SemiBaCon

Semi-Supervised Balanced Contrastive Learning for Multi-Modal Remote Sensing Image Classification

He, Yufei; Xi, Bobo; Li, Guocheng; Zheng, Tie; Li, Yunsong; Xue, Changbin; Shen, Ming

Published in:
IEEE Transactions on Geoscience and Remote Sensing

DOI (link to publication from Publisher):
10.1109/TGRS.2025.3589487

Publication date:
2025

Document Version
Accepted author manuscript, peer reviewed version

SemiBaCon: Semi-Supervised Balanced Contrastive Learning for Multi-Modal Remote Sensing Image Classification

Yufei He, *Student Member, IEEE*, Bobo Xi, *Member, IEEE*, Guocheng Li, *Student Member, IEEE*, Tie Zheng, Yunsong Li, *Member, IEEE*, Changbin Xue, *Member, IEEE*, and Ming Shen, *Senior Member, IEEE*

***Abstract*—The limited availability of annotated training data significantly constrains the classification accuracy of hyperspectral image (HSI) and LiDAR fusion approaches. Although contrastive learning has emerged as a potential solution, current implementations frequently neglect the critical class imbalance issues during unlabeled sample selection. To address the issue, we introduce a novel semi-supervised balanced contrastive learning (SemiBaCon) framework for multi-modal remote sensing image classification. First, we propose a superpixel-based balanced sampling (SPBS) mechanism that fundamentally addresses class imbalance through intelligent pseudo-label generation. By segmenting the HSI data into homogeneous superpixels and implementing intra-region label propagation, the method ensures statistically balanced pseudo-label selection across categories, effectively overcoming the bias introduced by conventional random sampling strategies. Second, our architecture integrates a dual-stream encoder combining convolutional neural networks (CNNs) with Transformers, enabling hierarchical feature extraction from spectral-spatial characteristics of HSI and elevation patterns of LiDAR. This design facilitates the construction of multi-modal positive sample pairs, achieving enhanced representation learning through inter-modal consistency constraints. Third, we develop a pseudo-label guided contrastive learning (PLCL) paradigm that synergistically combines pseudo-label confidence with feature similarity metrics, which effectively reduces intra-class variance and improves decision boundaries in the latent space. Comprehensive evaluations on three benchmark datasets demonstrate the framework's superior performance compared to the state-of-the-art methods.**

***Index Terms*—semi-supervised, contrastive learning, superpixels, multi-modal, balanced sampling.**

This work was supported by the National Nature Science Foundation of China under Grant 62401434, the China Postdoctoral Science Foundation under Grant 2023M732742, the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2024JC-YBQN-0641, Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515110504, the Postdoctoral Science Foundation of Shaanxi Province under Grant 2023BSHYDZZ97, the Climbing Program of National Space Science Center E3PD40013S. *(Corresponding authors: Bobo Xi, Changbin Xue)*

Yufei He and Guocheng Li are with the Key Laboratory of Electronics and Information Technology for Space Systems, National Space Science Center, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China (email: heyufei20@mails.ucas.ac.cn; liguocheng20@mails.ucas.ac.cn).

Bobo Xi is with the State Key Laboratory of Integrated Service Networks, School of Telecommunications Engineering, Xidian University, Xi'an 710071, China, and also with the National Space Science Center, Chinese Academy of Sciences, Beijing 100190, China (e-mail: xibobo1301@foxmail.com).

Tie Zheng and Changbin Xue are with the Key Laboratory of Electronics and Information Technology for Space Systems, National Space Science Center, Chinese Academy of Sciences, Beijing 100190, China (e-mail: zhengtie@nssc.ac.cn, xuechangbin@nssc.ac.cn).

Yunsong Li is with the State Key Laboratory of Integrated Service Networks, School of Telecommunications Engineering, Xidian University, Xi'an 710071, China (e-mail: ysli@mail.xidian.edu.cn).

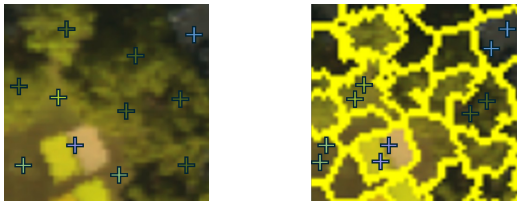
Ming Shen is with the Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark (e-mail: mish@es.aau.dk).

I. Introduction

MULTI-MODAL data fusion has emerged as a crucial technology in remote sensing (RS) and computer vision, significantly enhancing classification accuracy by leveraging complementary information from different modalities [1]–[4]. Among various multi-modal data sources, hyperspectral images (HSI) and light detection and ranging (LiDAR) data are widely utilized due to their distinct yet complementary characteristics [5]–[8]. The HSI provides rich spectral information for material identification [9]–[11], while the LiDAR captures elevation and structural details, aiding in differentiating objects with similar spectral properties but varying heights or shapes. The integration of the two modalities enables a more comprehensive scene understanding, leading to improved classification performance and robustness.

Traditional multi-modal classification approaches typically employ feature-level or decision-level fusion strategies [12]–[16]. Feature-level fusion involves concatenating features extracted from different modalities, while decision-level fusion combines the outputs of independent classifiers trained on each modality. However, these methods often fail to fully exploit the intricate interactions between different data sources. With the advent of deep learning, more advanced architectures have been introduced to address these limitations [17]. Particularly, convolutional neural networks (CNNs) have been widely adopted for extracting hierarchical features from both HSI and LiDAR data [18], [19]. Models such as coupled CNNs [20], dual-branch CNNs [21], and attention-based fusion networks [22]–[25] have demonstrated superior performance by effectively learning joint representations of multi-modal data. These approaches leverage the ability of deep learning to capture both spatial and spectral dependencies, resulting in state-of-the-art classification accuracy [26], [27].

Despite the success of deep learning-based multi-modal classification techniques, a major challenge persists—the reliance on large amounts of labeled data [28], [29]. Acquiring labeled data in the RS field is particularly challenging due to the high costs and time-intensive nature of manual annotation, which often requires domain expertise [30], [31]. As a result, the above supervised learning methods are constrained by



(a) (b)

Fig. 1. Different sampling strategies for unlabeled training data. (a) The random sampling strategy. (b) The SPBS strategy.

the scarcity of labeled samples, leading to overfitting and suboptimal generalization to unseen test data [32].

To address this challenge, self-supervised and semi-supervised learning techniques have emerged as promising alternatives [33]–[35]. Among them, contrastive learning (CL) has gained significant attention for its ability to learn meaningful feature representations from unlabeled data. The popular CL frameworks such as Momentum Contrast (MoCo) [36] and SimCLR [37] have been successfully applied in various domains, including natural image processing and RS. The FusDreamer [38] built a label-efficient RS framework using world-model-driven representation learning combined with contrastive regularization. By leveraging the CL, models can learn robust and discriminative feature representations without requiring extensive labeled data.

Although the CL methods have alleviated the problem of insufficient labeled samples to some extent, they still suffer from class imbalance among the unlabeled samples during training, which may lead to suboptimal classification performance for certain categories. This imbalance arises from the inherent skew in the unlabeled data distribution, where certain classes are overrepresented while others are severely underrepresented. Such bias results in a dominance of majority-class instances in the contrastive loss, ultimately degrading minority-class representation. Some researchers have recognized this issue and proposed corresponding solutions. For instance, Cai *et al.* [39] introduced a classification model based on Bole convolution and a three-directional attention mechanism, which addresses the challenges of small sample sizes and class imbalance through a feature reward-penalty mechanism and multi-directional attention weighting. However, this method adopts post-selection adjustment strategies after randomly selecting training samples and does not address the root cause of class imbalance at the sample selection stage. In contrast, we propose a superpixel-based balanced sampling (SPBS) strategy that tackles this issue at the source. By enforcing class-balanced selection of pseudo-labeled training samples, our approach ensures a fairer class distribution during the sampling process, leading to more stable and equitable representation learning. As shown in Fig. 1, the different-colored plus signs denote the ground-truth classes of the unlabeled samples. The traditional random sampling strategy (Fig. 1 (a)) often results in a severely imbalanced class distribution. By comparison, our SPBS strategy (Fig. 1 (b)) significantly alleviates this problem.

