN 的空谱分支提取出的两路特征拼接后的 HSI 数据特征，f_{LiDAR}是 Resnet18 提取过后的 LiDAR 数据特征。$f_{HSI+LiDAR}$是最后会被送到分类器的特征，但是此处会在训练阶段对参数a参数b进行迭代，通过反向传播不断更新，得到融合阶段的最终参数。

2.4 实验与分析

2.4.1 数据集介绍

（1）Houston 2013 数据集

表2.2 Houston2013 数据集介绍

类别序号	类别名称	训练样本	测试样本	样本数
1	健康草地	198	1053	1251
2	受压草地	190	1064	1254
3	人造草地	192	505	679
4	树木	188	1056	1244
5	土壤	186	1056	1242
6	水	182	143	325
7	住宅	196	1072	1268
8	商业	191	1053	1244
9	道路	193	1059	1252
10	高速公路	191	1036	1227
11	铁路	181	1054	1235
12	停车场 1	192	1041	1233
13	停车场 2	184	285	469
14	网球场	181	247	428
15	跑道	187	473	660
	总计	2832	12,197	15,029

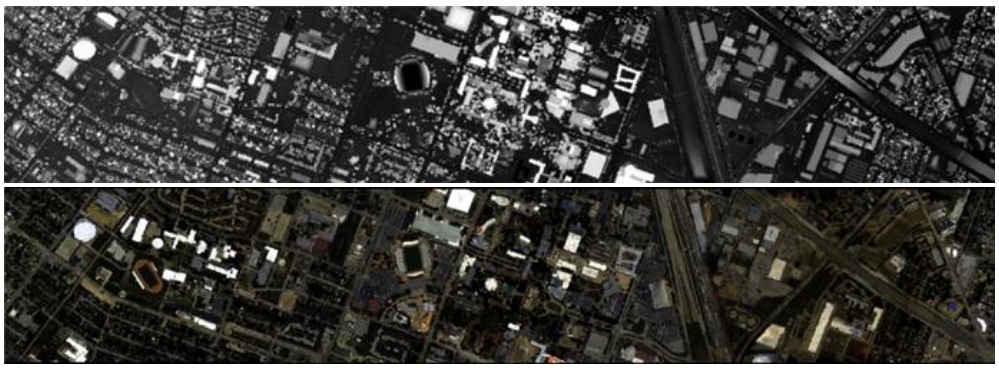


图2.10 Houston2013 数据集：从上到下是 LiDAR 衍生的数字表面模型（DigitalSurface Model，DSM）图像、HSI 伪彩色图像

该数据集是通过 ITRES CASI-1500 传感器获取的，覆盖了休斯顿大学校园及其相邻区域。其中 HSI 数据的光谱范围在 364 至 1046 纳米之间，具有 144 个光谱波段，波长范围从 0.38 到 1.05 微米。LiDAR 图像是一个 349×1905 像素的单波段数据。HSI 和 LiDAR 数据的空间尺寸均为 349×1905，空间分辨率为 2.5 米，空间分辨率为 2.5 米。该数据集共有 15 个地表覆盖类别，分别是健康草地、受压力草地、人造草地、树木、土壤、水、住宅、商业、道路、高速公路、铁路、停车场 1、停车场 2、网球场和跑道，具体的训练和测试样本数如表 2.2 所示。

（2）Trento 数据集：该数据集由 AISA Eagle 传感器和 Optech ALTM3100EA 传感器获取，覆盖了意大利特伦托南部的一个农村地区。其中的 HSI 数据包含 63 个光谱波段，波长范围从 0.42 到 0.99 微米，光谱分辨率为 9.2 纳米。空间尺寸为 166×600，空间分辨率为 1 米。数据共有六个地表覆盖类别，分别为：苹果树、建筑、地面、木材、葡萄园和道路，这 6 类数据中具体的训练和测试样本数如表 2.3 所示。

表2.3 Trento 数据集介绍

类别序号	类别名称	训练样本	测试样本	样本数
1	苹果树	129	3905	4034
2	建筑	125	2778	2903
3	地面	105	374	479
4	木材	154	8969	9123
5	葡萄园	184	10,317	10,501
6	道路	122	3052	3174
	总计	819	29,395	30,214

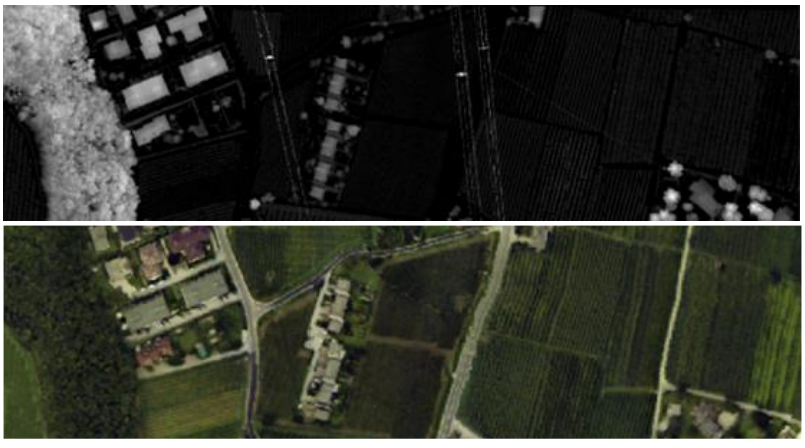


图2.11 Trento 数据集：从上到下是 LiDAR 衍生的 DSM 灰度图像、HSI 伪彩色图像

2.4.2 评价指标

本文根据三个常用指标定量评估每个模型的分类性能，即总体准确率（Overall Accuracy，OA）、平均准确率（Average Accuracy，AA）和 Kappa 系数（Kappa Coefficient，KC）。此外，将不同模型获得的分类图可视化以进行定性比较。

总体准确率 OA 是所有正确分类的样本数占总样本数的比例，反映了分类器在整个数据集上的整体性能。但需要注意的是，当数据集中各类别的样本数量分布不均衡时，OA 可能不能全面反映分类器的性能。OA 的计算公式如下：

$$OA = \frac{\sum_{i=1}^{k} m_{ii}}{N}, \tag{2-5}$$

其中，N 表示样本总数，m_{ii} 表示属于第 i 类的样本被正确分为 i 类的样本数，k 为总类别数。

平均准确率 AA 是各类别准确率的平均值。其中，每个类别的准确率是该类别正确分类的样本数占该类别总样本数的比例。AA 考虑了数据集中各类别的性能，尤其适用于样本分布不均衡的情况。通过 AA，可以了解分类器在不同类别上的表现差异。AA 的计算公式如下：

$$AA = \frac{\sum_{i=1}^{k} CA_i}{k}, \tag{2-6}$$

KC 是用于衡量分类精度的指标，它考虑了实际分类与随机分类之间的误差。KC 系数值越高，表示分类精度越高。KC 能够反映分类器性能的稳定性，KC 对于样本分布不均衡的情况也具有一定的鲁棒性。KC 的计算公式如下：

$$Kappa = \frac{\left(k \times \sum_{n=1}^{m} c_{nn} - \sum_{n=1}^{m} k_n \times c_{nn} \right)}{\left(k^2 - \sum_{n=1}^{m} k_n \times c_{nn} \right)}, \tag{2-7}$$

其中，c_{nn} 代表第 n 行和第 n 列，m 代表测试样本的种类。

2.4.3 对比算法

在本章中，基于双分支空谱残差网络的多模态遥感图像分类算法将与多个优秀的分类网络在 Houston2013、Trento 数据集上以相同的实验配置进行对比实验。具体的对比算法包括：ELM[76]、DeepCNN[24]、FusAtNet[77]、EndNet[78]和 HRWN[79]。ELM 算法随机选择隐藏节点并分析确定单隐藏层前馈神经网络的输出权重，使其能够快速学习泛化性能；DeepCNN 在特征提取阶段使用了两个耦合的 CNN 对 HSI 和 LiDAR 数据提取特征并使用特征级和决策级融合来整合异构特征；FusAtNet 使用了自注意力和交叉注意力来突出光谱和空间特征并增强交互，为了对比实验的公平性，这里本文去掉了 FusAtNet 的特征增强操作；EndNet 引入深度编码器-解码器网络架构，强制融合特征依次重建多模态输入来融合多模态信息；HRWN 也是关注了 HSI 的空间和光谱特征建立了双隧道 CNN 并针对 LiDAR 数据的高程信息类间关系提出像素级亲和力分支。性能评价指标则选择的为上文提及的 OA、AA、KC。

2.4.4 实验环境与参数设置

（1）本章实验环境详细配置如表 2.4 所示：

表2.4 本章实验环境配置

实验环境	版本号
CPU	Intel(R) Core(TM) i9-10900X CPU @ 4.50GHz 64GB
GPU	NVIDIA Geforce RTX 3090 24GB
编程语言	Python 3.9
操作系统	Linux Ubuntu 20,04 LTS
Cuda	12.0
深度学习框架	Pytorch 1.12.0

（2）本章的参数设定如下：

本章算法在 Houston2013、Trento 数据集上进行性能评估，训练过程批大小设定为 128；采用的优化器为 Adam，权重衰减参数为 0.9；初始学习率设定为$5e^{-3}$，而后采用线性衰减策略对学习率进行衰减；最大迭代次数为 100 代；使用了交叉熵损失作为损失函数进行训练；实验结果表示为 OA、AA、KC。每个实验重复 10 次取平均值和标准偏差。为了保证实验的公平性和可对比性，本章方法和使用的所有对比方法均采用上述配置。

2.4.5 对比实验结果与分析

（1）Houston2013 数据集对比实验结果

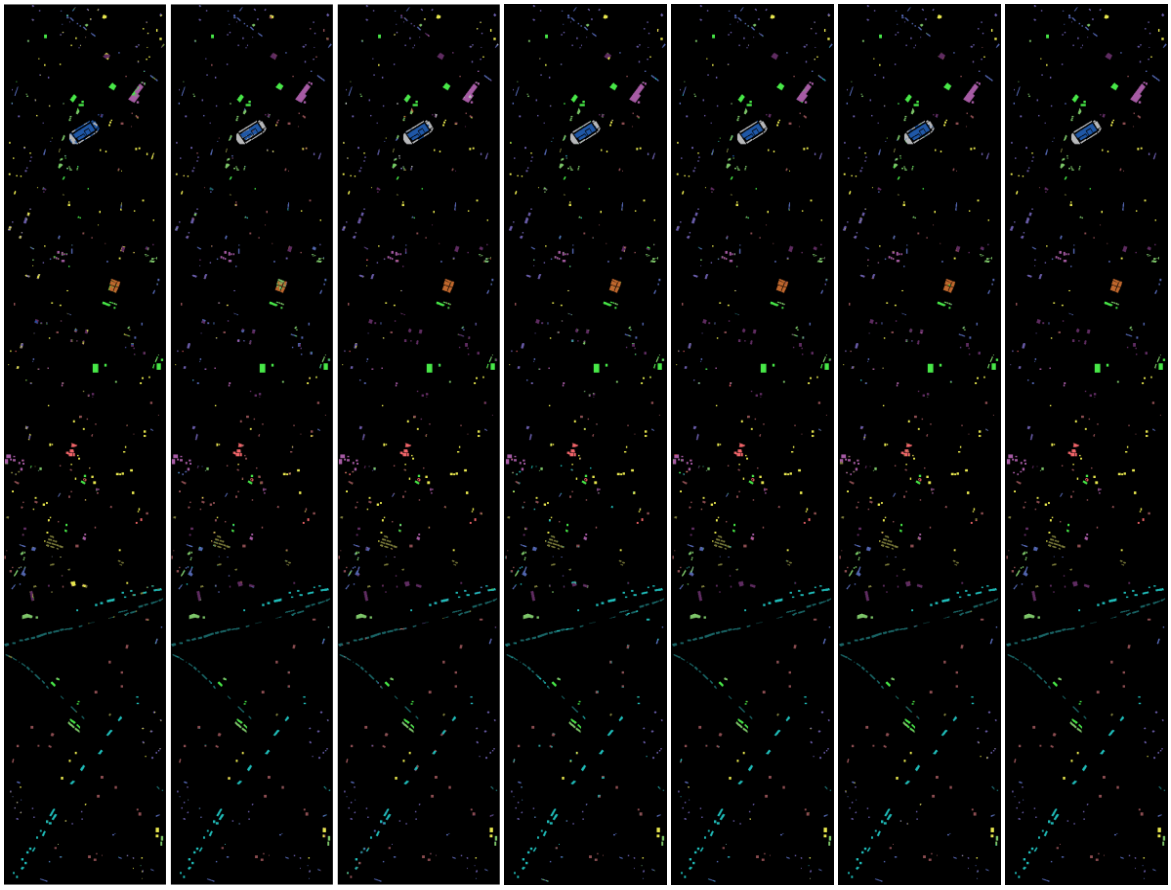
本章方法与对比方法的性能结果对比如表 2.5 所示，其对应的可视化结果图如图 2.12 所示，其中（a）~（g）分别代表 ELM、DeepCNN、FusAtNet、EndNet、HRWN、本章方法的分类可视化结果图和 Groundtruth 图。

表2.5 DBSSRN 在 Houston2013 数据集上的对比实验结果

No.	ELM	DeepCNN	FusAtNet	EndNet	HRWN	Ours
1	90.51±4.72	93.31±4.12	63.64±31.8	96.24±2.55	96.89±1.92	**97.01**±0.55
2	94.98±1.44	95.77±0.84	**98.05**±2.36	93.46±3.10	97.29±2.82	95.24±0.33
3	75.45±1.35	85.46±0.09	98.26±2.20	96.44±5.08	**99.64**±0.62	97.43±0.07
4	97.58±0.19	96.53±4.22	95.90±1.76	**98.51**±0.75	97.22±2.56	93.22±0.37
5	95.91±0.40	96.71±0.08	98.42±1.52	96.25±1.55	**99.45**±0.28	97.74±0.12
6	77.46±13.1	77.46±1.01	87.94±8.49	96.53±5.60	**100.00**±0.00	**100.00**±0.00
7	86.47±1.20	88.99±3.73	93.67±1.37	98.03±0.61	96.47±1.58	**98.55**±0.16
8	89.51±4.59	91.47±1.81	86.84±4.75	95.53±3.62	87.48±3.09	**95.26**±0.17
9	69.13±1.26	86.11±4.36	91.42±6.61	79.80±5.57	91.26±1.98	**92.13**±0.05
10	71.39±5.78	79.44±0.51	80.86±13.7	77.53±5.90	94.19±3.59	**95.74**±0.03
11	66.45±3.75	66.45±0.48	91.40±3.29	86.40±4.57	91.97±2.19	**93.89**±0.11
12	78.36±3.46	78.46±0.68	68.76±17.3	87.02±4.51	**96.77**±0.89	94.73±0.06
13	75.50±13.3	74.34±2.54	95.61±3.84	79.64±20.0	91.63±0.98	94.46±0.22
14	93.97±0.97	94.55±2.10	86.47±25.0	98.71±0.64	**100.00**±0.00	**100.00**±0.00
15	98.73±0.46	97.71±0.31	98.42±1.32	99.38±0.76	**100.00**±0.00	**100.00**±0.00
OA(%)	83.67±0.81	86.93±0.73	87.97±2.94	90.77±1.77	94.50±0.81	**95.23**±0.12
AA(%)	84.09±0.88	86.85±0.44	89.04±1.95	91.96±2.10	95.35±0.44	**96.36**±0.08
KC(%)	82.28±0.88	86.66±0.79	86.90±3.21	89.98±1.92	94.12±0.65	**95.07**±0.04