It is also worth noting that in the CL, constructing high-quality positive and negative samples plays a crucial role in learning meaningful feature representations. Most existing methods in the RS field follow the general paradigm established in natural image processing, where positive pairs are formed via data augmentations such as cropping, flipping, or spectral jittering [40]. While such approaches are effective in certain scenarios, they often fail to capture the semantic relationships inherent in multi-modal RS data. Specifically, they overlook the complementary nature of HSI and LiDAR modalities, which offer distinct but spatially aligned information.

Additionally, the conventional CL methods commonly treat augmented views of the same sample as positive pairs and all others as negatives, without considering the underlying class semantics [41]. This simple assumption may lead to suboptimal representation learning, especially when samples from the same class are mistakenly pushed apart in the latent space [42]. To address this, some recent works have attempted to incorporate semantic information into CL by leveraging pseudo-labels or clustering techniques [42], [43]. However, most of these approaches either ignore the confidence of the pseudo-labels or fail to integrate feature similarity metrics, limiting their effectiveness in reducing intra-class variance.

To address the aforementioned challenges, we propose a novel semi-supervised balanced contrastive learning (SemiBaCon) network, which comprises three key components: an SPBS module, a dual-stream encoder, and a pseudo-label guided contrastive learning (PLCL) module. In the SPBS module, the HSI is segmented into superpixels, within which the labels of a small number of labeled samples are propagated to generate pseudo-labels. These pseudo-labeled samples are then selected in a balanced manner to ensure a fair class distribution during training. The dual-stream encoder extracts multi-modal features from both labeled and unlabeled samples and constructs positive pairs by matching same-position patches from different modalities. The PLCL module integrates both supervised loss and the CL loss, and includes a pseudo-label guided mechanism, which leverages the pseudo-labels to further minimize intra-class feature distances. The main contributions of our network are summarized as follows:

1) An SPBS mechanism is introduced to address class imbalance among the unlabeled samples, where the pseudo-labels are generated via intra-region label propagation, ensuring a statistically balanced selection across categories.

2) To effectively extract the complementary features from the HSI and LiDAR data, a dual-stream encoder is designed, integrating CNNs and Transformers for hierarchical spectral-spatial and elevation feature representations, thereby enabling the construction of meaningful multi-modal positive pairs.

3) A PLCL strategy is proposed, in which both pseudo-label confidence and feature similarity are utilized to guide pair selection, thereby reducing the intra-class variance and improving the discriminative power of the learned representations.

4) The experimental evaluations conducted on three benchmark HSI-LiDAR datasets demonstrate the outstanding classification performance of the proposed SemiBaCon framework, especially under limited labeled data conditions.

The rest of the paper is structured as follows. Section II summarizes the related works on superpixels and the CL.

Section III details our proposed networks. Section IV introduces the datasets, parameter settings, ablation and comparison experiments. Section V gives the conclusion.

II. Related Work

A. Superpixels Preprocessing in RS Image Classification

Superpixels are perceptually meaningful, irregular regions formed by grouping adjacent pixels with similar texture, color, brightness, or other features. By aggregating pixels with high similarity, superpixels effectively reduce the number of primitives in an image, thereby lowering the computational complexity of subsequent image processing tasks. As such, they are widely employed as a preprocessing step in various image segmentation and classification algorithms.

Among the available superpixel generation methods, simple linear iterative clustering (SLIC) [44] is one of the most popular due to its simplicity, efficiency, and ability to produce compact and approximately uniform superpixels. The SLIC strikes a favorable balance between computational speed, boundary adherence, and regularity of shape. Another notable method, superpixels extracted via energy-driven sampling (SEEDS) [45], utilizes an energy-based color similarity function to iteratively generate superpixels, offering a robust and efficient segmentation approach.

In recent years, superpixel methods have become a common approach in RS image classification. For instance, Li *et al.* [46] adopted unsupervised superpixel embedding to provide additional shallow morphological spatial feature information for deep learning networks, reducing the pressure on the feature extraction network and enhancing the feature discrimination ability. Nartey *et al.* [47] proposed a graph convolutional network based on superpixels to extract HSI features through pixels at different levels. The above methods aim to provide additional discriminative features.

Unlike these methods, which primarily focus on enhancing feature representations, our approach leverages superpixels to generate pseudo-labels, enabling balanced selection of unlabeled samples. This not only improves the representativeness of the training data but also provides valuable distributional information to guide the CL process.

B. CL in RS Image Classification

The CL framework can learn representations from unlabeled data without relying on explicit labels. In recent years, several CL frameworks have gained prominent positions in various downstream tasks due to their effectiveness in learning powerful representations. MoCo [36] builds a dynamic dictionary that minimizes the distance between queries and positive keys and facilitates contrastive unsupervised learning. The loss function is shown as

$$\mathcal{L}_q = -\log \frac{\exp(q \cdot k_+/t)}{\sum_{i=0}^{k} \exp(q \cdot k_i/t)}. \tag{1}$$

where t is a temperature hyperparameter, q is the encoded query, k_+ is the positive key, and k_i are the keys in a dictionary. Additionally, SimCLR [37] is a powerful self-supervised CL framework. The core idea is to maximize the consistency between augmented views of the same instance while minimizing the consistency between views of different instances.

Building upon existing CL frameworks, Sheng *et al.* [48] proposed a novel approach that integrates manifold enhancement and self-supervised CL within a meta-learning paradigm. This approach addresses the challenge of limited labeled data in RS image classification by leveraging the underlying geometric structure of the data manifold. Wang *et al.* [49] introduced a nearest neighbor-based CL network that effectively exploits large-scale unlabeled data to learn more discriminative feature representations. Their approach incorporates a novel nearest neighbor-based data augmentation strategy, which enhances semantic consistency by utilizing local contextual information and better captures inter-modal semantic alignments.

Although the aforementioned methods show certain advantages in exploring the intrinsic features of unlabeled data, samples that should belong to the same class are treated as negative samples, which limits the performance of the model. Additionally, the CL in multi-modal RS image classification has rarely been explored. Therefore, we propose an improved multi-modal CL method. The same pixel in different modalities is set as a positive sample to better learn multi-modal features. Moreover, we consider the similarity between samples and design masks based on pseudo-labels generated by superpixels to further reduce the intra-class distance.

C. Pseudo-Labels in RS Image Classification

Pseudo-labels represent a pivotal semi-supervised learning technique that assigns labels to unlabeled data and then incorporates them into the training process [50]. According to the cluster assumption, data points located in high-density regions are more likely to belong to the same category, so their pseudo-labels are considered more reliable.

In scenarios with limited labeled RS data, pseudo-labels can effectively leverage the unlabeled data as additional supervision to enhance the generalization ability of the model on unseen samples. For instance, Chen *et al.* [51] proposed a hybrid distance pseudo-label generation method that uses iterative multi-scale superpixel segmentation to enhance the model's ability to learn more representative features and patterns. Similarly, Madeleine *et al.* [52] sought a robust and efficient label propagation method for semi-supervised learning. This label propagation-based method takes into account the data distribution of HSI, which is crucial for the subsequent construction of graphs.

Different from the aforementioned methods, our method incorporates pseudo-labels into the loss function to prevent same-class samples from forming negative pairs, thereby facilitating learning of the data distribution.