从表 2.5 中可以看出，本章算法在 9 个类别中产生了最高的分类准确率，其中在水、网球场和跑道类别上做到了 100%的分类准确率，且 OA、AA、KC 指标均超过对比方法结果达到最好，平均比对比方法的分类指标高 1%以上。与同样采取两个 CNN 分别提取 HSI 和 LiDAR 数据的特征的 DeepCNN 方法对比，本章的方法有 13 个类别的分类准确率超过 DeepCNN，并且两种方法均加入了耦合的 CNN，本文加入的为 SNNM 对两个模态数据特征相似性进行约束和训练，所以可以分析出，本章针对 HSI 数据的光谱特征和空间特征设计两条分支分别针对性提取特征的有效性，能

够帮助 CNN 进一步提取 HSI 数据的光谱特征。FusAtNet 方法是针对模态间的特征交互设计的算法，特征提取的网络用的是简单的 6 层 CNN 网络，与本文的方法主要区别在于 FusAtNet 针对光谱特征和空间特征分别设计对应注意模块，本文是放在了特征提取阶段设计光谱和空间两个特征提取分支，但是效果上本章的方法更能关注到 HSI 数据的光谱和空间的细节特征，在特征提取阶段更能关注到数据原始的特征信息，因此在分类指标上本章方法高出 7%左右。EndNet 是对特征分层提取，在特征融合阶段多次重建了多模态数据的输入，进而提升了模型分类效果，但是根源上还是没有解决 CNN 对光谱特征提取不充分的问题，对比看来本章方法所设计的光谱分支更好的关注到了 CNN 在 HSI 数据特征提取中的问题。HRWN 与本章方法同样关注到了以上问题，采取了双隧道的 CNN 进行 HSI 特征提取，在效果上有显著提升，其中有 6 类的分类准确率达到了最高，但是本章方法加入了残差模块，帮助模型更快收敛且避免梯度消失或者爆炸，保留和强调了光谱和空间特征中的重要信息，此外本章方法加入了 SNNM 和 LF，在特征融合阶段的融合策略和参数调整显得更合理，因此在指标表现上优于 HRWN 模型。



(a) (b) (c) (d) (e) (f) (g)

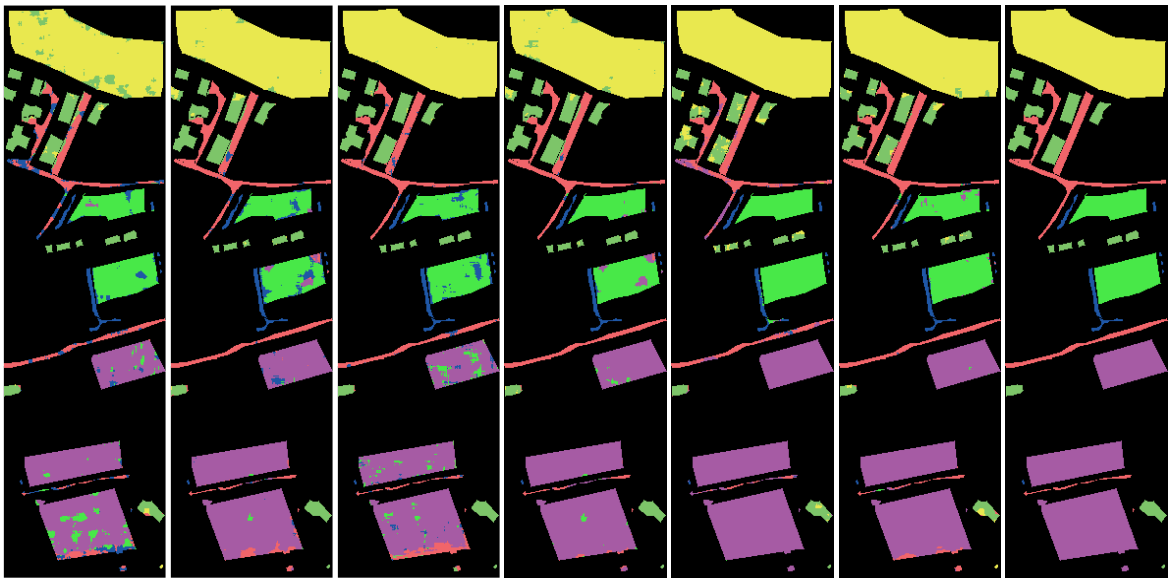
图2.12 DBSSRN 在 Houston2013 数据集上的可视化结果图

（2）Trento 数据集对比实验结果

本章方法与对比方法在 Trento 数据集上的性能结果对比如表 2.6 所示，训练样本是从表 2.3 中随机选择的，且训练数据集和测试数据集之间没有重叠，其对应的可视化结果图如 0 所示，其中（a）~（g）分别代表五个对比算法和本章的分类可视化结果图以及 Groundtruth 图。

表2.6 DBSSRN 在 Trento 数据集上的对比实验结果

No.	ELM	DeepCNN	FusAtNet	EndNet	HRWN	Ours
1	55.44±0.46	80.71±7.15	98.34±1.44	98.83±0.38	99.14±2.18	**99.30**±0.13
2	95.45±2.33	78.47±10.6	**100.00**±0.00	92.95±0.71	91.53±0.58	94.55±0.11
3	70.67±1.06	97.35±0.50	98.28±0.00	95.41±0.00	**99.41**±3.43	97.27±0.18
4	99.64±0.02	99.73±0.17	99.54±0.17	91.43±0.00	**99.90**±0.29	97.89±0.06
5	89.72±0.47	99.81±0.37	98.09±1.91	94.93±0.09	**99.31**±0.58	98.53±0.15
6	95.06±1.31	90.57±4.84	91.65±1.96	90.88±0.70	91.35±1.03	**97.76**±0.02
OA(%)	86.95±0.32	94.29±1.53	97.80±0.35	94.21±0.52	97.87±0.29	**98.02**±0.03
AA(%)	84.33±0.44	91.11±2.44	97.65±0.22	94.07±0.55	96.90±0.31	**97.55**±0.04
KC(%)	82.79±0.41	92.27±2.09	97.43±0.46	94.19±0.42	97.54±0.36	**98.01**±0.02



(a) (b) (c) (d) (e) (f) (g)

图2.13 DBSSRN 在 Trento 数据集上的可视化结果图

从表 2.6 中可以看出，在三个分类指标上本章方法都达到了最优，从数据中可以看出，传统的 ELM 方法对特征的细节关注不足，从可视化结果图也可看出，对于 6

个类别的细节区域很容易分类错误。HRWN 方法和本章方法因为都使用了针对光谱特征和空间特征的 HSI 特征提取网络，所以整体的分类性能都很好，HRWN 在 3 个类别上取得了分类性能最优，本章方法在 2 给类别上取得最优，但是 HRWN 在建筑类别和道路类别上的分类准确率很低，分别低于本章方法 3.02%和 6.41%，HRWN 的特征提取网络没有，加入残差网络结构，导致在面临复杂的特征信息时会没有有效收敛，此外本章加入的 LFM 部分辅助在特征融合阶段更好的利用了提取器提供的特征。结合数据和可视化图可以看出，本章的方法对 HSI 的光谱特征提取更加充分，SNN 的加入帮助了特征间的相关性联系，自适应的 LF 帮助特征融合时的结果更加适用于分类器。

2.4.6 消融实验结果与分析

本章主要利用消融实验来验证 HSI 和 LiDAR 两种模态数据融合对分类效果的影响，以及本章所设计的 DBSSRN 的空间和光谱两个分支的有效性、SNNM 和 LFM 这几个模块对分类效果的影响。故分为以下两部分进行消融实验分析：

（1）不同模态数据输入的消融分析

表2.7 DBSSRN 方法中不同模态数据在 Trento 数据集上的消融实验结果

数据源	Trento		
	OA(%)	AA(%)	KC(%)
HSI	95.33±0.05	94.76±0.03	94.59±0.01
LiDAR	97.30±0.01	97.18±0.03	96.37±0.02
HSI+LiDAR	**98.02**±0.03	**97.55**±0.04	**98.09**±0.02

表 2.7 展示了本章方法在 Trento 数据集上的不同模态数据间的分类效果，表中数据说明多模态方法对 HSI 和 LiDAR 的单模态分类性能都有提升，尤其对于数据维度高、信息量大的 HSI 数据，OA 提升了 2.69%、AA 提升了 2.79%、KC 提升了 3.5%，LiDAR 数据的高程信息对 HSI 数据起到了很好的补充作用，尤其在建筑物和道路等类别上。进一步证明本文采取的多模态遥感图像分类的有效性。

（2）不同模块的消融分析

在对 SNNM 和 LF 模块消融的时候特征提取网络使用了 DBSSRN，仅仅消融了 SNNM 和 LF，在没有 LF 时，采取直接拼接的简单特征融合方案，来验证 LF 对于特征融合的作用和有效性。对 DBSSRN 网络模型中的空间分支和光谱分支消融时，特征融合部分没有做任何处理，去掉了 SNNM 对特征间相似性和相关性的判断也去掉了 LF 对融合阶段权值训练和调整，目的是验证两个分支是否进一步提取到更值得关

注的特征，具体实验数据如表 2.8 所示，此处以在 Houston2013 数据集上的结果为例。

表 2.8 中数据表明，光谱分支为整个分类任务提供了较大的特征贡献，加入光谱分支提取后，OA 提升了 5.33%、AA 提升了 4.44%、KC 提升了 3.54%，可见本章针对 CNN 对于 HSI 数据的光谱特征提取不足缺点提出的 DBSSRN 中的两个分支的有效性。在没有空间分支情况下分类准确率高于没有光谱分支的原因是另一模态的 LiDAR 数据提供了一定的高程特征和空间特征信息。从对 SNNM 和 LF 的消融数据可知，每一模块对融合后的特征分类都有一定的提升，在两个联合使用的时候，特征融合阶段可以关注到两模态数据间的相关性，在预训练阶段更好的调整 LF，使得两个模态特征以更合理的权重融合。

表2.8 DBSSRN 方法中不同模块在 Houston2013 数据集上的消融实验结果

模块				Houston2013		
空间分支	光谱分支	SNNM	LF	OA(%)	AA(%)	KC(%)
✓				86.42±0.14	87.57±0.17	86.63±0.19
	✓			89.74±0.04	89.15±0.13	88.33±0.09
✓	✓			91.75±0.05	92.01±0.08	90.17±0.10
✓	✓	✓		92.79±0.85	93.29±0.12	92.23±0.16
✓	✓		✓	93.53±0.23	94.52±0.21	93.98±0.08
✓	✓	✓	✓	**95.23**±0.12	**96.36**±0.08	**95.07**±0.04

2.5 本章小结

本章提出了基于双分支空谱残差网络的多模态遥感图像分类算法，该算法由特征提取网络、孪生神经网络辅助预训练的 SNNM、LFM 的特征融合模块及分类器四部分组成，引入了 LiDAR 数据进行多模态遥感图像分类，经过消融实验分析发现，效果明显优于单一模态的分类效果；在特征提取阶段，针对 CNN 对 HSI 数据的光谱特征难提取充分的缺点设计了两条分支分别提取光谱特征和空间特征的 DBSSRN，提取后组合得到 HSI 模态数据的特征，丰富了整个特征提取阶段的特征信息；在特征融合阶段加入了 SNNM 和 LFM，使得融合的两组特征间的相关性更强，且通过对具体特征学习后的系数进行融合，整个过程更加合理和适应于对应任务。最终通过对比实验和消融实验证明，本章的方法能够很好的提取到光谱特征和空间特征，有效解决 CNN 对 HSI 数据提取的缺陷，结合了多模态数据后的分类指标有明显的提升，证明了多模态遥感图像分类方法的有效性，比所对比的其他方法表现出更好的分类性能。

第三章 基于双向交互融合的多模态遥感图像分类

3.1 引言

本文的第二章针对 HSI 单模态遥感数据特征信息有限引入 LiDAR 数据进行空间和高程特征补充，同时还针对 CNN 对 HSI 光谱特征提取不充分的问题提出 DBSSRN 进行对空间特征及光谱特征的双分支提取，但还存在一些因为 CNN 固有结构而存在的特征提取模块的问题及多模态之间交互的问题。如 CNN 虽然在图像分类任务中提取空间结构信息和局部上下文信息方面表现出了强大的能力，但是 CNN 对像 HSI 这样维度高、信息丰富的数据难以提取充分，且 CNN 对长距离像素和全局上下文信息的处理能力上，卷积操作感受野受限，难以关注到图像中的长距离依赖关系，CNN 很难很好地捕获像中长期依赖性的序列属性。另一方面，经典的 CNN 过度关注空间内容信息，这会在光谱上扭曲所学习特征中的顺序信息。这在很大程度上给挖掘诊断光谱属性带来了更多困难。对于多模态遥感图像分类任务，异构特征之间很难做到保持特征一致性和两种模态特征充分交互。

为解决上述问题，本章主要进行以下几项工作：

（1）针对 CNN 对 HSI 这样维度高、信息丰富的数据难以提取充分，对长距离像素和全局上下文信息的处理能力不足，难以关注到图像中的长距离依赖关系，CNN 很难很好地捕获像中长期依赖性的序列属性的问题，在两个 CNN 做特征提取之后进行特征融合的多模态遥感图像分类任务的基本思路上，将光谱特征和空间特征较为复杂的 HSI 支路的特征提取网络从 CNN 置换成 SpectralFormer，进而在关注到全局信息的同时，保持对光谱特征的充分提取。

（2）针对多模态遥感图像分类任务，异构特征之间很难做到保持特征一致性的问题，本章引入一致性损失来保证两个支路提取到的特征的一致性。

（3）针对多模态遥感图像分类任务中特征融合阶段，两种模态特征间没有充分交互的问题，本章引入交叉注意力机制，在一致性损失的基础上，特征能够在融合过程中进行交互和补充，提高分类的准确率。

3.2 相关理论基础

3.2.1 Transformer 的基本结构

Transformer 最初因为在 NLP 领域的显著效果而备受关注，后被 Niki 等人在 Image Transformer 中迁移到计算机视觉领域，且效果显著，随后出现的 Vision Transformer

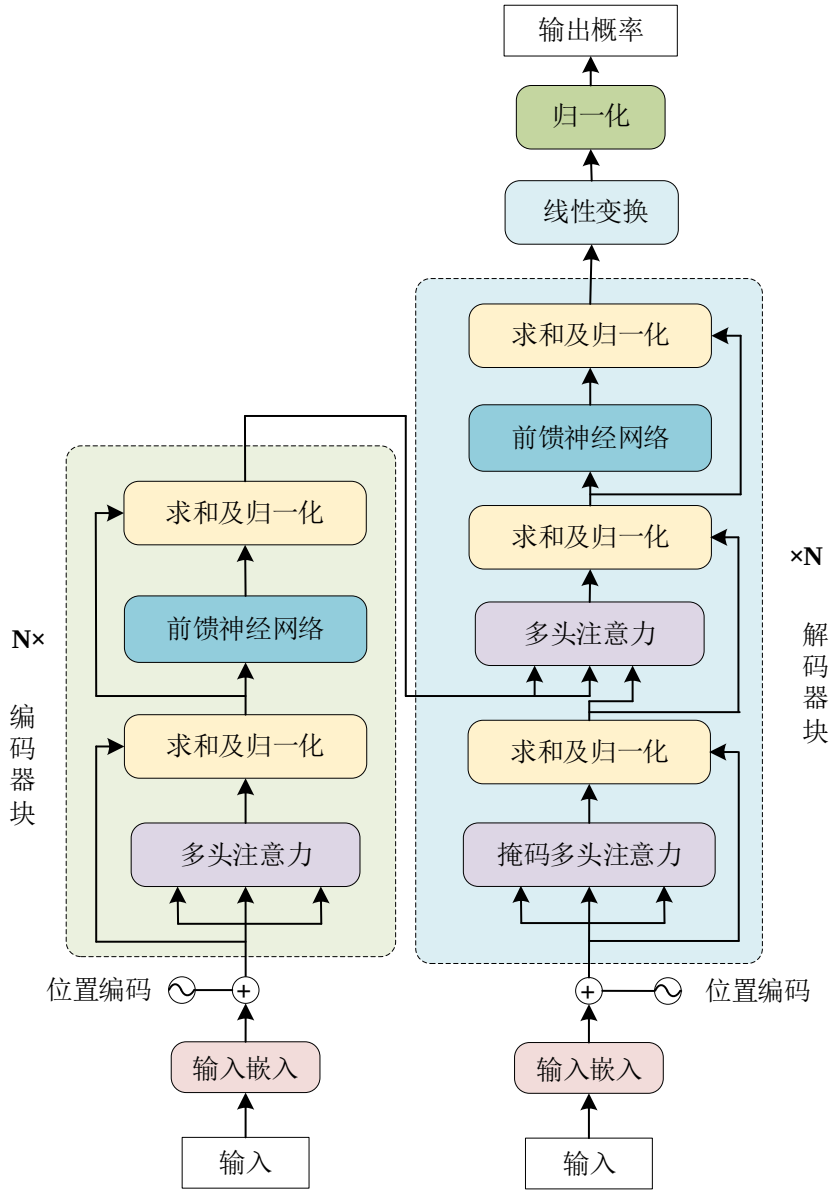


图3.1 Transformer 模型结构图

彻底奠定了 Transformer 在计算机视觉中的地位。原始的 Transformer 是一个基于自注意力机制的模型，一系列的标记被用作模型的输入，并且多头注意力被用来在输入字符序列中绘制全局相关性。