III. Methodology

The overall framework of the proposed SemiBaCon method is illustrated in Fig. 2. It begins with the SPBS module, where HSI data is segmented into superpixels, and pseudo-labels are generated via intra-region label propagation. A balanced subset

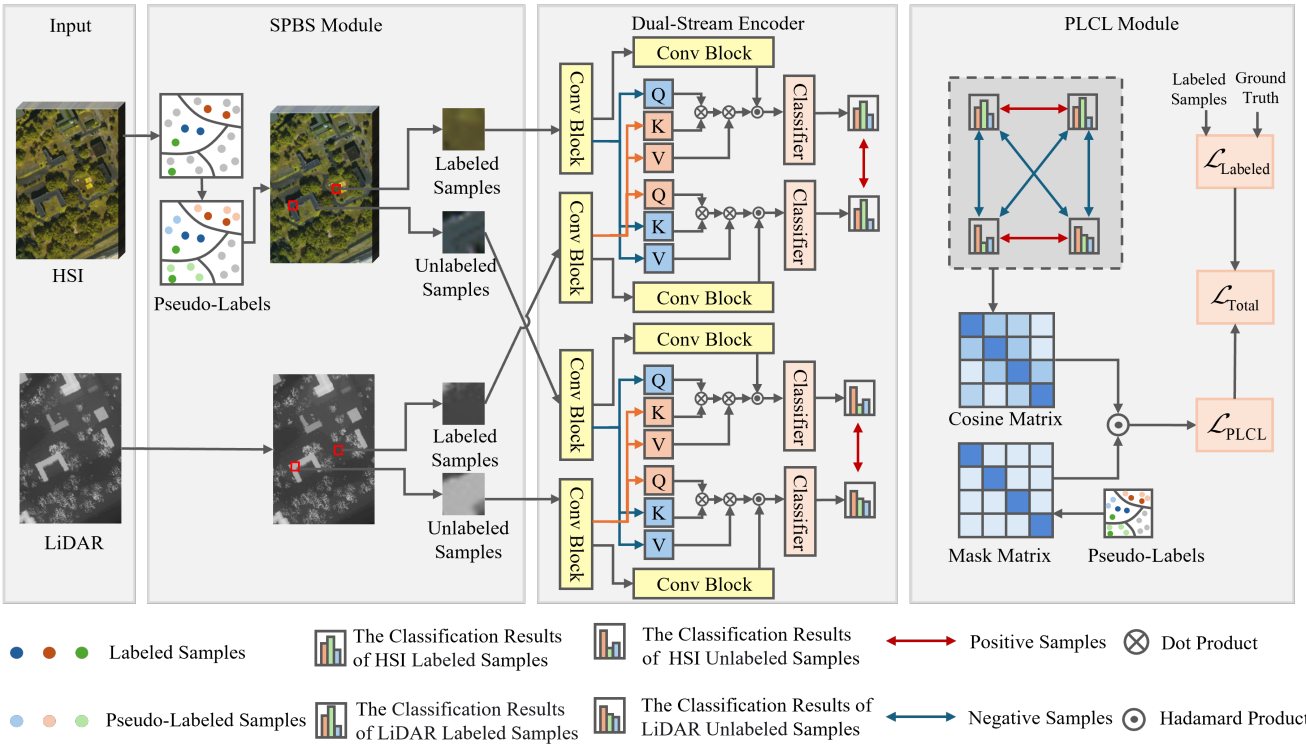


Fig. 2. Flowchart of the proposed SemiBaCon. The network consists of several modules: the SPBS module, the dual-stream encoder, and the PLCL module.

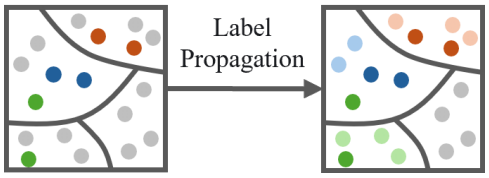


Fig. 3. Illustration of intra-region label propagation based on superpixels in the SPBS module.

of these pseudo-labeled samples is selected to address class imbalance. Subsequently, a dual-stream encoder that integrates CNNs and Transformers is employed to extract spectral-spatial and elevation features, which are also used to construct semantically meaningful positive pairs for the subsequent PLCL module. In the PLCL module, the supervised loss is computed from labeled samples, and the self-supervised loss leverages modality-consistent predictions at the same location as positive pairs to enhance cross-modal alignment. Additionally, the PLCL module refines features by integrating pseudo-label confidence and feature similarity, promoting intra-class compactness, thereby improving the discriminative power of the learned feature representations.

A. Superpixel-Based Balanced Sampling Module

To enhance the classification performance of minority-class samples, the SPBS module generates pseudo-labels based on superpixels and employs a balanced sampling strategy for unlabeled samples. Firstly, the SPBS module utilizes the SLIC algorithm [44] to segment the HSI into non-overlapping superpixels, which are homogeneous regions of varying shapes and sizes. Then, the superpixels containing labeled samples are selected. According to the characteristic that each superpixel block has high homogeneity inside, we roughly classify the samples in the superpixel into one category. Specifically, the indices of the superpixels with labeled pixels are recorded in a list with dimensions of $numClass \times num\_sp$. Subsequently, the number and category of labeled pixels in each superpixel are counted. Within a superpixel, the label of the majority category is propagated to other pixels, as shown in Fig. 3. An optimal number of superpixels is essential, as too many can increase model complexity, while too few may reduce accuracy by merging dissimilar pixels, as discussed in the parameter tuning section.

Based on superpixel segmentation and label propagation, an equal number of labeled samples and newly labeled samples from each class are selected for training. This process can effectively alleviate the class imbalance problem and improve the classification performance of the model. It is important to note that the number of unlabeled samples affects model performance: too few restrict feature learning, while too many may increase training time and reduce efficiency.

B. Dual-Stream Encoder

To effectively extract complementary features from the HSI and LiDAR data, a dual-stream encoder is designed, integrating CNNs and Transformers for hierarchical spectral-spatial and elevation feature representation, which enables the construction of meaningful multi-modal positive pairs.

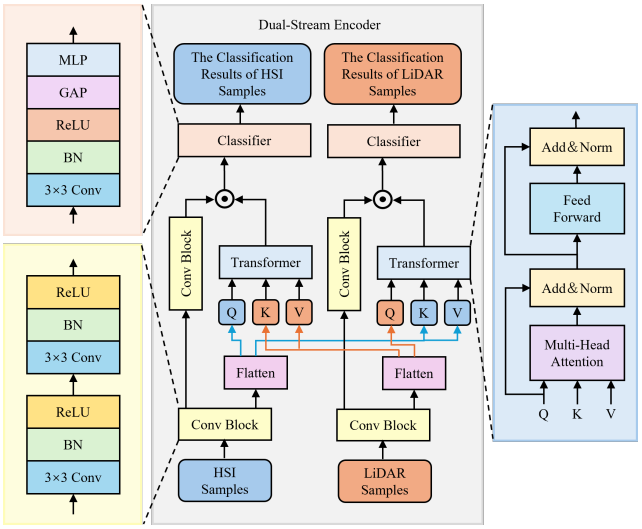


Fig. 4. Structure of the dual-stream encoder.