经典 Transformer 模型的结构图如图 3.1 所示，采用编码器-解码器结构，编码器和解码器都堆叠了 N 层，是避免循环的模型结构，输入的数据经过 N 层的编码器之后输出到每一层的解码器上计算注意力。编码器每层包含多头注意力层和前馈连接层，解码器有三个子层结构，掩码多头注意力层、多头注意力层和前馈连接层。为了能够更好的优化深度学习网络和避免针对序列不同位置关注度不同而导致特征向量的规模不一，整个网络使用了残差连接、映射到高维空间寻找分类面和归一化的手段。

3.2.2 注意力机制

注意力机制的灵感来源于人类的视觉注意力，即人们在观察图像时，往往会选择性地关注图像的某些部分，而忽略其他不重要的信息。在深度学习中，注意力机制可以被看作是一种权重分配策略，通过为输入数据的不同部分分配不同的权重，使得模型能够聚焦于对当前任务更为关键的信息。在传统的神经网络中，每个神经元的输出只依赖于前一层的所有神经元的输出，而在注意力机制中，每个神经元的输出不仅仅取决于前一层的所有神经元的输出，还可以根据输入数据的不同部分进行加权，即对不同部分赋予不同的权重。如图 3.2 是注意力机制的本质思想示意图。Q 、K 、V 分别代表查询向量（Query）、键向量（Key）和值向量（Value），Q 用于获取与其他向量的相关性，键向量 K 用于计算查询和值之间的相似性，值向量 V 包含了需要根据查询进行加权聚合的信息，最终输出重加权的值向量[75]。

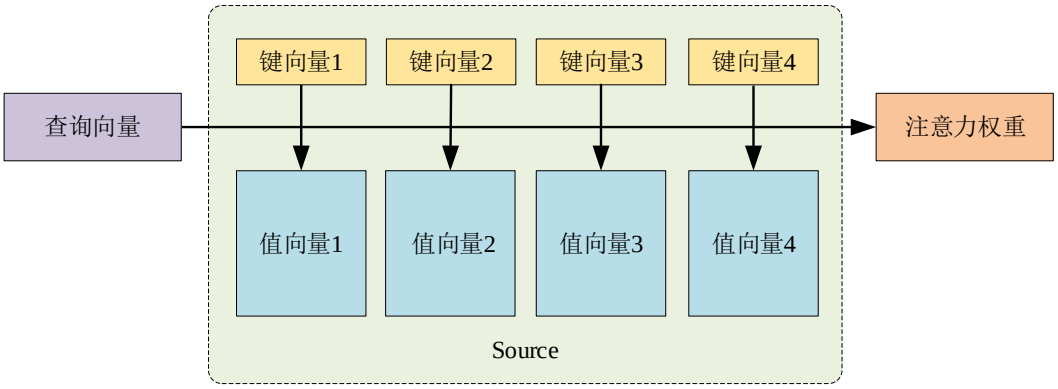


图3.2 注意力机制示意图

可以概括为下式：

$$\text{Attention}(Query, Source) = \sum_{i=1}^{L_x} \text{Similarity}\left(Query, Key_i\right) * Value_i, \tag{3-1}$$

其中，L_x 代表 source 的长度，抽象概括可以总结为两个过程，第一个过程是根据 Q 和 K 计算权重系数，其中包括根据 Q 和 K 计算两者的相似性或者相关性和对原始分值进行归一化处理。第二个过程是根据权重系数对 V 进行加权求和。两个过程对应的公式如下所示：

$$a_i = \text{softmax}\left(Sim_i\right) = \frac{e^{Sim_i}}{\sum_{j=1}^{L_x} e^{Sim_j}}, \tag{3-2}$$

$$\text{Attention}(Query, Source) = \sum_{i=1}^{L_x} a_i \cdot Value_i, \tag{3-3}$$

神经网络接收很多大小不一的向量输入，并且不同向量之间有一定的关系，然而在实际训练的时候没有充分利用这些输入之间的关系，进而会导致模型训练的效果不好，自注意力的引入主要就是为了解决这样的问题。Transformer 中的自注意力机制的公式如下：

$$\text{Attention}(Q,K,V)=\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V\ , \tag{3-4}$$

其中，Q 、K 、V 分别为查询矩阵、键矩阵和值矩阵，d_k 是它们对应的维数，$\sqrt{d_k}$ 是比例因子，用来对注意力矩阵进行归一化操作，避免 d_k 过大导致计算出的注意力结果过度膨胀。

这种方法的具体架构如图 3.4 所示，它允许同时处理多样化的信息，以获取更为丰富且深度的特性

多头注意力的核心在于通过对 Q 、K 、V 这三种参数进行多次分解，使得每个分解后的参数能够被映射至不同的子空间内并产生相应的注意力权重，这样就能有效地捕捉输入中的各个元素，结构如图 3.3 所示。多个头的结构可以关注到多种信息，以捕获更为丰富且深度的特征。

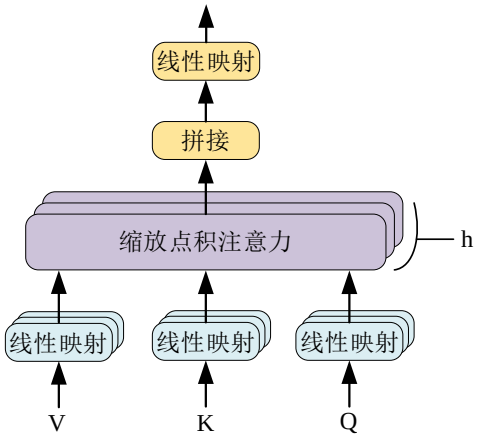


图3.3 多头注意力机制结构图

3.3 基于双向交互融合的多模态遥感图像分类算法

3.3.1 整体网络架构

为解决 CNN 对 HSI 这样维度高、信息丰富的数据难以提取充分及长距离像素和全局上下文信息的处理能力的不足的问题，CNN 在光谱上扭曲所学习特征中的顺序信息，及对于多模态遥感图像分类任务，异构特征之间很难做到保持特征一致性和两种模态特征充分交互的问题，本章提出的基于双向交互融合的多模态遥感图像分类的

网络整体结构如图 3.4 所示。网络模型主要包含双支路 CNN 与 Transformer 的特征提取网络、一致性损失、交叉注意力特征融合和分类网络四个部分。其中双支路 CNN 与 Transformer 的特征提取网络结合了 CNN 强大的特征提取能力及上下文建模能力和针对全局特征提取充分且对 HSI 数据复杂特殊的光谱信息敏感有效的 SpectralFormer，在 CNN 关注到 LiDAR 数据的高程信息、空间结构信息和局部上下文信息的同时，SpectralFormer 可以挖掘和表示 HSI 的光谱特征及其序列属性并提取到 HSI 的全局信息；一致性损失在这里既保证了双路特征的一致性，又提高了数据表示的稳定性和模型的鲁棒性；与此同时，在特征融合阶段加入交叉注意力机制，可以进一步自适应的融合 HSI 和 LiDAR 的异构特征；分类网络就采取简单的全连接网络进行分类了。

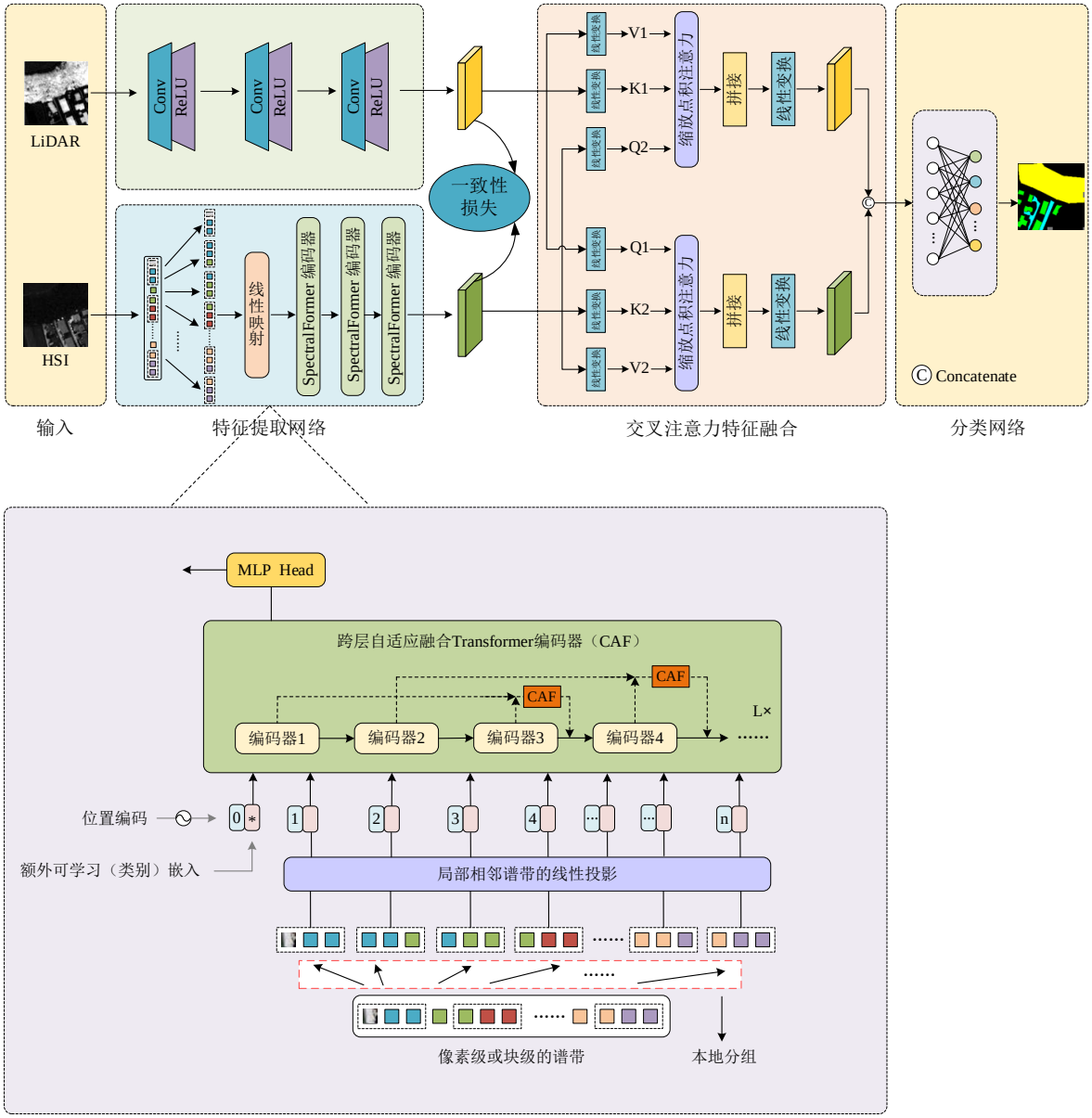


图3.4 基于双向交互融合的多模态遥感图像分类算法结构图

3.3.2 双支路 CNN 与 Transformer 的特征提取网络

（1）Transformer 支路

Transformer 是目前最先进的骨干网络之一，因为其核心是自注意力机制，后续基于 Transformer 提出的 ViT 更是适用于计算机视觉领域。本章在针对 HSI 数据的特征提取支路引入的 SpectralFormer 就是基于 ViT 设计的。SpectralFormer 更加关注 HSI 的光谱特征，使其非常适用于 HSI 的高精度和精细分类，SpectralFormer 中设计了分组谱嵌入（GSE）和跨层自适应融合（CAF）两个模块。

（a）GSE 模块

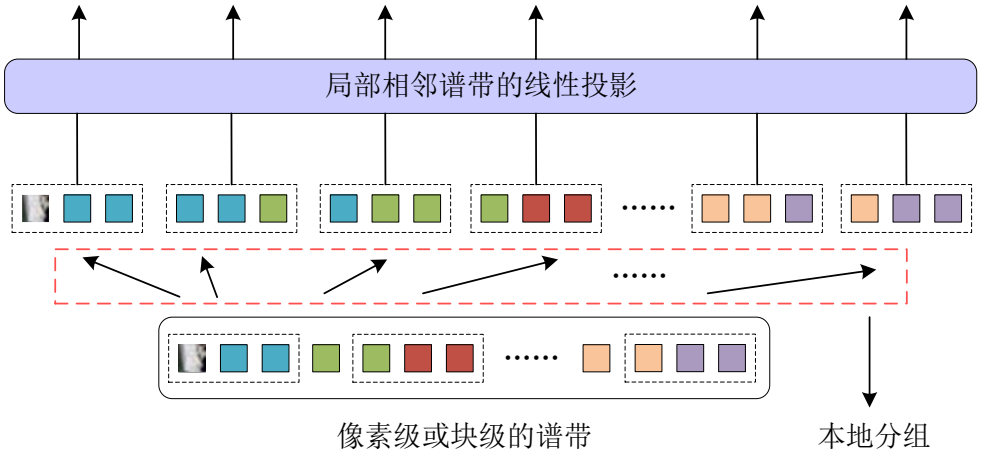


图3.5 分组谱嵌入 GSE 模块结构图

HSI 数据中不同位置的光谱信息反映不同波长的不同吸收特性，展示其对应的物理特性。捕捉这些光谱特征的局部吸收或变化是高光谱场景分类的关键因素，GSE 即是解决这样问题的模块，用来替代传统的基于单一波段输入和表示的方法。GSE 的结构如图 3.5 所示，和传统 Transformer 在此处的区别就是这里多了一个本地分组的操作，从局部光谱轮廓或相邻波段中学习特征嵌入。假设 $\boldsymbol{x}=\left[x_1,x_2,\ldots,x_m\right]\in\mathbb{R}^{\mathrm{I}\times m}$ 是一个光谱特征，即 HSI 中的一个像素，GSE 获得的特征嵌入如下式所示：

$$\dot{\boldsymbol{A}}=\boldsymbol{W}\boldsymbol{X}=\boldsymbol{W}g(\boldsymbol{x}), \tag{3-5}$$

其中，$\boldsymbol{W}\in\mathbb{R}^{d\times n}$ 和 $\boldsymbol{X}\in\mathbb{R}^{n\times m}$ 分别对应于变量 $\boldsymbol{w}$ 和 $\boldsymbol{x}$ 的分组表示，n 表示相邻波段的数量。变量 $\boldsymbol{W}$ 可以简单地看作是网络的一层，可以通过更新整个网络来进行优化。函数 $g(\cdot)$ 表示与变量 $\boldsymbol{x}$ 相关的重叠分组操作，即：

$$\boldsymbol{X}=g(\boldsymbol{x})=\left[\boldsymbol{x}_1,\ldots,\boldsymbol{x}_q,\ldots,\boldsymbol{x}_m\right], \tag{3-6}$$