Let us denote the labeled inputs of the HSI as $\mathbf{X}_{\mathrm{LH}} \in \mathbb{R}^{ps\times ps\times C_{\mathrm{H}}}$, and those of the LiDAR as $\mathbf{X}_{\mathrm{LL}} \in \mathbb{R}^{ps\times ps\times C_{\mathrm{L}}}$, where ps represents the patch size of the inputs, and C_{H} and C_{L} represent the number of channels of the multi-modal data. The unlabeled inputs are $\mathbf{X}_{\mathrm{UH}} \in \mathbb{R}^{ps\times ps\times C_{\mathrm{H}}}$ for the HSI, and $\mathbf{X}_{\mathrm{UL}} \in \mathbb{R}^{ps\times ps\times C_{\mathrm{L}}}$ for the LiDAR. These patches are fed into the dual-stream encoder, as shown in Fig. 4. The dual-stream encoder has two levels of Conv Blocks. The basic component of the Conv Block is the 2D convolution with a kernel size of 3×3, batch normalization, and rectified linear unit (ReLU) activation. The number of convolutional kernels is empirically set to 64 for each block. The process of the single layer Conv Block can be represented as

$$outputs = \mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}(inputs))), \tag{2}$$

where the inputs can be $\mathbf{X}_{\mathrm{LH}}$, $\mathbf{X}_{\mathrm{LL}}$, $\mathbf{X}_{\mathrm{UH}}$, and $\mathbf{X}_{\mathrm{UL}}$.

Then, the outputs of the previous layer are converted into tokens, denoted as Q, K, and V. For the HSI branch, $\mathbf{Q} \in \mathbb{R}^{ps^2\times C_{\mathrm{E}}}$ corresponds to the HSI, while $\mathbf{K} \in \mathbb{R}^{ps^2\times C_{\mathrm{E}}}$ and $\mathbf{V} \in \mathbb{R}^{ps^2\times C_{\mathrm{E}}}$ correspond to the LiDAR data. For the LiDAR branch, $\mathbf{Q} \in \mathbb{R}^{ps^2\times C_{\mathrm{E}}}$ corresponds to the LiDAR data, while $\mathbf{K} \in \mathbb{R}^{ps^2\times C_{\mathrm{E}}}$ and $\mathbf{V} \in \mathbb{R}^{ps^2\times C_{\mathrm{E}}}$ correspond to the HSI. In this part, cross-attention is applied to handle the semantic relationship between the two modalities, which are fused through the multi-head attention as

$$\mathrm{MHA}(\mathbf{Q}_h, \mathbf{K}_h, \mathbf{V}_h) = \mathrm{Softmax}(\frac{\mathbf{Q}_h\mathbf{K}_h^T\mathbf{V}_h}{\sqrt{d/head}}), \tag{3}$$

where $\mathbf{Q}_h$, $\mathbf{K}_h$, and $\mathbf{V}_h$ are the multi-head forms of $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$, respectively. The $\mathrm{Softmax}$ denotes the $\mathrm{Softmax}$ function. T represents the transpose operation. $head$ is the number of heads, and d is the spectral dimension of the input feature vector, which is set to 64 according to experience. Through the cross-attention mechanism in the dual-stream network, the model can form meaningful multi-modal positive pairs, which is crucial for the CL.

After the Hadamard product between the Transformer and another Conv Block, the resulting features are passed through a classifier consisting of three sequential blocks. The first block sequentially includes 2D convolutions with 1×1 kernel size, batch normalization, and ReLU activation operations. The number of convolutional filters is set to 32 to reduce the number of channels. The second block performs global average pooling to reduce the spatial dimensions to 1×1 and capture pixel-level category information. The third block is a fully connected layer that maps the channels to the predefined number of land cover classes Num. The outputs of the classification module are denoted as y, including the classification results of both labeled and unlabeled multi-modal samples.

In summary, the proposed dual-stream encoder, which integrates CNNs and Transformers, enables the effective extraction of complementary features from HSI and LiDAR data. This architecture not only enhances the ability to form meaningful multi-modal positive pairs but also improves the overall classification performance through a hierarchical feature representation strategy.

C. Pseudo-Label Guided Contrastive Learning Module

To address the problem of label restriction, a PLCL module is designed, combining supervised learning and self-supervised CL. The supervised learning component is trained using high-quality labeled samples. The cross-entropy loss between the classification results for labeled samples and the true labels is denoted as $\mathcal{L}_{\mathrm{Labeled}}$. The self-supervised CL part learns rich semantic information from unlabeled samples by constructing a contrastive loss function. This combination effectively leverages the limited labeled samples while extracting latent feature information from the unlabeled samples, thereby enhancing the generalization capability of the model.

Different from the data augmentation method, our proposed PLCL sets the data of different modalities at the same location as positive samples, achieving enhanced representation learning through inter-modal consistency constraints. The loss function for a positive pair of examples (i, j) can be written as

$$\ell_1(i,j) = -\log\frac{\exp(sim(y_i, y_j)/t)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]}\exp(sim(y_i, y_k)/t)}, \tag{4}$$

where t is the temperature parameter, $sim(\cdot)$ is the cosine similarity, $exp(\cdot)$ is an exponential function, N is batch size, and $\mathbb{1}$ is an indicator function evaluating to 1 if $k = i$. The loss function for the CL component is defined as

$$\mathcal{L}_{\mathrm{CL}} = \frac{1}{2N}\sum_{k=1}^{N}[\ell_1(2k-1, 2k) + \ell_1(2k, 2k-1)]. \tag{5}$$

While the traditional CL effectively minimizes the distance between augmented views of the same sample, it does not account for intra-class similarity among different samples. To address this limitation, we propose a PLCL method that refines the contrastive loss using pseudo-labels generated for

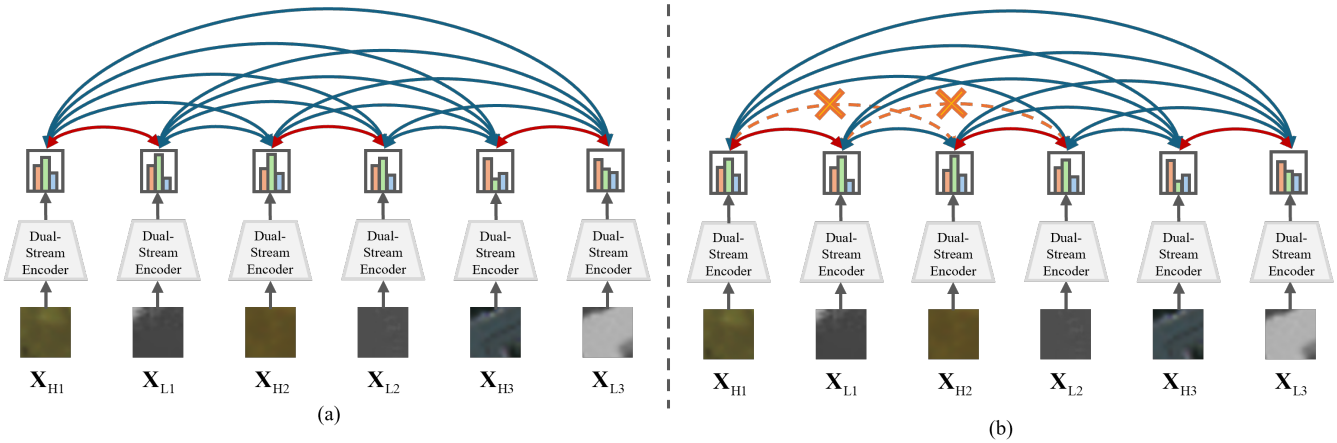


Fig. 5. Illustration of the difference between the ordinary CL and the proposed PLCL. (a) The ordinary CL. (b) The PLCL.

unlabeled data. This allows the model to avoid treating semantically similar samples as negatives, which in turn reduces intra-class variance and improves classification performance.