其中，$\boldsymbol{x}^q=[x_{q-\lfloor(n/2)\rfloor},\ldots,x_q,\ldots,x_{q+\lfloor(n/2)\rfloor}]^{\top}\in\mathbb{R}^{n\times 1}$，$\lfloor\cdot\rfloor$表示四舍五入操作。

（b）CAF 模块

Transformer 中的残差连接（SC）可以增强各层之间的信息交换，减少网络学习过程中的信息损失，但是也会造成高低层特征差距大、特征融合不充分等问题。而所提出的 CAF 可以自适应学习跨层特征融合，从浅层到深层传递特征信息，CAF 的结构图如图 3.6 所示：

如果$\boldsymbol{z}^{(l-2)} \in \mathbb{R}^{1\times d_z}$和$\boldsymbol{z}^{(l)} \in \mathbb{R}^{1\times d_z}$表示第$(l-2)$层和第$l$层的输出，则 CAF 的公式表示如下：

$$\hat{\boldsymbol{z}}^{(l)} \leftarrow \ddot{\boldsymbol{w}}\begin{bmatrix} z^{(l)} \\ z^{(l-2)} \end{bmatrix}, \tag{3-7}$$

其中，$\hat{\boldsymbol{z}}^{(l)}$表示使用 CAF 得到的第$l$层的融合表示，$\ddot{\boldsymbol{w}} \in \mathbb{R}^{1\times 2}$是用于自适应融合的可学习网络参数，另外 CAF 是只跳过一个编码器，这也是为了契合 HSI 的数据特点。

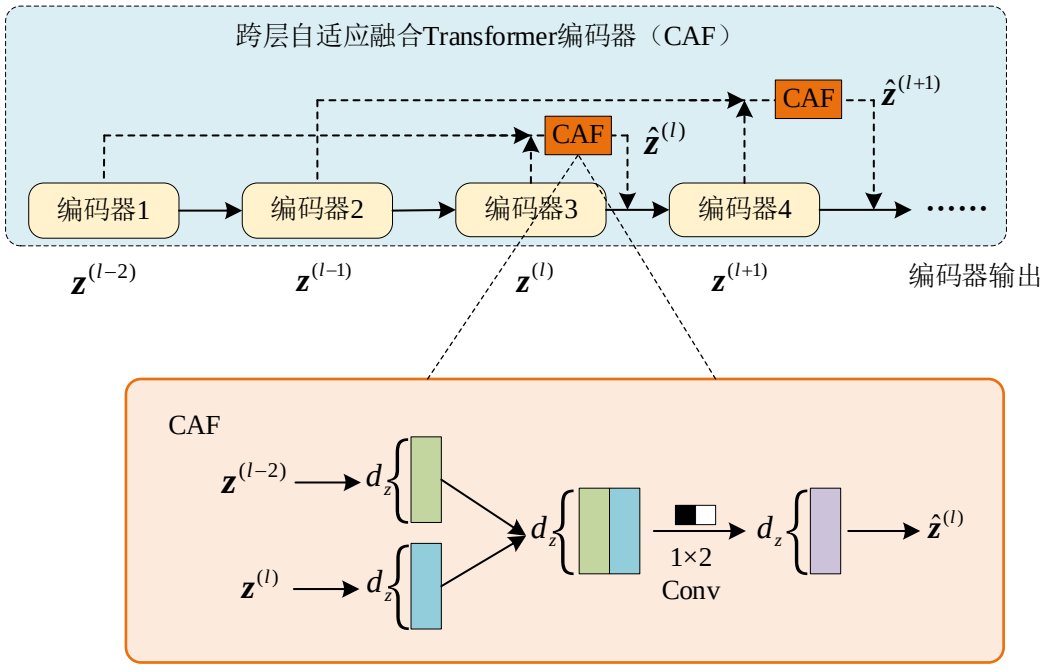


图3.6 跨层自适应融合 CAF 模块结构图

（2）CNN 支路

提取 LiDAR 数据特征的支路使用的仍为第二章所使用的 Resnet18，具体的网络模型参数结构见 2.3.2 的表 2.1 所示。

3.3.3 交叉注意力模块

在特征融合阶段，加入交叉注意力模块主要是为了进行 HSI 特征和 LiDAR 的特征的交互。首先，本章将双支路提取到的 HSI 和 LiDAR 数据特征 A_1、A_2 进行 Linear

线性变换，生成查询特征向量Q_1和Q_2、键特征向量K_1和K_2值特征向量V_1和V_2。本章将两个模态遥感图像数据得到的查询特征向量Q_1和Q_2进行交叉，构造交叉注意力机制。具体实现过程如下面公式所示：

$$K_1 = A_1 W^{K_1}，\ K_2 = A_2 W^{K_2}， \tag{3-8}$$

$$V_1 = A_1 W^{V_1}，\ V_2 = A_2 W^{V_2}， \tag{3-9}$$

$$Q_1 = A_2 W^{Q_1}，\ Q_2 = A_1 W^{Q_2}， \tag{3-10}$$

其中，W^{K_i}, W^{V_i}和W^{Q_i}（$i=1,2$）都是可学习的参数，在本章采用了一个简单的线性层实现。然后计算不同模态数据间的交叉注意力，公式如下：

$$A^1 = \text{softmax}\left(\frac{Q_2 K_1^{\ T}}{\sqrt{D}}\right) V_1， \tag{3-11}$$

$$A^2 = \text{softmax}\left(\frac{Q_1 K_2^{\ T}}{\sqrt{D}}\right) V_2， \tag{3-12}$$

其中，A^1和A^2表示经过交叉注意力机制引导后的特征。然后，将交叉注意力后得到的特征进行拼接得到一个更加稳健和鲁棒的特征表征，由以下公式得到：

$$A = \left[A^1, A^2\right]， \tag{3-13}$$

3.3.4 一致性损失模块

一致性损失函数的主要目的是确保模型在面对数据的变化或扰动时，其输出能够保持一致。这种损失函数在计算机视觉和深度学习的多个领域中都有应用，特别是在无监督学习、数据增强和域适应等任务中。一致性损失通常被定义为模型在原始输入和变换后输入上的输出差异。这种差异可以通过各种方式来衡量，例如使用 L1 范数、L2 范数或者更复杂的损失函数。具体的损失函数形式会根据应用任务、模型架构和数据特性等因素进行调整。

本章使用的一致性损失函数是均方误差（Mean Squared Error，MSE）损失函数，主要作用是使得模型更加稳定，模型能够充分利用 HSI 和 LiDAR 两种模态的互补信息，并进行双向特征交互，从而提高分类的准确性。即使其中一种模态的数据存在噪声或缺失，另一种模态的信息也可以在一定程度上弥补这一不足；通过优化一致性损失，模型能够学习到如何将不同模态的特征有效地结合起来，从而充分利用多模态数据提供的丰富信息。MSE 损失的函数如下：

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2, \tag{3-14}$$

其中，$\hat{Y}$ 可以看作是模态一的特征，Y 为模态二的特征。

MSE 计算 HSI 特征和 LiDAR 特征对应的输出之间的差异，量化两个模态在分类任务上的不一致性。计算出的损失值用于反向传播算法中，通过梯度下降或其变种优化模型的权重和参数。随着训练的进行，模型会根据 MSE 损失来调整内部参数，以最小化 HSI 和 LiDAR 输出之间的差异，同时提高分类准确性。

3.4 实验与分析

3.4.1 实验环境与参数设定

（1）本章实验环境详细配置如 2.4.3 小节的表 2.4 所示。

（2）本章的参数设定如下：

为保证实验公平性，对实验过程中的可变因素进行定量控制，在本章所有使用的对比实验和消融实验中均采用以下配置：评价指标方面使用与上一章相同的总体正确率 OA、平均准确率 AA 和 KC。

本章算法在 Houston2013、Trento 数据集上进行性能评估，数据集设置信息同 2.4.1 小节。训练过程批大小设定为 128；采用的优化器为 Adam，权重衰减参数为 0.9；初始学习率设定为 $5e^{-3}$，而后采用线性衰减策略对学习率进行衰减；最大迭代次数为 100 代；使用了交叉熵损失作为损失函数。每个实验重复 10 次取平均值和标准偏差。

3.4.2 对比实验结果与分析

本章的对比实验所使用的方法为 2.4.4 小节所提及的：ELM、DeepCNN、FusAtNet、EndNet、HRWN 和第二章的 DBSSRN 六种方法，且六种方法均采用以上介绍的同等实验参数设置。

（1）Houston2013 数据集对比实验结果

本章方法与对比方法的性能结果对比如表 3.1 所示，训练样本是从表 2.2 中随机选择的，并且训练数据集和测试数据集之间没有重叠，其对应的可视化结果图如图 3.7 所示，其中（a）~（h）分别代表 ELM、DeepCNN、FusAtNet、EndNet、HRWN、DBSSRN、本章方法的分类可视化结果图和 Groundtruth 图。

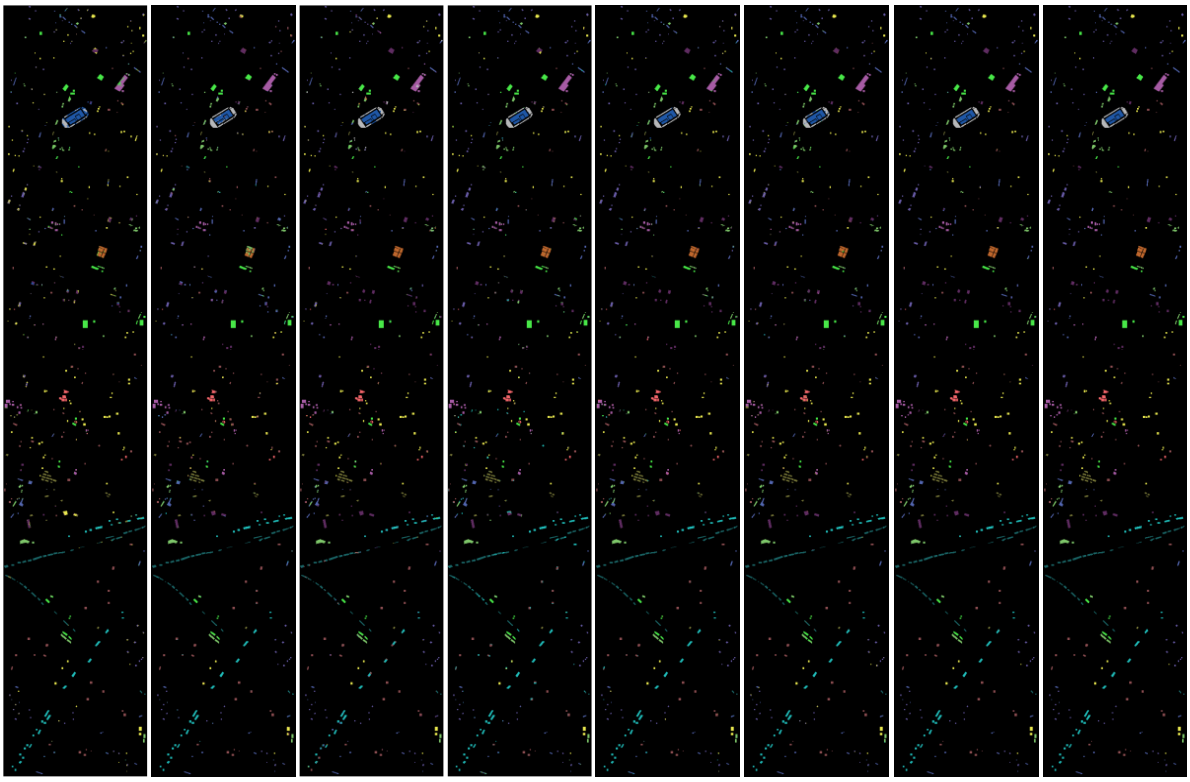
从表 3.1 中可以看出，本章算法在 6 个类别中产生了最高的分类准确率，且 OA、AA、KC 指标均超过对比方法结果达到最好，其中 OA 比分类效果第二好的 HRWN

高 1.79%，AA 高出 1.72%，KC 高出 1.85%。与使用两个 CNN 对 HSI 和 LiDAR 进行特征提取的 DeepCNN 相比，OA 提升了约 10%，LiDAR 支路的特征提取本章也是使用了 CNN 进行提取，这点上与 DeepCNN 并无过大区别，但是本章方法加入了 SpectralFormer 对 HSI 的光谱信息进行了充分提取且更加关注了全局特征信息，同时本章加入的交叉注意力模块与一致性损失，所以特征间的交互较为充分且能够保障两模态数据间的特征一致性，指标的提升也证明了其有效性。此外，对比于 FusAtNet，三个评价指标也有较大的提升，FusAtNet 与本章方法都使用了交叉注意力来增强样本间的特征交互，但是 FusAtNet 没有关注到两个模态的特征间的一致性，在有噪声或者数据扰动时，不能保证两种模态在一定程度上互相弥补，从而导致融合阶段效果不理想，指标差于本章方法。HRWN 和 DBSSRN 中针对 HSI 的光谱和空间特征分别设计了两个分支进行特征提取，与本章的 SpectralFormer 一样都关注到了数据的光谱特征，但是对于模态间的特征一致性和特征间的交互部分还是有不足之处，具体可参

表3.1 本章算法在 Houston2013 数据集上的对比实验结果

No.	ELM	DeepCNN	FusAtNet	EndNet	HRWN	DBSSRN	Ours
1	90.51±4.72	93.31±4.12	63.64±31.8	96.24±2.55	96.89±1.92	**97.01**±0.55	96.90±0.44
2	94.98±1.44	95.77±0.84	**98.05**±2.36	93.46±3.10	97.29±2.82	95.24±0.33	91.82±0.16
3	75.45±1.35	85.46±0.09	98.26±2.20	96.44±5.08	**99.64**±0.62	97.43±0.07	98.97±0.23
4	97.58±0.19	96.53±4.22	95.90±1.76	**98.51**±0.75	97.22±2.56	93.22±0.37	93.62±0.50
5	95.91±0.40	96.71±0.08	98.42±1.52	96.25±1.55	**99.45**±0.28	97.74±0.12	98.54±0.13
6	77.46±13.1	77.46±1.01	87.94±8.49	96.53±5.60	**100.00**±0.00	**100.00**±0.00	**100.00**±0.00
7	86.47±1.20	88.99±3.73	93.67±1.37	98.03±0.61	96.47±1.58	**98.55**±0.16	96.43±0.17
8	89.51±4.59	91.47±1.81	86.84±4.75	**95.53**±3.62	87.48±3.09	95.26±0.17	94.70±0.40
9	69.13±1.26	86.11±4.36	91.42±6.61	79.80±5.57	91.26±1.98	92.13±0.05	**94.07**±0.14
10	71.39±5.78	79.44±0.51	80.86±13.7	77.53±5.90	94.19±3.59	95.74±0.03	**98.89**±0.10
11	66.45±3.75	66.45±0.48	91.40±3.29	86.40±4.57	91.97±2.19	93.89±0.11	**98.67**±0.06
12	78.36±3.46	78.46±0.68	68.76±17.3	87.02±4.51	**96.77**±0.89	94.73±0.06	94.49±0.22
13	75.50±13.3	74.34±2.54	95.61±3.84	79.64±20.0	91.63±0.98	94.46±0.22	**100.00**±0.00
14	93.97±0.97	94.55±2.10	86.47±25.0	98.71±0.64	**100.00**±0.00	**100.00**±0.00	**100.00**±0.00
15	98.73±0.46	97.71±0.31	98.42±1.32	99.38±0.76	**100.00**±0.00	**100.00**±0.00	98.99±0.08
OA(%)	83.67±0.81	86.93±0.73	87.97±2.94	90.77±1.77	94.50±0.81	95.23±0.12	**96.29**±0.02
AA(%)	84.09±0.88	86.85±0.44	89.04±1.95	91.96±2.10	95.35±0.44	96.36±0.08	**97.07**±0.02
KC(%)	82.28±0.88	86.66±0.79	86.90±3.21	89.98±1.92	94.12±0.65	95.07±0.04	**95.97**±0.02