The difference between ordinary CL and PLCL is shown in Fig. 5. $(\mathbf{X}_{\mathrm{H1}}, \mathbf{X}_{\mathrm{L1}})$, $(\mathbf{X}_{\mathrm{H2}}, \mathbf{X}_{\mathrm{L2}})$, and $(\mathbf{X}_{\mathrm{H3}}, \mathbf{X}_{\mathrm{L3}})$ denote three groups of samples from different modalities. Among them, $(\mathbf{X}_{\mathrm{H1}}, \mathbf{X}_{\mathrm{L1}})$ and $(\mathbf{X}_{\mathrm{H2}}, \mathbf{X}_{\mathrm{L2}})$ belong to the same class. The red solid line, blue solid line, and orange dashed line represent positive, negative, and nonexistent sample pairs, respectively. Fig. 5 (a) depicts the standard CL scenario, while Fig. 5 (b) shows how the PLCL adjusts the contrastive relationships based on pseudo-labels. In the PLCL, the negative samples that share the same pseudo-label with the anchor are masked out from the denominator in Eq. (4). This step is equivalent to the process of Hadamard product of the cosine matrix produced by the CL and the mask matrix produced by pseudo-labels. The modified loss for a positive pair is given as

$$\ell_2(m,n) = -\log \frac{\exp(sim(y_m, y_n)/t)}{\sum_{k=1}^{2N} \mathbb{1}_{[q \neq m \vee pl_q \neq pl_m]} \exp(sim(y_m, y_q)/t)}, \tag{6}$$

where pl_q is the pseudo-label of the sample q, and pl_m is the pseudo-label of the sample m. The loss function for the pseudo-label guided CL component is computed as

$$\mathcal{L}_{\mathrm{PLCL}} = \frac{1}{2N} \sum_{k=1}^{N} [\ell_2(2k-1, 2k) + \ell_2(2k, 2k-1)]. \tag{7}$$

The overall loss function of the model is defined as

$$\mathcal{L}_{\mathrm{Total}} = \mathcal{L}_{\mathrm{Labeled}} + \mathcal{L}_{\mathrm{PLCL}}. \tag{8}$$

Among them, the CL is only used in the training stage. During the test stage, only the HSI branch is used, including multi-modal cross-attention.

IV. Experiments and Analysis

In this section, we conduct experiments on three widely adopted HSI and LiDAR datasets to evaluate the capability of the proposed method and compare it with other advanced models.

TABLE I
THE NUMBER OF TRAINING AND TEST SAMPLES FOR EACH CLASS IN AUGSBURG DATASET

No.	Class	Training	Test
1	Forest	20	13487
2	Residential Area	20	30309
3	Industrial Area	20	3831
4	Low Plants	20	26837
5	Allotment	20	555
6	Commercial Area	20	1625
7	Water	20	1510

TABLE II
THE NUMBER OF TRAINING AND TEST SAMPLES FOR EACH CLASS IN TRENTO DATASET

No.	Class	Training	Test
1	Apple Trees	20	4014
2	Building	20	2883
3	Ground	20	459
4	Woods	20	9103
5	Vineyard	20	10481
6	Roads	20	3154

A. Data Description

1) Augsburg Dataset: This dataset consists of two different data sources over the city of Augsburg, Germany, including an HSI and a DSM image. Both the HSI and DSM data were acquired by the DAS-EOC of DLR. They are collected by the HySpex sensor and the DLR-3 K system, respectively. To evaluate the performance of multi-modal fusion classification effectively, the spatial resolution of all images was downsampled to a uniform 30m GSD. The scene comprises 332×485 pixels and 180 spectral bands ranging from 0.4μm to 2.5μm for the HS image, and 1 band for the DSM image. Detailed information regarding the training and test sets is shown in Table I.

2) Trento Dataset: This dataset was collected in a rural area

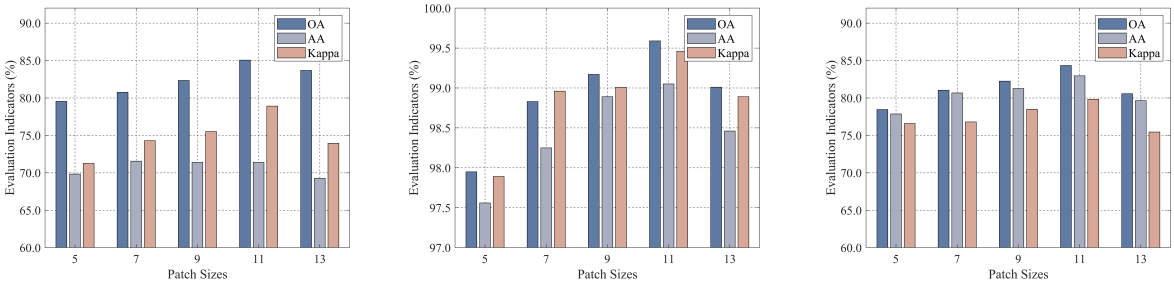


Fig. 6. Comparison of model performance across different patch sizes. (a) Augsburg. (b) Trento. (c) MUUFL.

TABLE III
THE NUMBER OF TRAINING AND TEST SAMPLES FOR EACH CLASS IN MUUFL DATASET

No.	Class	Training	Test
1	Trees	20	23226
2	Mostly Grass	20	4250
3	Mixed Ground Surface	20	6862
4	Dirt and Sand	20	1806
5	Road	20	6667
6	Water	20	446
7	Building Shadow	20	2213
8	Building	20	6220
9	Sidewalk	20	1365
10	Yellow Curb	20	163
11	Cloth Panels	20	249

in the southern part of Trento. It consists of a DSM derived from HSI and LiDAR data. The HSI data was acquired using the AISA Eagle sensor, which consists of 600×166 pixels with 63 spectral channels. The spectral range of the channels is from 0.42 μm to 0.99 μm. The LiDAR data was collected using the optical ALTM 3100 EA sensor. Both the HSI and LiDAR-derived DSM have a spatial resolution of 1 meter. This dataset focuses on land cover classification tasks for 6 different categories. Table II provides the training and test sample counts for the Trento dataset.

3) MUUFL Dataset: This dataset was captured in November 2010 over the Gulf Park campus of the University of Southern Mississippi. It consists of a DSM derived from HSI and LiDAR data. The HSI data was acquired using the ITRES CASI-1500 sensor with 64 available spectral channels ranging from 0.38 μm to 1.05 μm and a spatial resolution of 0.54×1.0 m. The LiDAR data was collected using the Gemini Airborne LiDAR System with a spatial resolution of 0.60×0.78 m. This dataset contains a total of 53687 ground-truth samples for 11 different categories. Table III provides the training and test sample counts for the MUUFL dataset.

B. Parameter Tuning

The experimental equipment is a Lenovo computer with an RTX3060 graphics card, 6GB of video random-access memory (VRAM). The version of CUDA is 11.7, and PyTorch is 2.0.1. For the optimization, we adopt the Adam algorithm, incorporating a learning rate decay factor of 0.9. Based on the observations, the number of training epochs is set to 200, as the model shows stable performance under this setting. We set the batch size to 64 because the same settings were used in most of the comparative experiments and achieved excellent performance. To objectively reflect the classification performance, we use overall accuracy (OA), average accuracy (AA), and Kappa coefficient as the evaluation criteria. All the evaluation metrics are averaged over ten classification results.

1) Input Patch Size: The patch size significantly impacts the performance of neural networks. Smaller patch sizes can capture finer local details and texture information in images. However, they may overlook global contextual information, resulting in a less comprehensive understanding of the overall scene. Additionally, there is an increased risk of overfitting to the training data, which can potentially affect the generalization ability of the model. Conversely, larger patch sizes can capture more global information, but they may dilute the distinction between positive and negative sample pairs. This is because larger patches may encompass multiple types of objects or features, making it harder for the model to learn discriminative boundaries between classes. Given these considerations, selecting an appropriate patch size is crucial. To determine the optimal patch size, we conduct experiments on three datasets. As shown in Fig. 6, the results indicate that a patch size of 11×11 yields the best network performance on all three datasets. Therefore, we chose this setting in the following experiments.