照表 3.3 结果。但是 HRWN 中也是有着 6 个类别的方法分类准确率达到最高，分别是 3（人造草地）、5（土壤）、6（水）、12（停车场 1）、15（跑道）这 5 个类别，其中停车场 1 和跑道的效果与本章高出稍多，可见 CNN 对于局部的特征关注程度确实高于 Transformer，Transformer 在关注全局特征的同时容易忽略局部特征。



(a) (b) (c) (d) (e) (f) (g) (h)

图3.7 在 Houston2013 数据集上的可视化结果图

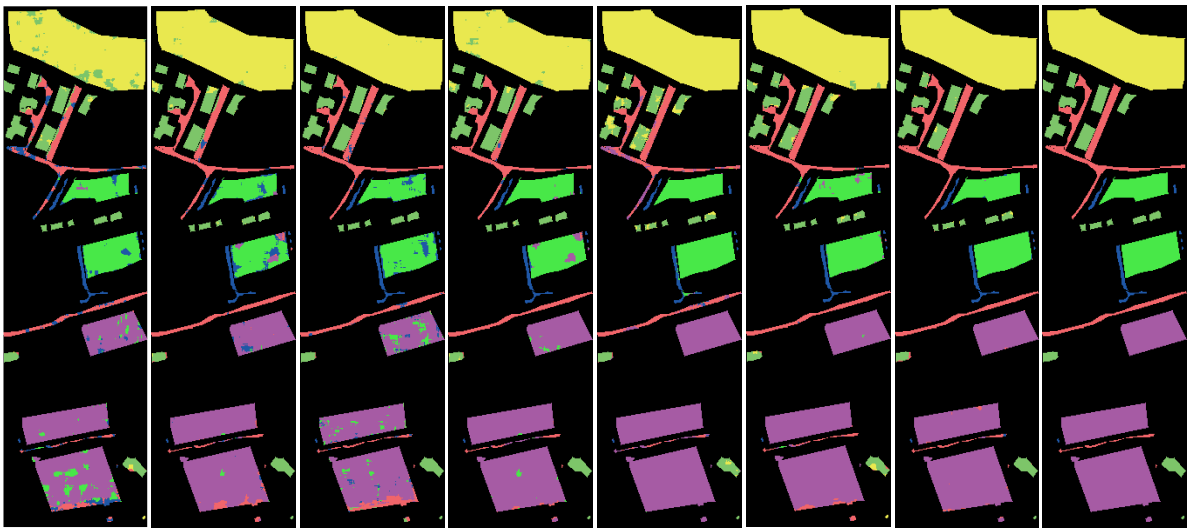
（2）Trento 数据集对比实验结果

本章方法与对比方法的性能结果对比如表 3.2 所示，训练样本是从表 2.3 中随机选择的，并且训练数据集和测试数据集之间没有重叠，其对应的可视化结果图如图 3.8 所示，其中（a）~（h）分别代表六个对比算法和本章的分类可视化结果图以及 Groundtruth 图。

从表 3.2 可以看出，在 OA、AA 和 KC 三个指标上，本章的方法均是最高。但是在各类别分类准确率中仅有两个达到了最高，而使用两个 CNN 并且关注到光谱特征的 HRWN 有三个类别的准确率达到了最高，同类型的 DBSSRN 只有一个类别达到了最高，再次证明了 CNN 对局部特征的关注比较充分。但是综合 6 个对比方法的分类效果来看，增加了特征交互融合的 FusAtNet 效果较于其他的效果更好一些，横向对比可以看出，本章的方法在进行了特征双向交互融合之后（同时也更多的关注到光谱和全局特征），鲁棒性更优，具备更好的分类性能。

表3.2本章算法在 Trento 数据集上的对比实验结果

No.	ELM	DeepCNN	FusAtNet	EndNet	HRWN	DBSSRN	Ours
1	55.44±0.46	80.71±7.15	98.34±1.44	98.83±0.38	99.14±2.18	**99.30**±0.13	97.00±0.11
2	95.45±2.33	78.47±10.6	**100.00**±0.00	92.95±0.71	91.53±0.58	94.55±0.11	95.54±0.18
3	70.67±1.06	97.35±0.50	98.28±0.00	95.41±0.00	**99.41**±3.43	97.27±0.18	96.58±0.11
4	99.64±0.02	99.73±0.17	99.54±0.17	91.43±0.00	**99.90**±0.29	97.89±0.06	98.89±0.05
5	89.72±0.47	99.81±0.37	98.09±1.91	94.93±0.09	99.31±0.58	98.53±0.15	**99.88**±0.03
6	95.06±1.31	90.57±4.84	91.65±1.96	90.88±0.70	91.35±1.03	97.76±0.02	**98.95**±0.05
OA(%)	86.95±0.32	94.29±1.53	97.80±0.35	94.21±0.52	97.87±0.29	98.02±0.03	**98.65**±0.02
AA(%)	84.33±0.44	91.11±2.44	97.65±0.22	94.07±0.55	96.90±0.31	97.55±0.04	**97.81**±0.01
KC(%)	82.79±0.41	92.27±2.09	97.43±0.46	94.19±0.42	97.54±0.36	98.01±0.02	**98.19**±0.03



(a) (b) (c) (d) (e) (f) (g) (h)

图3.8 在 Trento 数据集上的可视化结果图

3.4.3 消融实验结果与分析

本章主要利用消融实验来验证 HSI 和 LiDAR 两种模态数据进行的多模态融合对分类效果的影响，以及本章所设计的一致性损失模块 MSE 和交叉注意力模块 CAM 对分类效果的影响。故分为以下两部分进行消融实验分析：

（1）不同模态数据输入的消融分析

表 3.3 展示了本章对于数据集的消融实验结果，具体类别中的准确率在这里就不详细展开了。

表3.3 不同模态数据在 Trento 数据集上的消融实验结果

数据源	Trento		
	OA(%)	AA(%)	KC(%)
HSI	93.72±0.01	92.56±0.05	91.56±0.01
LiDAR	97.30±0.00	97.20±0.05	96.38±0.01
HSI+LiDAR	**98.65**±0.02	**97.81**±0.01	**98.19**±0.03

由表 3.3 可以看出，相较于在两种模态数据上各自单独进行分类任务时的准确率，融合后的多模态方法有着明显的提高，尤其是对于 HSI 数据的分类结果，各指标提升了 1%~7%不等，足以证明对于高程信息和空间结构特征的补充，弥补了 HSI 单模态分类中的特征缺失。

（2）不同模块的消融分析

表 3.4 针对本章提出的算法模型中的部分模块（MSE、CAM）进行了消融分析，来进一步证明本章所加入的 MSE 和 CAM 对整体分类任务的正向作用。

表3.4 不同模块在 Houston2013 数据集上的消融实验结果

模块		Houston2013		
MSE	CAM	OA(%)	AA(%)	KC(%)
		87.35±2.24	88.13±1.02	86.47±0.88
✓		93.37±0.32	94.79±0.36	92.56±0.16
	✓	94.63±0.22	95.84±0.18	94.17±0.24
✓	✓	**96.29**±0.02	**97.07**±0.02	**95.97**±0.02

由表 3.4 可以看出，当不考虑两个模态数据的特征一致性和特征交互时，直接把特征提取网络输出的两个模态特征拼接时，整体的分类准确率是很低的，主要原因也如本章在 3.1 中分析的，当不考虑两个模态数据的特征间相关性时，LiDAR 中的高程信息和空间特征信息无法准确的相关到 HSI 数据的特征，从而无法实现异构特征间的互补，对应到表中的数据则为当加入 MSE 模块后，本章算法在 Houston2013 数据集上的分类性能明显提升。在只有 CAM 模块时，两个模态间虽然没有通过一致性损失关注到异构特征间的相关性，但是两模态特征间进行了特征的交互，因此，两个模态特征可以进行理解，并且进一步的自适应融合异构特征，由结果可以看出 CAM 模块的有效性。当然，在充分考虑到异构数据间的双向交互融合后，对比于单一的 MSE 和 CAM 模块，这里的效果提升也比较明显，OA、AA、KC 指标基本能提升 1%~2%。

3.5 本章小结

本章提出了基于双向交互融合的多模态遥感图像分类算法，该算法由特征提取网络、一致性损失、基于交叉注意力的特征融合模块及分类器四部分组成，首先在特征提取阶段用 Resnet18 对 LiDAR 数据提取，HSI 支路使用能够充分提取复杂光谱序列信息和全局特征的 SpectralFormer 进行提取，获得两种模态的特征，并将其输入到交叉注意力模块进行模态间的特征交互融合。最后输入分类器进行分类，预训练过程中使用一致性损失持续反向传播两模态特征间的异构特征信息，保障其一致性与相关性。最终通过对比实验和消融实验证明，本章的方法能够充分提取到 HSI 数据的丰富信息、保证两模态特征间的一致性和充分交互，比所对比的其他方法表现出更好的分类性能。

第四章　基于对比学习的多模态遥感图像分类

4.1 引言

本文的第一章和第二章针对特定模态数据特征提取问题和模态间交互问题，在CNN 层面和 Transformer 层面针对性的进行了系列研究并提出相应的解决方案，虽然已经在结果上取得了不错的效果，但是仍然存在一些不足。因为以上的方法都是监督学习方法，而遥感图像数据的标注样本量很少，且标注成本高，很多依赖大量数据的深度学习方法在遥感图像分类任务中很难在有限的实例下达到更高的性能水平。随着对比学习在计算机视觉领域的成功应用，自监督对比学习已被证明在解决这一问题上具有显著的效果。虽然最近的一些研究已经探讨了对比学习在遥感图像分类中的应用，但现有基于对比学习的方法主要通过一组随机的数据增强实现不变性来促进单一模态遥感数据的表示学习。这些方法没有充分考虑遥感图像的异构特征表示，而这些特征表示可以提供更加全面的信息来描述地面特性并改善分类性能，即，可以考虑采取对比学习与多模态遥感图像分类任务相结合，进一步解决遥感图像分类任务的性能进一步提升。

针对遥感图像数据的标注样本量稀缺、标记成本高的问题，引入 SSL 中的对比学习，并且应用到多模态遥感图像分类任务中，在单一模态内利用未标注数据中的内在结构和关系来学习有用的特征表示，帮助缓解数据标注问题，降低对标注数据的依赖；在模态间通过对比学习，在预训练期间更好地挖掘多模态特征表示，同时在模态内部和模态之间保持语义一致。

4.2 自监督学习与对比学习

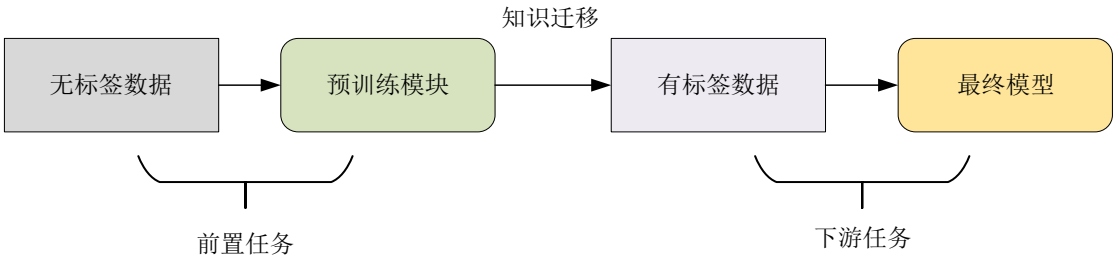


图4.1 SSL 算法示意图

SSL 通过无监督方式理解提供的未标记数据，生成监督标签，然后在下一次迭代中使用其中高置信度的数据标签，通过反向传播等方法像任何其他监督学习模型一样对模型进行训练。每次迭代使用的数据标签都是不同的。SSL 的结构图如图 4.1 所示。

SSL 最广泛用于解决计算机视觉问题，如图像分类、目标检测、语义分割或实例分割。

SSL 任务分为两类：pretext 任务和下游任务。前者采用无监督学习来学习表示，标签是从数据本身生成的。当学习完成后，模型将先前学习的表示应用到后续任务中。图 4.2 中的(a)和(b)分别描述了 pretext 任务和下游任务执行的各种任务。

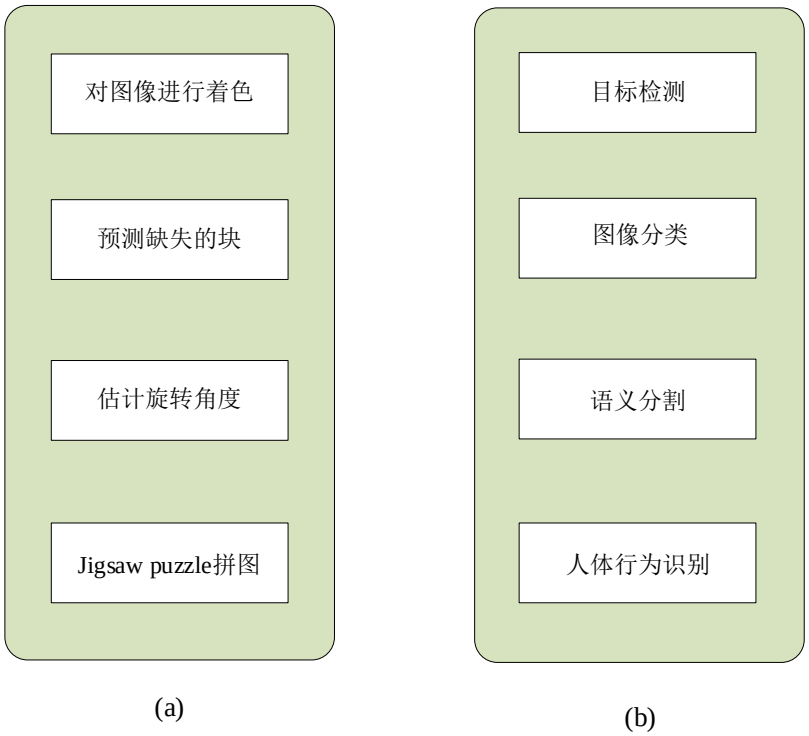


图4.2 SSL 任务分类：(a) pretext 执行的各种任务，(b)下游执行的各种任务

在 pretext 任务中，使用可见部分来预测数据的隐藏部分。pretext 任务可以应用于任何类型的数据，例如图像、音频、视频等。这项任务允许机器直接从数据中获得监督，而无需使用标签，从而实现自动学习。设计合适的 pretext 任务需要领域知识。

下游任务是定义模型目的的主要任务。pretext 任务，也称为辅助任务，允许模型学习用于完成下游任务的有用特征表示信息。为了确保下游任务的质量，应该计算在 pretext 任务中学习到的特征表示。下游的主要任务是在数据标签不足的情况下执行分类。下游任务可以通过两种方式完成：微调或使用线性分类器。为了获得良好的性能，下游任务需要少量的数据标记。当自监督预训练与下游任务之间的领域差距较小时，下游任务的性能通常会更好。

对比学习（Contrastive Learning，CL）是 SSL 的一个分支，专注于学习同一类实例之间的共同特征并区分不同类实例之间的差异，通过自动构造相似实例和不相似实例，学习一个表示学习模型，通过这个模型，使得相似的实例在投影空间中比较接近，而不相似的实例在投影空间中距离比较远。对比学习的基本思路是给定数据 ，对比学习的目标是学习一个编码器 使得：

$$score\left(f(x), f\left(x^{+}\right)\right) >> score\left(f(x), f\left(x^{-}\right)\right), \tag{4-1}$$

其中，x被称为针点数据，x^{+}是和x相似的正样本，x^{-}是和x不相似的负样本，$score()$是一个度量函数，来衡量正负样本的相似度，经常采用欧氏距离、余弦相似度等。为了优化编码器$f(\cdot)$，对比学习一般构造 softmax 分类器对正样本和负样本进行分类，损失函数此处介绍 Info NCELoss。

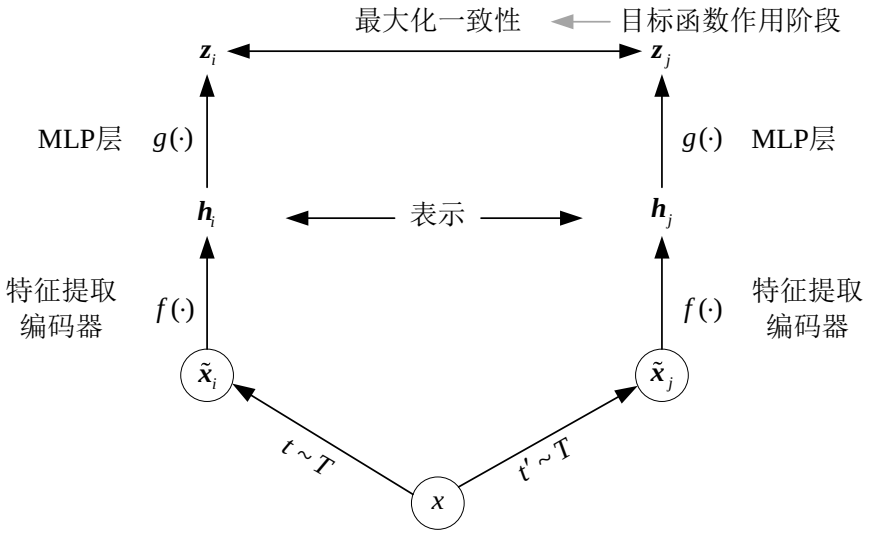


图4.3 SimCLR 示意图

以 SimCLR[80]为例，说明一下对比学习的结构，结构图如图 4.3 所示：

函数$f(\cdot)$：在编码得到表示后通过 MLP 将表示映射到对比学习损失的空间，目标是希望同一张图片的不同 augmentation 表示相近，与 mini-batch 中其他图片的 augmentation 表示较远。

（1）给定输入的锚点数据为$\boldsymbol{X}$，首先通过数据增强（随机裁剪、颜色失真、高斯模糊等）生成正负样本对$\tilde{\boldsymbol{x}}_i$和$\tilde{\boldsymbol{x}}_j$。

（2）特征提取编码器。函数$f(\cdot)$就是一个编码器，CL 中用什么编码器不做限制，SimCLR 中使用的是 ResNet50，$\tilde{\boldsymbol{x}}_i$和$\tilde{\boldsymbol{x}}_j$通过分别得到$\boldsymbol{h}_i$和$\boldsymbol{h}_j$。

（3）MLP 层。通过特征提取之后，进入 MLP 层，MLP 层的输出就是对比学习的目标函数作用的地方，通过 MLP 层输出$\boldsymbol{z}_i$和$\boldsymbol{z}_j$。

（4）目标函数作用阶段。对比学习中的损失函数一般是 infoNCE Loss，$\boldsymbol{z}_i$和$\boldsymbol{z}_j$的损失函数定义如下：

$$l_{i,j}=-\log\frac{\exp\left(sim\left(z_i,z_j\right)/\tau\right)}{\sum_{k=1}^{2N}\Gamma_{[k\neq i]}\exp\left(sim\left(z_i,z_k\right)/\tau\right)},\tag{4-2}$$

其中，N 代表的是一个 batch 的样本数，即对于一个 batch 的 N 个样本，通过数据增强的得到 N 对正样本对，此时共有 $2N$ 个样本。对于一个给定的正样本对，剩下的 $2(N-1)$ 个样本都是负样本，即负样本都基于这个 batch 的数据生成。上式中 $sim\left(z_i,z_k\right)$ 其实就是余弦相似度，其计算公式如下所示。$\Gamma_{[k\neq i]}$ 输入 0 或 1，当 k 不等于 i 时，结果就为 1 否则为 0，τ 是温度系数。

$$sim(u,v)=u^{T}v/\|u\|\cdot\|v\|,\tag{4-3}$$

从公式(4-2)可以看出，分子中只计算正样本对的距离，负样本只会在对比损失的分母中出现，当正样本对距离越小，负样本对距离越大，损失越小。

4.3 基于对比学习的多模态遥感图像分类算法

4.3.1 整体网络架构

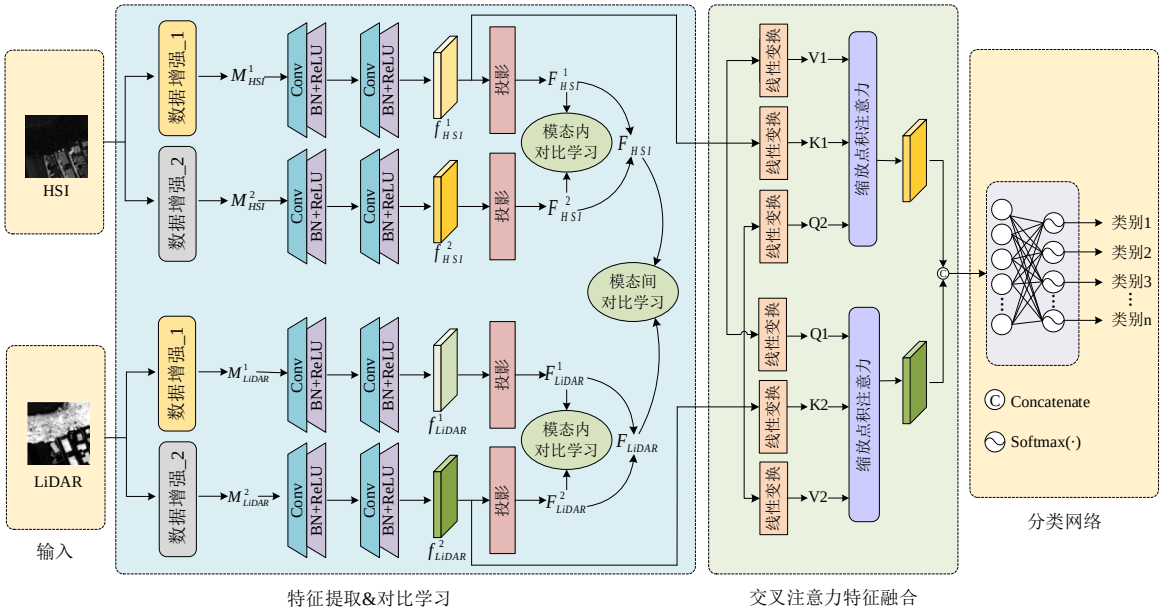


图4.4 基于对比学习的多模态遥感图像分类算法结构图

为解决遥感图像数据的标注样本量稀缺、标记成本高的问题，本章引入自监督学习中的对比学习，所实现的基于对比学习的多模态遥感图像分类网络整体结构图如图 4.4 所示，网络模型主要包含数据增强、CNN 特征提取、对比学习预训练、交叉注意

力辅助特征融合和分类网络五个部分。其中数据增强结合大量未标记样本构建正负样本对，为对比学习提供数据基础。特征提取所使用的 CNN 比较简略，降低了模型参数量。主要集中在通过对比学习实现对模型的预训练上，使用对比损失分别对两个模态内及两模态间的数据特征进行对比学习预训练。后续的特征融合阶段和分类阶段同第三章使用的结构一样，详情见 3.3.2 小节。

4.3.2 数据增强

数据增强是对比学习算法中的重要一环，对比学习的标签信息来源于数据本身的，核心模块便是数据增强。有效的数据增强操作可以在保留原始信息的同时，增加样本之间的差异性，使模型更容易区分不同类别的样本，从而提高分类准确率。因此本章设计了以下两种数据增强方法：

（1）添加噪声

通过添加随机高斯噪声构建参与训练的正负样本对，均值为零的高斯噪声可以在一定程度上扭曲原始的高频特征，添加适当的噪声可以增强网络的学习能力。生成增强数据的方式如下式所示：

$$X' = \frac{1}{\alpha_1 + \alpha_2}\left(\alpha_1 X + \alpha_2 X\right) + \beta N \text{ ,} \tag{4-4}$$

其中，X 是需要增强的数据，α_1 和 α_2 是 0.01 到 1 范围内均匀分布的随机数，N 是服从标准正态分布的随机高斯噪声，β 设置为了 1/25。

（2）数据旋转

基于（1）中得到的噪声数据 X'，本章进一步进行旋转操作，随机旋转 θ 角度，生成 X''，其中 $\theta \in [0, \pi/2]$。原始数据 X 、X' 和 X'' 可以共同构建正负样本。

4.3.3 特征提取器

表4.1 特征提取器参数表

名称	输出尺寸	卷积核大小	卷积核个数
Conv1	11*11	3*3	32 个
Conv2	11*11	3*3	64 个
Conv3	11*11	3*3	128 个

在特征提取阶段，为减小网络参数，将模型轻量化，本章将以上两个工作点中的 CNN 或 Transformer 简化为具有三层相同结构的卷积网络，图 4.4 绘出两组卷积结构，

每个卷积层使用了 3×3 卷积核，卷积核后连接 BN 和 ReLU，网络设置如表 4.1 所示。

4.3.4 模态内对比学习

从大的方向看，整个算法分为两个阶段：预训练和微调。第一阶段，通过对比学习方法对模型进行预训练，然后保存模型参数。在分类任务的第二阶段，利用监督学习方法进行微调，得到最终的分类结果。第一阶段中的预训练和优化的最终目的是最小化损失函数，拉近正样本距离，拉远负样本距离。本章采用了模态内对比学习和模态间对比学习策略，模态内对比学习的具体实现过程如下：

将增强后的数据和原始数据输入到前文构建的特征提取网络中，得到特征 f_i^1 和 f_i^2，然后把特征通过投影头 $g(\cdot)$ 投影到新的嵌入空间，得到 F_i^1 和 F_i^2：

$$F_i^1 = g\left(f_i^1\right),\quad F_i^2 = g\left(f_i^2\right), \tag{4-5}$$

然后计算 f_i 和 F_i 的余弦相似度：

$$S(f_i, F_i) = -\frac{F_i}{\left\|F_i\right\|_2}\cdot\frac{f_i}{\left\|f_i\right\|_2}, \tag{4-6}$$

则对于 HSI 或者 LiDAR 单一模态内部的某一正样本对的对比损失函数表示为：

$$l_h\left(i, F_i^1, F_i^2\right) = -\log\frac{\exp\left(S\left(F_i^1, F_i^2\right)/\tau\right)}{\sum_{k=1}^{2N}\Gamma_{[k\neq i]}\exp\left(S\left(F_i^1, F_k\right)/\tau\right)}, \tag{4-7}$$

其中，$\Gamma_{[k\neq i]}\in\{0,1\}$ 表示指标函数，温度系数 τ 默认设置为 0.1，N 为 mini-batch 的大小。则 mini-batch 大小的模态内对比学习损失为：

$$l_{intra} = \frac{1}{2N}\sum_{i=1}^{N}\left[l_h\left(i, F_i^1, F_i^2\right) + l_h\left(i, F_i^2, F_i^1\right)\right], \tag{4-8}$$

4.3.5 模态间对比学习

模态内的对比学习旨在学习同一模态的类别特征信息，模态间的对比学习被用来提取不同模态间隐藏的互补信息，拉近不同模态特征之间的距离。模态间的对比学习损失函数计算如下：

先计算 F_i^{HSI1} 和 F_i^{HSI2} 的均值 F_i^{HSI}，以及 F_i^{LiDAR1} 和 F_i^{LiDAR2} 的均值 F_i^{LiDAR}：

$$F_i^{HSI}=\frac{1}{2}\left(F_i^{HSI1}+F_i^{HSI2}\right),\quad F_i^{LiDAR}=\frac{1}{2}\left(F_i^{LiDAR1}+F_i^{LiDAR2}\right), \tag{4-9}$$

则两模态间的正样本对 F_i^{HSI} 和 F_i^{LiDAR} 之间的损失函数为：

$$l_c\left(i,F_i^{HSI},F_i^{LiDAR}\right)=-\log\frac{\exp\left(S\left(F_i^{HSI},F_i^{LiDAR}\right)/\tau\right)}{\sum_{k=1}^{2N}\Gamma_{[k\neq i]}\exp\left(S\left(F_i^{HSI},F_k\right)/\tau\right)}, \tag{4-10}$$

则在 mini-batch 大小时的模态间对比学习损失为：

$$l_{cross}=\frac{1}{2N}\sum_{i=1}^{N}\left[l_c\left(i,F_i^{HSI},F_i^{LiDAR}\right)+l_c\left(i,F_i^{LiDAR},F_i^{HSI}\right)\right], \tag{4-11}$$

以上就是本章使用的模态内和模态间的对比学习的损失函数。

4.4 实验与分析

4.4.1 实验环境与参数设定

（1）本章实验环境详细配置如 2.4.3 小节的表 2.4 所示。

（2）本章的参数设定如下：

为保证实验公平性，对实验过程中的可变因素进行定量控制，在本章所有使用的对比实验和消融实验中均采用以下配置：评价指标方面使用总体正确率 OA、平均准确率 AA 和 KC；训练过程中批大小设定为 256；采用的优化器为 Adam，权重衰减参数为 0.9；初始学习率设定为 $5e^{-3}$，而后采用线性衰减策略对学习率进行衰减；最大迭代次数为 100 代；使用了交叉熵损失作为损失函数。每个实验重复 10 次取平均值和标准偏差。

本章算法在 Houston2013、Trento 数据集上进行性能评估，数据集设置不同于二三两章的是，预训练阶段使用 50%的未标记样本作为训练样本；在微调阶段每类使用 10 个样本进行微调，其余作为测试样本，训练和微调过程中批大小设定为 256。

4.4.2 对比实验结果与分析

对比实验在前两章对比的五种方法基础上添加了两种基于自监督学习的算法：

SimCLR[80]、SSFR[81]，监督学习的五种算法在实验过程中使用文章中的原始参数，本章方法和对比自监督学习对比方法在预训练阶段每类随机选择 10 个有标签样本进行预训练，本章方法参数同上所述。在两个数据集上的对比结果和分析如下：