2) The Number of Superpixels and Unlabeled Samples: The number of superpixels and unlabeled samples both affect the accuracy of the model. If the number of superpixels is too small, irrelevant pixels may be grouped together, resulting in misclassification and unbalanced sample selection. On the other hand, too many superpixels can make label propagation more difficult and increase computational complexity. In addition, if the number of unlabeled samples is too small, the model may overfit the labeled data. Conversely, an excessive number of unlabeled samples may introduce noise and incorrect labels, resulting in degraded model performance. We conduct experiments to determine the appropriate hyperparameters. As shown in Fig. 7, when the number of superpixels is 600 and the number of unlabeled samples is 50, the model

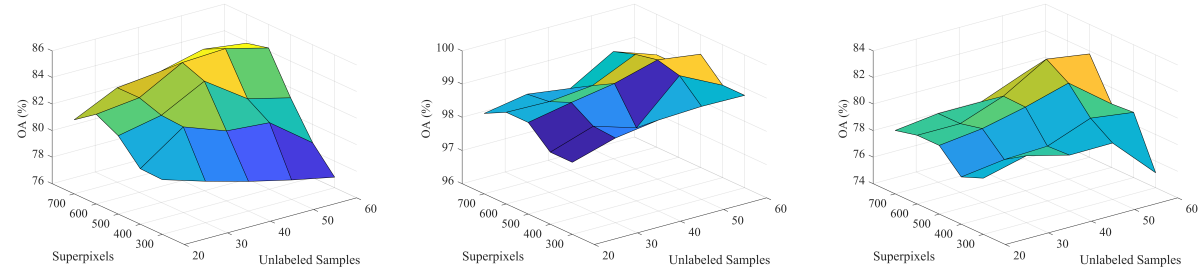


Fig. 7. Comparison of model performance across varying parameters. (a) Augsburg. (b) Trento. (c) MUUFL.

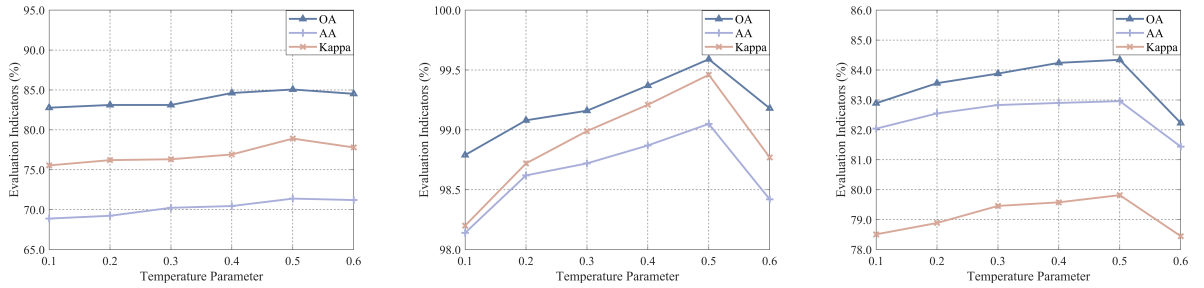


Fig. 8. Comparison of model performance across varying temperature parameters. (a) Augsburg. (b) Trento. (c) MUUFL.

TABLE IV
COMPARISON OF THE PERFORMANCE OF DIFFERENT METHODS ON THE AUGSBURG DATASET

Class No.	TBCNN [21]	EndNet [16]	MDL-RS [19]	MAHiDFNet [22]	DSHFNet [23]	GLT-Net [24]	MFT [25]	NNCNet [49]	FusDreamer [38]	SemiBaCon
1	90.27±03.57	72.72±02.17	96.38±01.67	96.93±00.86	**98.09±00.14**	98.03±00.56	93.04±02.27	97.38±00.62	96.46±00.71	77.89±02.62
2	74.36±03.24	74.85±09.68	83.33±05.14	80.95±03.59	78.10±04.56	85.11±02.03	80.42±02.87	84.86±01.79	82.46±01.99	**86.92±01.05**
3	37.39±06.36	26.79±12.36	44.33±08.12	44.47±05.89	46.88±09.71	89.44±01.52	85.02±01.47	46.70±07.06	49.75±06.59	**89.61±01.23**
4	56.19±04.31	56.26±04.82	79.03±04.16	69.48±05.82	69.73±05.82	56.12±06.81	71.28±01.11	73.17±01.15	80.39±00.89	**88.68±00.68**
5	62.07±12.05	**79.01±16.58**	71.17±10.44	76.52±02.36	75.12±02.36	78.86±02.44	25.45±00.50	78.74±01.75	65.95±02.56	52.19±08.52
6	52.27±02.66	47.56±10.43	48.15±09.16	51.48±07.21	**54.21±05.86**	28.65±17.98	22.46±02.17	49.60±05.21	50.83±06.79	51.62±03.56
7	51.82±03.50	61.65±03.78	61.46±04.26	60.21±01.73	60.33±02.57	59.07±02.56	61.84±01.73	61.85±01.82	61.88±01.02	**62.32±02.68**
OA (%)	68.07±01.57	64.82±02.75	81.03±02.65	76.94±01.44	76.29±01.90	75.75±02.60	77.49±00.98	79.91±00.72	81.51±00.65	**85.06±00.84**
AA (%)	60.62±02.00	60.75±02.39	69.69±02.42	68.60±01.53	68.93±01.71	69.18±02.21	60.66±01.06	70.33±01.45	70.53±01.23	**71.39±01.02**
Kappa×100	56.87±01.81	54.85±02.49	73.82±03.24	68.34±01.96	67.75±02.35	66.06±02.14	68.80±01.06	72.45±01.67	74.58±01.20	**78.91±01.53**

performs best on the Augsburg and MUUFL datasets. For the Trento dataset, the model performs best when the number of superpixels is 500 and the number of unlabeled samples is 50. These results suggest that adaptive adjustment of the number of superpixels based on dataset characteristics is necessary to achieve optimal performance. Furthermore, selecting an appropriate proportion of unlabeled samples can help maintain a good balance between generalization and robustness.

3) Temperature Parameter: The temperature parameter t in the contrastive loss plays a crucial role in controlling the smoothness of the similarity distribution between positive and negative pairs. A lower temperature sharpens the distribution, emphasizing harder negatives, while a higher temperature produces a softer distribution but may reduce discriminative learning. An inappropriate temperature may either suppress meaningful contrast or amplify noise. To determine the optimal value of t, we perform a grid search over the set $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$ on the validation set of the Augsburg dataset. As shown in Fig. 8, the performance across all datasets is compared under different values of t. The results indicate that $t = 0.5$ consistently provides the best trade-off between stability and discriminability. This selected value is therefore used across all experiments to ensure consistent and robust contrastive learning.

C. Comparison Experiments

To verify the advancement and effectiveness of the proposed SemiBaCon, we compare it with several state-of-the-art HSI and LiDAR fusion classification models, including the TBCNN [21], EndNet [16], multi-modal deep learning

TABLE V
COMPARISON OF THE PERFORMANCE OF DIFFERENT METHODS ON THE TRENTO DATASET

Class No.	TBCNN [21]	EndNet [16]	MDL-RS [19]	MAHiDFNet [22]	DSHFNet [23]	GLT-Net [24]	MFT [25]	NNCNet [49]	FusDreamer [38]	SemiBaCon
1	98.31±00.51	97.23±00.82	99.79±00.12	99.63±00.12	99.55±00.13	98.93±00.71	97.21±00.66	**99.83±00.09**	99.78±00.12	99.68±00.12
2	77.24±02.37	73.47±01.65	97.38±00.97	92.10±01.42	**99.31±00.06**	97.40±01.63	89.76±03.43	98.33±00.88	98.72±00.68	99.20±00.02
3	**99.52±00.21**	99.51±00.18	98.21±00.43	98.21±00.40	98.91±00.39	98.10±00.72	95.71±01.80	98.26±00.45	98.26±00.44	97.82±00.68
4	99.99±00.00	97.47±00.47	**100.00±00.00**	**100.00±00.00**	99.57±00.04	**100.00±00.00**	99.97±00.06	**100.00±00.00**	**100.00±00.00**	**100.00±00.00**
5	98.42±00.62	92.27±01.86	99.89±00.11	99.93±00.03	99.90±00.02	99.98±00.03	99.76±00.21	99.84±00.12	99.98±00.01	**100.00±00.00**
6	76.74±02.19	78.44±01.70	94.89±01.45	93.15±01.08	82.66±02.15	95.45±01.01	93.01±01.94	93.21±01.96	92.49±01.26	**97.59±00.52**
OA (%)	94.60±00.22	91.37±01.13	99.12±00.14	98.42±00.19	97.88±00.16	99.11±00.17	97.76±00.54	99.02±00.13	99.03±00.12	**99.59±00.11**
AA (%)	91.71±00.27	89.74±00.96	98.36±00.18	97.17±00.31	96.65±00.19	98.33±00.31	95.90±00.93	98.25±00.23	98.20±00.26	**99.05±00.13**
Kappa×100	92.81±00.29	88.55±01.35	98.83±00.19	97.89±00.25	97.16±00.21	98.80±00.23	97.01±00.74	98.70±00.24	98.71±00.24	**99.46±00.12**

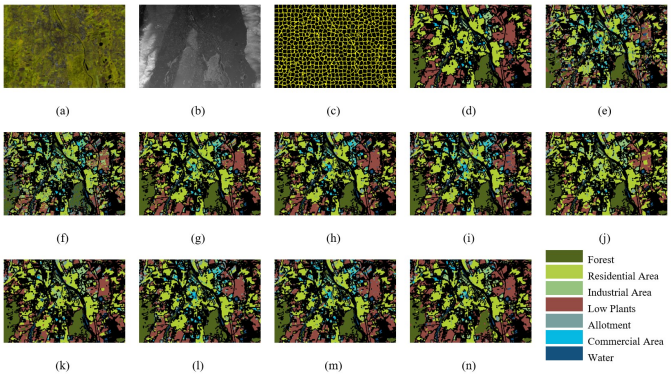


Fig. 9. Classification maps for the Augsburg dataset obtained with different methods. (a) Pseudo-color image for HSI. (b) LiDAR-based DSM. (c) Superpixel boundaries. (d) Groundtruth. (e) TBCNN (68.07%). (f) EndNet (64.82%). (g) MDL-RS (81.03%). (h) MAHiDFNet (76.94%). (i) DSHFNet (76.29%). (j) GLT-Net (75.75%). (k) MFT (77.49%). (l) NNCNet (79.91%). (m) FusDreamer (81.51%). (n) SemiBaCon (85.06%).

framework (MDL-RS) [19], multi-attentive hierarchical fusion net (MAHiDFNet) [22], dynamic-scale hierarchical fusion network (DSHFNet) [23], GLT-Net [24], MFT [25], NNCNet [49], and the FusDreamer [38]. Among them, the NNCNet and FusDreamer are semi-supervised methods. In order to ensure the fairness of the experiments, all methods adopt the optimal parameters mentioned in the papers. We adopt the same training and test samples, with experiments in each network running 200 epochs and ten independent executions. The mean results with the standard deviation are presented, which can reduce the error caused by individual experimental changes.

The experimental results include both quantitative analysis and visualizations. The quantitative analysis results are presented from Table IV to VI, with the best accuracy for each row highlighted in bold. The visual results are depicted from Fig. 9 to Fig. 11, displaying the pseudo-color images for the HSI, LiDAR images, ground-truth images, and classification maps obtained by different methods. The colors used to represent each class are also indicated in the figures.

The Augsburg dataset consists of seven classes and exhibits complex scenes. The number of pixels in the Industrial Area, Allotment, Commercial Area, and Water categories is relatively small, making them more difficult to classify compared to categories with a larger number of pixels. From Fig. 9 and Table IV, we can draw the following conclusions. Firstly, TBCNN and EndNet pay more attention to the characteristics of the single input pixel, ignoring the global features, so there are obvious noise points in the classification results. Secondly, the three CNN-based methods, MDL-RS, MAHiDFNet, and DSHFNet, have achieved certain progress in local feature extraction, which is manifested in the reduction of noise points. However, there is a significant issue as some pixels in the Industrial Area are incorrectly classified as belonging to the Commercial Area. This might be due to the inability of the network to effectively distinguish between similar features. Thirdly, the GLT-Net and MFT utilized Transformers to obtain global features, achieving progress in distinguishing the Industrial Area from the surrounding pixels. However, the classification effectiveness of these two methods is still limited for categories with a smaller number of pixels. Moreover, the two semi-supervised methods enhance multi-modal feature alignment and discriminability through contrastive learning, but still exhibit limited robustness in scenarios with severe class imbalance or noisy pseudo-labels.

Notably, our proposed SemiBaCon achieves an OA of 85.06%, AA of 71.39%, and Kappa of 78.91%, which are 3.55%, 0.86%, and 4.33% higher than the suboptimal approach, respectively. The classification results for the Residential Area, Industrial Area, Low Plants, and Water are superior to those of other methods. It is worth noting that our method performs well in categories with fewer pixels, benefiting from the balanced sampling strategy. Moreover, our CL strategy can effectively extract features of pixels from different categories.

For the Trento dataset, as shown in Table V, the SemiBaCon achieves an OA of 99.59%, AA of 99.05%, and Kappa of 99.46%, which are 0.47%, 0.69%, and 0.63% higher than the suboptimal results, respectively. As illustrated in Fig. 10, the features of the Building and the Roads are spatially close, and both contain relatively few pixels, which increases the difficulty of classification. However, our method can effectively distinguish between these two categories. This is possibly because the PLCL strategy effectively leverages the inherent features of the training pixels.

Table VI presents the classification accuracies of all comparisons on the MUUFL dataset. The SemiBaCon achieves the highest OA, AA, and Kappa of 84.34%, 82.96%, and 79.82%,

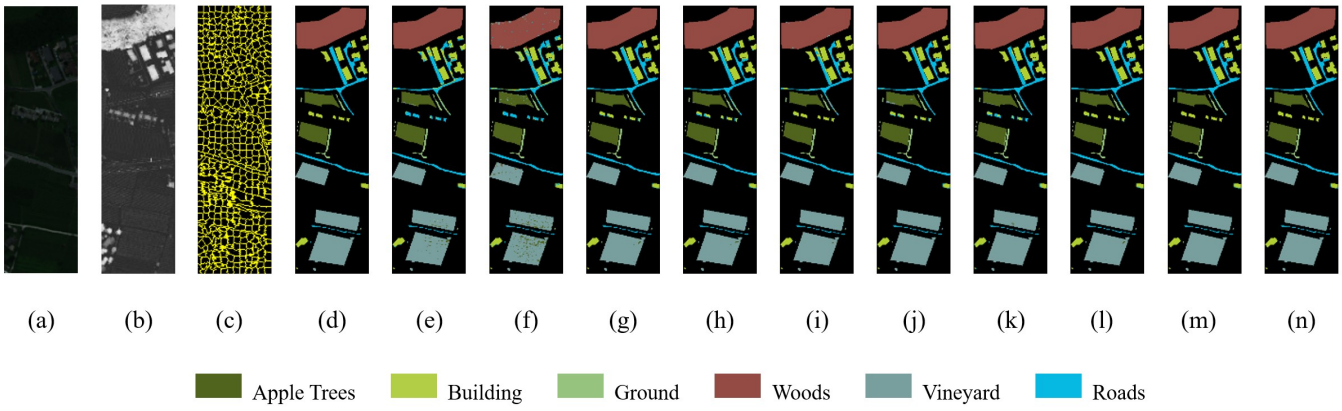


Fig. 10. Classification maps for the Trento dataset obtained with different methods. (a) Pseudo-color image for HSI. (b) LiDAR-based DSM. (c) Superpixel boundaries. (d) Ground-truth. (e) TBCNN (94.60%). (f) EndNet (91.37%). (g) MDL-RS (99.12%). (h) MAHiDFNet (98.42%). (i) DSHFNet (97.88%). (j) GLT-Net (99.11%). (k) MFT (97.76%). (l) NNCNet (99.02%). (m) FusDreamer (99.03%). (n) SemiBaCon (99.59%).