（1）Houston2013 数据集对比实验结果

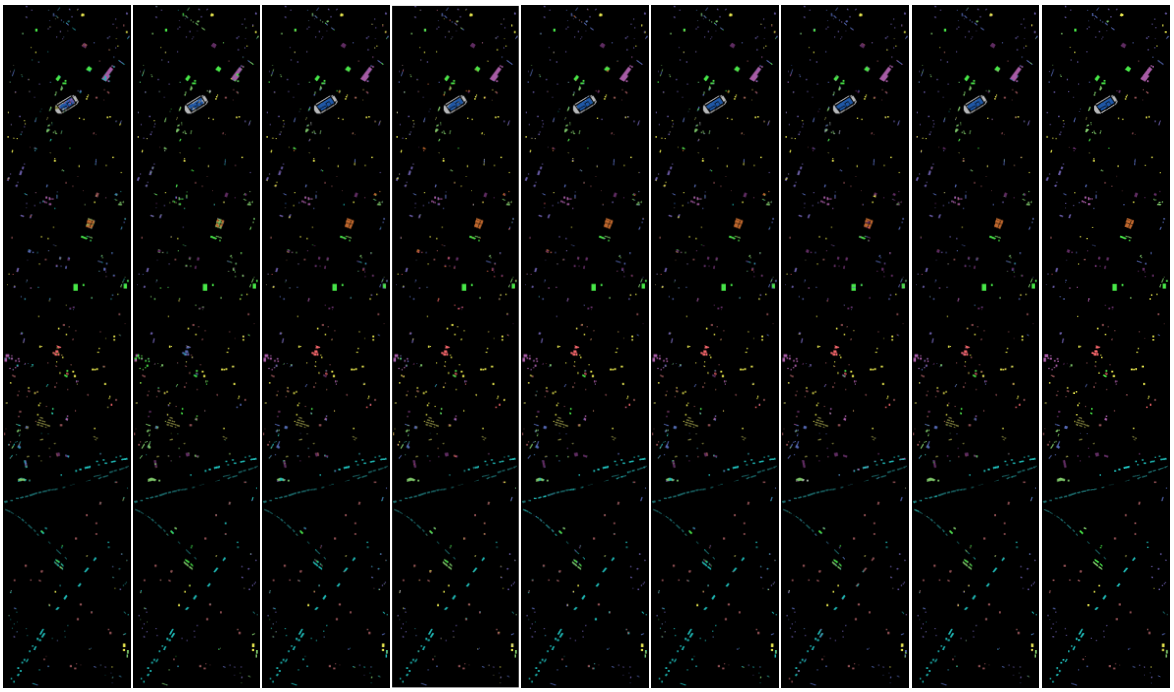
本章方法与对比方法的性能结果对比如表 4.2 所示，其对应的可视化结果图如图 4.5 所示，其中（a）~（i）分别代表 ELM、DeepCNN、FusAtNet、EndNet、HRWN、SimCLR、SSFR、本章方法的分类可视化结果图和 Groundtruth 图。

表4.2 基于对比学习的算法在 Houston2013 数据集上的对比实验结果

No.	ELM	DeepCNN	FusAtNet	EndNet	HRWN	SimCLR	SSFR	Ours
1	79.31±6.48	82.43±5.59	54.85±10.1	85.89±3.01	88.07±3.59	94.01±3.01	**95.65**±3.49	87.01±3.13
2	83.68±3.51	85.38±5.04	**89.71**±3.32	83.44±3.77	87.11±2.76	74.78±3.88	86.96±4.61	85.24±1.21
3	67.26±2.79	76.66±3.75	89.22±2.63	85.77±1.95	90.21±3.22	**99.18**±0.95	99.10±1.54	89.73±1.81
4	89.39±0.53	85.97±1.22	84.63±5.25	89.51±0.83	86.43±5.04	**95.12**±1.43	88.50±5.65	87.92±3.50
5	84.15±3.56	87.02±0.98	88.74±2.76	86.16±2.33	89.96±2.41	91.31±3.96	**94.37**±3.86	93.22±0.97
6	68.52±6.83	67.85±6.33	77.67±3.90	86.55±1.87	90.56±1.57	**96.30**±4.41	95.57±3.22	91.74±1.23
7	78.44±5.88	78.02±3.64	84.23±2.42	**88.73**±0.59	85.88±2.61	88.21±4.03	83.12±5.96	85.65±2.21
8	79.93±6.04	80.54±5.23	77.12±4.31	85.64±1.21	77.53±3.09	83.36±4.17	78.37±6.49	**86.20**±2.78
9	60.79±3.55	78.77±4.53	82.01±3.55	70.08±6.31	81.69±3.04	72.69±5.04	78.63±5.82	**83.23**±1.55
10	61.21±2.80	69.41±8.55	70.66±6.13	67.36±5.45	84.01±3.28	62.76±4.90	66.97±4.76	**89.77**±1.02
11	57.72±4.42	57.56±9.01	82.42±1.37	76.33±3.31	81.21±1.17	77.67±3.95	**88.57**±6.03	81.47±3.10
12	69.51±7.49	67.44±4.91	57.51±11.2	76.99±2.57	**86.33**±3.95	63.02±6.18	78.74±5.04	83.66±2.31
13	64.40±4.29	63.93±6.66	**85.75**±0.97	68.27±4.12	81.57±2.47	60.26±4.84	62.32±14.1	83.83±1.33
14	84.64±5.62	85.81±1.95	75.96±3.87	88.52±0.92	89.99±0.73	**96.55**±2.87	91.99±5.16	94.78±0.56
15	90.42±5.69	89.37±2.99	87.53±1.58	91.48±0.42	90.13±1.06	**98.76**±1.26	98.46±1.28	96.44±0.72
OA(%)	73.36±1.74	77.51±1.21	78.54±1.21	83.23±1.53	84.71±0.32	80.80±1.17	84.15±1.03	**87.34**±0.34
AA(%)	74.62±1.63	77.08±3.01	79.20±2.23	82.05±1.39	86.04±0.44	83.60±0.61	85.82±0.69	**87.99**±0.29
KC(%)	73.29±1.87	76.76±2.11	78.89±1.03	81.79±0.98	83.60±0.36	79.26±1.26	82.88±1.11	**85.72**±0.51

从表 4.2 中可以看出，本文的五个基于深度学习的监督学习算法，CNN 等网络有强大的特征提取能力，整体的分类效果较为可观，但是同第二章和第三章中的实验参数情况下的实验结果对比发现，在少样本情况下，即仅有每类 10 个样本进行预训练，很明显的发现各类别及对应的 OA、AA、KC 指标都比大量样本进行预训练的情况下明显下降，可见监督学习的方法对于样本存在依赖性。对比发现，本文的对比学习方

法在使用 50%无标签样本进行学习和表示后，仅使用 10 个有标签样本微调时，OA、AA、KC 指标都高于五个监督学习的对比方法。与 SimCLR、SSFR 两种自监督学习算法对比，虽然在 15 个类别中只有 3 类的准确率达到最优，但是不难发现，本章算法对应的各类别准确率比较稳定偏高，证明本章算法模型的学习表征能力较强且稳定。主要是因为模态内对比学习有效的学习了两个模态内的特征表示，同时模态间对比学习可以在预训练阶段捕获两个模态数据之间的对应关系，使学习到的特征表示更具有一致性和相关性，并在微调中通过第三章所提出的交叉注意力进一步集成两个模态的互补特征信息，从而产生了更好的分类性能。综上，基于对比学习的方法可以有效的减少网络对样本数量和质量的依赖，在样本量不足时完全能够根据没有标签的数据自主学习，生成标签数据。



(a) (b) (c) (d) (e) (f) (g) (h) (i)

图4.5 基于对比学习的算法在 Houston2013 数据集上的可视化结果图

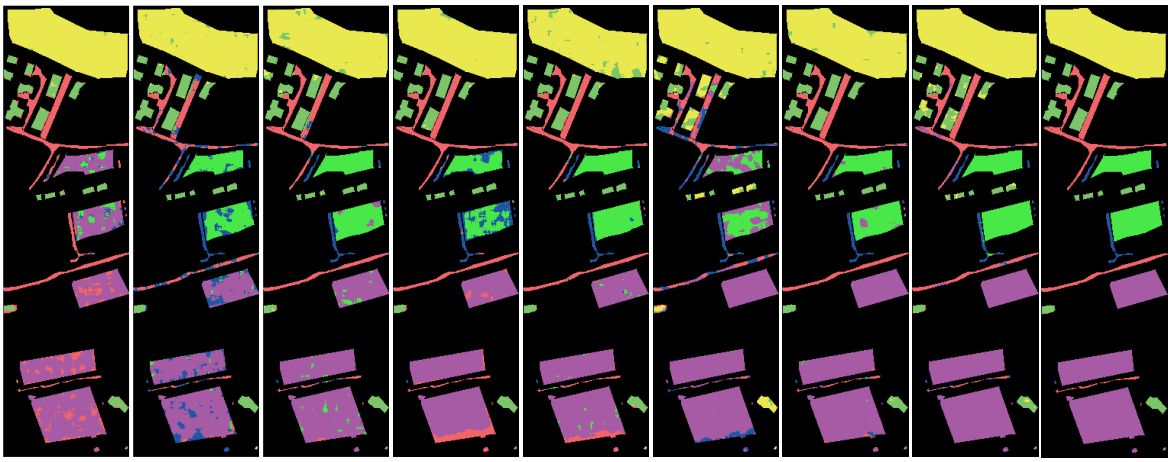
（2）Trento 数据集对比实验结果

从表 4.3 中可以看出，在 OA、AA 和 KC 三个指标上，本章的方法均是最高。本章的方法在三个指标上分别领先五个监督学习对比算法中效果最好的方法 3.1%、0.65%、2.42%。对于 Trento 数据集，其类别数和样本量是比较少的，也都是较为集中的区域，所以对于监督学习的几个方法，分类效果大多不错，但是与这五种监督学习的算法对比发现，对于 6 个类别，本章方法在 4 个类别中的分类效果达到最优，可见少样本情况下对比学习方法的学习能力和效果是明显优于监督学习的方法的，在更复杂的分类任务中会有更明显的表现。具体的原因和优势就不赘述了，同 Houston2013

数据集实验分析。对应的可视化结果图如图 4.6 所示，其中（a）~（i）分别代表 ELM、DeepCNN、FusAtNet、EndNet、HRWN、SimCLR、SSFR、本章方法在 Trento 数据集上的分类可视化结果图和 Groundtruth 图。

表4.3 基于对比学习的算法在 Trento 数据集上的对比实验结果

No.	ELM	DeepCNN	FusAtNet	EndNet	HRWN	SimCLR	SSFR	Ours
1	53.32±1.43	77.59±2.81	95.21±1.94	94.99±3.35	97.05±3.18	67.47±6.55	87.52±1.09	**98.17**±1.13
2	92.92±3.21	75.33±4.31	96.66±3.90	90.55±4.71	89.78±1.55	75.78±4.61	**97.64**±0.60	96.23±1.58
3	67.88±2.95	94.25±1.56	95.22±3.06	92.33±4.40	**96.63**±2.43	64.58±10.2	81.52±6.50	90.14±1.74
4	96.43±2.44	96.43±4.17	96.73±3.17	89.69±5.07	97.25±1.79	82.56±4.37	**99.95**±0.05	99.28±0.05
5	86.54±1.55	96.47±1.37	95.05±3.31	92.65±3.19	97.52±2.18	88.37±4.37	**99.41**±0.13	98.53±0.93
6	93.19±1.31	88.06±3.24	87.98±2.46	88.43±3.74	89.77±2.06	76.70±4.35	**91.66**±0.74	89.55±3.11
OA(%)	82.43±3.32	89.95±2.55	93.89±1.35	91.57±2.52	93.74±1.69	79.73±2.58	96.44±0.34	**96.99**±0.40
AA(%)	81.71±3.41	88.02±2.64	94.48±2.23	91.44±1.75	94.67±2.31	75.91±2.57	92.95±1.21	**95.32**±0.32
KC(%)	80.64±2.46	89.33±1.89	93.56±1.56	93.73±2.02	93.91±1.76	73.21±3.32	95.27±0.45	**96.33**±0.61



(a) (b) (c) (d) (e) (f) (g) (h) (i)

图4.6 基于对比学习的算法在 Trento 数据集上的可视化结果图

4.4.3 消融实验结果与分析

本章消融实验主要分为以下四个模型：只有 HSI 数据输入、只有 LiDAR 数据输入、结合 HSI 和 LiDAR 的多模态数据输入以及本章在在预训练阶段使用的对比学习方法。前三个消融实验设置的主要目的是证明本文一直有强调的核心问题多模态遥感图像分类的有效性，在单一模态数据的分类实验中，只进行单一模态内的对比学习，无模态间的对比学习部分工作；预训练部分的消融是为了证明对比学习在样本受限情

况下的有效性。对于本章在特征融合阶段使用的交叉注意力模块，在第三章中已经进行了消融，说明了模块的有效性，故在这里就不进一步进行消融研究了。具体消融实验结果如表 4.4 所示。

表4.4 基于对比学习的算法在 Houston2013 数据集上的消融实验结果

模型			Houston2013		
HSI	LiDAR	预训练	OA(%)	AA(%)	KC(%)
✓			77.24±3.42	78.53±2.36	77.91±3.27
	✓		81.31±2.15	82.06±3.21	81.47±3.09
✓	✓		83.25±2.14	83.87±1.78	82.19±2.11
✓	✓	✓	**87.34**±0.34	**87.99**±0.29	**85.72**±0.51

从表 4.4 中可以发现，单独的 HSI 或者 LiDAR 数据，在分类指标上呈现出较低的效果，原因之一是此处消融时当不使用对比学习进行预训练时仅有简单的三层 CNN 作为特征提取器进行特征提取，会导致特征提取不充分，尤其对于光谱特征复杂、维度高的 HSI 数据，所以单模态时 HSI 的指标普遍低于 LiDAR。此外，会发现在两个模态数据进行融合和分类时，特征信息之间有一定的互补和交互，能够实现更好的融合，输入分类器中的融合特征能够有更好的分类表现。最后，对于对比学习预训练部分的消融，能够从数据中明显发现，对比于没有对比学习的样本受限实验条件下，OA、AA、KC 分别提升了 4.09%、4.12%、3.53%，主要是因为预训练阶段，本文设置的对比学习方法从大量无标签数据中很好的学习到了特征表示，且在模态间的对比学习中，很好的捕获和保证了两个模态特征间的关系和语义一致性。

4.5 本章小结

本章提出了基于对比学习的多模态遥感图像分类算法，该算法由简单 CNN 特征提取器、模态内和模态间对比学习预训练模块、基于交叉注意力的特征融合模块及分类器四部分组成。首先在特征提取阶段用简单的三层 CNN 对 LiDAR 和 HSI 数据进行特征提取，避免使用复杂的网络模型，降低了模型的复杂度。其次在预训练阶段设计使用了针对模态内的对比学习和模态间的对比学习，主要通过对比损失函数实现，对两种模态特征表示的学习同时保证了模态间特征的相关性和一致性，为融合模块提供基础。最后，通过与第三章相同的交叉注意力模块中进行特征融合，此处进一步进行了特征间的交互，保障了分类的效果。最终通过对比实验和消融实验证明，本章引入对比学习的模型很好的解决了遥感图像样本受限和 CNN 等监督学习对样本过度依赖的问题，在同样少量的微调样本情况下，比所对比的其他监督学习和对比学习自监

督算法表现出更好的分类性能。

第五章 总结与展望

5.1 总结

遥感图像分类有助于人们更直观、更系统地获取感兴趣的信息。遥感图像在资源调查、自然灾害观测、大气气象预报等多个领域都发挥着重要作用。通过对遥感图像进行分类处理，可以更准确地识别地面物体的类属及其分布特征，为相关领域的决策提供科学依据。多模态特征融合帮助遥感图像分类进一步提升分类性能，更全面的反映真实的地物特征。然而，目前常用的 CNN 对如 HSI 数据的光谱特征提取不充分，对维度高、信息过于丰富的数据特征提取不充分；异构数据间特征不一致、交互不充分；遥感数据样本量稀缺、标注困难，CNN 和 Transformer 等监督学习的方法过于依赖样本数量和质量，在遥感图像数据受限情况下容易过拟合。因此，本文旨在结合多模态遥感图像数据，通过针对性设计特征提取网络和加强数据间交互融合等方法，使分类性能进一步提升。具体工作总结如下：

（1）针对遥感图像数据单一模态的分辨率和提供的特征信息有限、缺乏空间结构特征和高程特征信息，引入了 LiDAR 数据进行多模态数据融合；针对 CNN 对 HSI 数据的光谱特征提取不充分等问题，对应 HSI 数据空间特征和光谱特征特点设计了对应的双分支空谱残差网络，进行两条支路提取空间-光谱特征后融合，弥补了 CNN 对光谱特征提取不充分的缺陷；同时引入 SNN 和可学习参数进一步改善特征融合阶段的效果，提升了模型的特征提取能力和特征丰富程度。实验证明，该方案对比于单一模遥感图像分类及一些比较前沿的多模态遥感图像分类方法的准确率有所提高。

（2）针对 CNN 对维度高、光谱信息复杂的 HSI 数据特征提取不充分且难以关注长距离依赖信息等网络结构固有的问题，本文使用了 Transformer 针对 HSI 数据进行特征提取，这里使用的是对光谱特征提取效果较好的 SpectraFormer；针对多模态数据异构特征间难以保持一致性、特征未充分交互，本文引入了一致性损失来保障了异构特征之间的双向交互和特征一致性，在特征融合阶段引入交叉注意力机制，进一步增强两个模态间的特征的交互融合，从而提升了模型的分类效果。

（3）针对监督方法过于依赖样本数量和质量，而现存的多模态遥感图像数据不足，样本量少，且标注成本高，很容易出现过拟合情况，本文在多模态遥感图像分类任务中引入了对比学习方法，除了在单一模态内使用对比学习利用未标注数据学习特征表示，缓解了有监督学习对样本依赖性的问题，还在两模态间增加了对比学习方法，增加了模态间特征的语义一致，提出的方法在少样本条件下有很高的准确率。

5.2 展望

本文主要研究了基于多模态特征融合的遥感图像分类任务系列方法，针对目前多模态遥感图像分类方法中存在的问题做出了多方面的改进，但由于 CNN 和 Transformer 的技术体系比较庞大，另外对自监督学习的研究时间有限，本文的研究还存在一定的不足，仍有一些方面问题需要进一步进行研究：

（1）本文主要针对了 HSI 和 LiDAR 两种模态的数据融合作用到分类任务中，未来可以尝试更多模态数据的融合。此外，二三两章针对维度高、光谱复杂的 HSI 数据使用了 DBSSRN 和 SpectralFormer 进行特征提取，未来可以考虑使用 Transformer 对 LiDAR 数据或者其他模态遥感数据进行特征提取和融合，辅助分类任务的完成，并设计具备泛化性的模型。

（2）本文使用的交叉注意力机制是最原始的结构，可以设计其他的交叉规则，让模态间的特征更充分的交互，例如 Q 值加权后共享等。此外可以尝试更多的一致性损失函数，使得模态间特征更好的对齐，提升模型预训练的效果。

（3）本文基于对比学习的多模态遥感图像分类方法的预训练时间比较长，不适合在一些实时或短时实际场景中应用，所以未来需要进一步考虑在不影响性能的前提下对预训练时间的压缩。此外，本文只尝试和实现了基于特征级的对比学习的方法研究，更具备细粒度的基于像素级的生成式自监督学习方法可能会有更好的效果，这也是未来可以研究和实现的方向。

参考文献

[1] Peng C, Li Y, Jiao L, et al. Efficient convolutional neural architecture search for remote sensing image scene classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 59(7): 6092-6105.

[2] 焦李成, 尚荣华, 刘芳等. 稀疏学习、分类与识别[M]. 北京: 科学出版社, 2017.

[3] Ghamisi P, Rasti B, Yokoya N, et al. Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art[J]. IEEE Geoscience and Remote Sensing Magazine, 2019, 7(1): 6-39.

[4] 焦李成, 张向荣, 侯彪等. 智能 SAR 图像处理与解译[M]. 北京: 科学出版社, 2008.

[5] 焦李成, 李阳阳, 刘芳等. 量子计算、优化与学习[M]. 北京: 科学出版社, 2017.

[6] Ali N, Zafar B, Riaz F, et al. A hybrid geometric spatial image representation for scene classification[J]. PloS one, 2018, 13(9).

[7] Latif A, Rasheed A, Sajid U, et al. Content-based image retrieval and feature extraction: a comprehensive review[J]. Mathematical problems in engineering, 2019, 2019.

[8] Yadav K, Yadav M, Saini S. Stock values predictions using deep learning based hybrid models[J]. CAAI Transactions on Intelligence Technology, 2022, 7(1): 107-116.

[9] Shakya A, Biswas M, Pal M. Parametric study of convolutional neural network based remote sensing image classification[J]. International Journal of Remote Sensing, 2021, 42(7): 2663-2685.

[10] Hu W S, Li H C, Pan L, et al. Spatial–spectral feature extraction via deep ConvLSTM neural networks for hyperspectral image classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(6): 4237-4250.

[11] Mei S, Li X, Liu X, et al. Hyperspectral image classification using attention-based bidirectional long short-term memory network[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-12.

[12] Zhang F, Bai J, Zhang J, et al. An optimized training method for GAN-based hyperspectral image classification[J]. IEEE Geoscience and Remote Sensing Letters, 2020, 18(10): 1791-1795.

[13] Shabbir A, Ali N, Ahmed J, et al. Satellite and scene image classification based on transfer learning and fine tuning of ResNet50[J]. Mathematical Problems in Engineering, 2021, 2021: 1-18.

[14] Yan P, He F, Yang Y, et al. Semi-supervised representation learning for remote sensing image classification based on generative adversarial networks[J]. IEEE Access, 2020, 8: 54135-54144.

[15] Li Z, Huang L, He J. A multiscale deep middle-level feature fusion network for hyperspectral classification[J]. Remote Sensing, 2019, 11(6): 695.

[16] Hong D, Yokoya N, Chanussot J, et al. CoSpace: Common subspace learning from hyperspectral-multispectral correspondences[J]. IEEE Transactions on Geoscience and Remote Sensing, 2019, 57(7): 4349-4359.

[17] Huo L Z, Silva C A, Klauberg C, et al. Supervised spatial classification of multispectral LiDAR data in urban areas[J]. PloS one, 2018, 13(10).

[18] Heiden U, Heldens W, Roessner S, et al. Urban structure type characterization using hyperspectral remote sensing and height information[J]. Landscape and urban Planning, 2012, 105(4): 361-375.

[19] Moreira A, Prats-Iraola P, Younis M, et al. A tutorial on synthetic aperture radar[J]. IEEE Geoscience and remote sensing magazine, 2013, 1(1): 6-43.

[20] Huang B, Li Y, Han X, et al. Cloud removal from optical satellite imagery with SAR imagery using sparse representation[J]. IEEE Geoscience and Remote Sensing Letters, 2015, 12(5): 1046-1050.

[21] 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.

[22] Khodadadzadeh M, Li J, Prasad S, et al. Fusion of hyperspectral and LiDAR remote sensing data using multiple feature learning[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2015, 8(6): 2971-2983.

[23] Wu X, Hong D, Chanussot J. Convolutional neural networks for multimodal remote sensing data classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-10.

[24] Hang R, Li Z, Ghamisi P, et al. Classification of hyperspectral and LiDAR data using coupled CNNs[J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(7): 4939-4950.

[25] Wang J, Li W, Gao Y, et al. Hyperspectral and SAR image classification via multiscale interactive fusion network[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022.

[26] Xue Z, Tan X, Yu X, et al. Deep hierarchical vision transformer for hyperspectral and LiDAR data classification[J]. IEEE Transactions on Image Processing, 2022, 31: 3095-3110.

[27] Roy S K, Deria A, Hong D, et al. Hyperspectral and LiDAR data classification using joint CNNs and morphological feature learning[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-16.

[28] Bromley J, Guyon I, LeCun Y, et al. Signature verification using a "siamese" time delay neural network[J]. Advances in neural information processing systems, 1993, 6.

[29] Angelovska M, Sheikholeslami S, Dunn B, et al. Siamese neural networks for detecting complementary products[C]//Proceedings of the 16th conference of the European chapter of the association for computational linguistics: student research workshop. 2021: 65-70.

[30] Nair V, Hinton G E. Rectified linear units improve restricted boltzmann machines[C]//Proceedings of the 27th international conference on machine learning (ICML-10), 2010: 807-814.

[31] Melekhov I, Kannala J, Rahtu E. Image patch matching using convolutional descriptors with euclidean distance[C]//Asian Conference on Computer Vision. Cham: Springer International Publishing, 2016: 638-653.

[32] Neculoiu P, Versteegh M, Rotaru M. Learning text similarity with siamese recurrent networks[C]//Proceedings of the 1st Workshop on Representation Learning for NLP, 2016: 148-157.

[33] Guo Q, Feng W, Zhou C, et al. Learning dynamic siamese network for visual object tracking[C]//Proceedings of the IEEE international conference on computer vision, 2017: 1763-1771.

[34] An N, Qi Yan W. Multitarget tracking using Siamese neural networks[J]. ACM Transactions on Multimidia Computing Communications and Applications, 2021, 17(2s): 1-16.

[35] Liu C F, Padhy S, Ramachandran S, et al. Using deep Siamese neural networks for detection of brain asymmetries associated with Alzheimer's disease and mild cognitive impairment[J]. Magnetic resonance imaging, 2019, 64: 190-199.

[36] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[37] Parmar N, Vaswani A, Uszkoreit J, et al. Image transformer[C]//International conference on machine learning. PMLR, 2018: 4055-4064.

[38] Chen M, Radford A, Child R, et al. Generative pretraining from pixels[C]//International conference on machine learning. PMLR, 2020: 1691-1703.

[39] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transfor mers for image recognition at scale[J]. 2020. https://arxiv.org/abs/2010.11929.

[40] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention[C]//International conference on machine learning. PMLR, 2021: 10347-10357.

[41] 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: 10012-10022.

[42] Zhang H, Li F, Liu S, et al. Dino: Detr with improved denoising anchor boxes for end-to-end object detection[J]. 2022. https://arxiv.org/abs/2203.03605.

[43] Aleissaee A A, Kumar A, Anwer R M, et al. Transformers in remote sensing: A survey[J]. Remote Sensing, 2023, 15(7): 1860.

[44] He J, Zhao L, Yang H, et al. HSI-BERT: Hyperspectral image classification using the bidirectional

encoder representation from transformers[J]. IEEE Transactions on Geoscience and Remote Sensing, 2019, 58(1): 165-178.

[45] Hong D, Han Z, Yao J, et al. SpectralFormer: Rethinking hyperspectral image classification with transformers[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-15.

[46] Zhong Z, Li Y, Ma L, et al. Spectral–spatial transformer network for hyperspectral image classification: A factorized architecture search framework[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-15.

[47] Liu B, Yu A, Gao K, et al. DSS-TRM: Deep spatial–spectral transformer for hyperspectral image classification[J]. European Journal of Remote Sensing, 2022, 55(1): 103-114.

[48] Z. Zhao, D. Hu, H. Wang, et al. Convolutional Transformer Network for Hyperspectral Image Classification[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 1-5.

[49] Yang X, Cao W, Lu Y, et al. Hyperspectral image transformer classification networks[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-15.

[50] Jia S, Wang Y. Multiscale convolutional transformer with center mask pretraining for hy perspectral image classification[J]. 2022. https://arxiv.org/abs/2203.04771

[51] Sun L, Zhao G, Zheng Y, et al. Spectral–spatial feature tokenization transformer for hyperspectral image classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-14.

[52] Roy S K, Deria A, Hong D, et al. Multimodal fusion transformer for remote sensing image classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-20.

[53] Xue Z, Tan X, Yu X, et al. Deep hierarchical vision transformer for hyperspectral and LiDAR data classification[J]. IEEE Transactions on Image Processing, 2022, 31: 3095-3110.

[54] Rani V, Nabi S T, Kumar M, et al. Self-supervised learning: A succinct review[J]. Archives of Computational Methods in Engineering, 2023, 30(4): 2761-2775.

[55] Zhou M, Li Z, Xie P. Self-supervised regularization for text classification[J]. Transactions of the Association for Computational Linguistics, 2021, 9: 641-656.

[56] Chen T, Liu S, Chang S, et al. Adversarial robustness: From self-supervised pre-training to fine-tuning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020: 699-708.

[57] Lin D, Fu K, Wang Y, et al. MARTA GANs: Unsupervised representation learning for remote sensing image classification[J]. IEEE Geoscience and Remote Sensing Letters, 2017, 14(11): 2092-2096.

[58] Stojnić V, Risojević V. Evaluation of split-brain autoencoders for high-resolution remote sensing scene classification[C]//2018 International Symposium ELMAR. IEEE, 2018: 67-70.

[59] Gidaris S, Singh P, Komodakis N. Unsupervised representation learning by predicting image

rotations[J]. 2018. https://arxiv.org/abs/1803.07728.

[60] Shu Q, Liu S, Wang J, et al. Image Classification Algorithm Named OCFC Based on Self-supervised Learning[C]//2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC). IEEE, 2020: 589-594.

[61] Li Y, Chen J, Zheng Y. A multi-task self-supervised learning framework for scopy images[C]//2020 IEEE 17th international symposium on biomedical imaging (ISBI). IEEE, 2020: 2005-2009.

[62] Zhao Z, Luo Z, Li J, et al. When self-supervised learning meets scene classification: Remote sensing scene classification based on a multitask learning framework[J]. Remote Sensing, 2020, 12(20): 3276.

[63] Jung H, Jeon T. Self‐supervised learning with randomised layers for remote sensing[J]. Electronics Letters, 2021, 57(6): 249-251.

[64] Scheibenreif L, Mommert M, Borth D. Contrastive self-supervised data fusion for satellite imagery[J]. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2022, 3: 705-711.

[65] Liu X, Zhang F, Hou Z, et al. Self-supervised learning: Generative or contrastive[J]. IEEE transactions on knowledge and data engineering, 2021, 35(1): 857-876.

[66] Caron M, Bojanowski P, Joulin A, et al. Deep clustering for unsupervised learning of visual features[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 132-149.

[67] Zhuang C, Zhai A L, Yamins D. Local aggregation for unsupervised learning of visual embeddings[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 6002-6012.

[68] Le-Khac P H, Healy G, Smeaton A F. Contrastive representation learning: A framework and review[J]. IEEE Access, 2020, 8: 193907-193934.

[69] Jaiswal A, Babu A R, Zadeh M Z, et al. A survey on contrastive self-supervised learning[J]. Technologies, 2020, 9(1): 2.

[70] He K, Fan H, Wu Y, et al. Momentum contrast for unsupervised visual representation learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 9729-9738.

[71] Huynh T, Kornblith S, Walter M R, et al. Boosting contrastive self-supervised learning with false negative cancellation[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2022: 2785-2795.

[72] Saad A B, Drouyer S, Hell B, et al. A review on contrastive learning methods and applications to roof-type classification on aerial images[C]//2021 IEEE International Geoscience and Remote

Sensing Symposium IGARSS. IEEE, 2021: 4960-4963.

[73] Shorfuzzaman M, Hossain M S. MetaCOVID: A Siamese neural network framework with contrastive loss for n-shot diagnosis of COVID-19 patients[J]. Pattern recognition, 2021, 113: 107700.

[74] Su J, Ahmed M, Lu Y, et al. Roformer: Enhanced transformer with rotary position embedding[J]. Neurocomputing, 2024, 568: 127063.

[75] Lu S, Liu M, Yin L, et al. The multi-modal fusion in visual question answering: a review of attention mechanisms[J]. PeerJ Computer Science, 2023, 9: e1400.

[76] Huang G B, Zhu Q Y, Siew C K. Extreme learning machine: theory and applications[J]. Neurocomputing, 2006, 70(1-3): 489-501.

[77] Mohla S, Pande S, Banerjee B, et al. Fusatnet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020: 92-93.

[78] Hong D, Gao L, Hang R, et al. Deep encoder–decoder networks for classification of hyperspectral and LiDAR data[J]. IEEE Geoscience and Remote Sensing Letters, 2020, 19: 1-5.

[79] Zhao X, Tao R, Li W, et al. Joint classification of hyperspectral and LiDAR data using hierarchical random walk and deep CNN architecture[J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(10): 7355-7370.

[80] Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations[C]//International conference on machine learning. PMLR, 2020: 1597-1607.

[81] Xue Z, Liu B, Yu A, et al. Self-supervised feature representation and few-shot land cover classification of multimodal remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-18.