TABLE VI
COMPARISON OF THE PERFORMANCE OF DIFFERENT METHODS ON THE MUUFL DATASET

Class No.	TBCNN [21]	EndNet [16]	MDL-RS [19]	MAHiDFNet [22]	DSHFNet [23]	GLT-Net [24]	MFT [25]	NNCNet [49]	FusDreamer [38]	SemiBaCon
1	83.70±01.41	80.77±02.49	86.01±01.29	84.61±01.58	83.54±00.87	87.65±03.21	83.20±02.18	87.01±00.93	86.07±01.02	**87.96±00.89**
2	74.81±03.51	**78.86±04.86**	72.55±03.23	78.42±03.26	72.18±00.74	76.54±16.85	75.02±03.63	69.86±04.56	75.32±03.66	76.19±01.31
3	68.79±03.99	56.29±08.71	66.46±04.83	68.07±04.24	70.85±03.45	70.06±08.80	65.23±04.32	70.56±02.31	67.02±04.82	**71.33±02.27**
4	83.73±04.21	77.58±07.90	88.13±05.29	92.58±01.02	84.64±00.58	89.24±15.65	**95.61±02.67**	88.26±01.49	91.86±01.89	89.22±01.48
5	78.86±01.57	82.04±01.28	84.10±01.81	86.70±01.29	85.07±04.67	**89.34±04.50**	82.22±02.31	83.46±02.35	86.19±01.78	87.23±01.07
6	83.27±01.60	94.15±01.47	99.96±00.08	98.26±00.12	**100.00±00.00**	99.30±01.12	99.71±00.40	**100.00±00.00**	**100.00±00.00**	98.86±00.83
7	80.11±01.41	78.37±05.18	86.80±01.55	84.63±01.28	86.40±03.89	79.78±10.58	83.03±04.25	**87.39±01.20**	86.35±01.42	86.28±01.09
8	78.04±01.59	81.11±02.65	92.37±01.23	90.50±00.67	93.86±01.25	**95.59±01.66**	93.17±00.84	90.26±00.68	91.40±01.11	92.89±00.66
9	48.09±03.77	**52.74±05.56**	48.40±02.97	48.13±04.81	49.81±07.51	33.79±08.40	37.17±04.98	49.08±01.32	52.38±04.86	49.13±01.10
10	76.87±01.86	76.49±05.64	73.39±04.70	76.69±02.74	69.88±07.32	56.72±09.36	54.29±07.26	62.58±08.26	70.01±05.63	**76.93±01.14**
11	97.59±02.50	94.06±01.46	99.18±00.30	96.75±00.81	98.80±01.00	98.92±01.06	98.47±01.50	99.20±00.32	99.27±00.03	**99.30±00.05**
OA (%)	77.38±00.49	73.25±00.83	81.68±00.70	82.46±00.45	81.77±01.77	82.66±01.50	79.07±01.13	82.65±00.47	82.89±00.52	**84.34±00.31**
AA (%)	76.71±00.42	72.86±01.30	81.27±00.89	82.73±01.63	81.23±02.37	79.23±02.92	77.01±00.80	80.70±01.65	82.65±00.32	**82.96±00.39**
Kappa×100	74.72±00.55	72.39±01.00	77.26±00.88	77.54±00.56	77.89±02.09	77.54±01.84	72.20±01.32	77.70±00.89	78.06±00.70	**79.82±00.37**

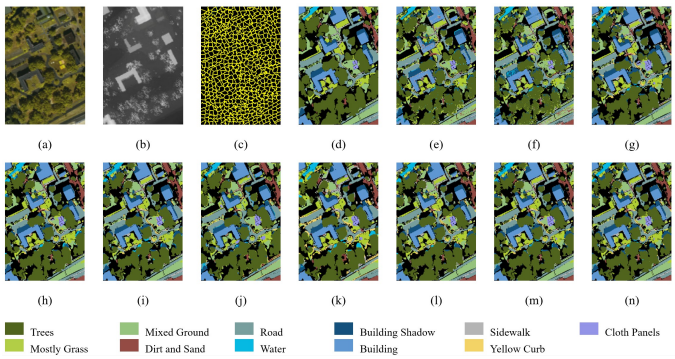


Fig. 11. Classification maps for the MUUFL dataset obtained with different methods. (a) Pseudo-color image for HSI. (b) LiDAR-based DSM. (c) Superpixel boundaries. (d) Ground-truth. (e) TBCNN (77.38%). (f) EndNet (73.25%). (g) MDL-RS (81.68%). (h) MAHiDFNet (82.46%). (i) DSHFNet (81.77%). (j) GLT-Net (82.66%). (k) MFT (79.07%). (l) NNCNet (82.65%). (m) FusDreamer (82.89%). (n) SemiBaCon (84.34%).

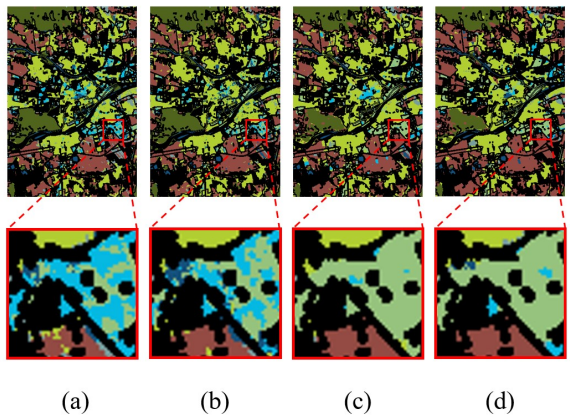


Fig. 12. Visualization results of the ablation study on the Augsburg dataset. (a) The supervised network. (b) The network with ordinary CL. (c) The network with the SPBS strategy and ordinary CL. (d) SemiBaCon.

respectively, among all methods. It is evident that the Yellow Curb and Cloth Panels have a very limited number of pixels in this dataset. Our method achieves the best classification results for these two categories. This is because the balanced sampling

TABLE VII
ABLATION ANALYSIS OF BS, CL AND PL IN TERMS OF OA (%), AA (%), AND KAPPA (×100) ON AUGSBURG, TRENTO, AND MUUFL DATASETS

Module			Augsburg			Trento			MUUFL		
BS	CL	PL	OA (%)	AA (%)	Kappa (×100)	OA (%)	AA (%)	Kappa (×100)	OA (%)	AA (%)	Kappa (×100)
×	×	×	81.84±00.98	69.56±01.15	77.43±01.58	98.01±00.15	95.25±00.20	97.89±00.16	82.56±00.37	81.84±00.42	78.99±00.41
×	✓	×	83.19±00.89	70.61±01.12	78.01±01.55	98.56±00.13	97.03±00.17	98.41±00.14	83.01±00.35	82.08±00.36	78.34±00.42
✓	✓	×	84.52±00.82	70.89±01.05	78.45±01.49	99.12±00.13	98.42±00.12	98.87±00.15	82.79±00.30	82.19±00.38	78.91±00.39
✓	✓	✓	**85.06±00.84**	**71.39±01.02**	**78.91±01.53**	**99.59±00.11**	**99.05±00.13**	**99.46±00.12**	**84.34±00.31**	**82.96±00.39**	**79.82±00.37**

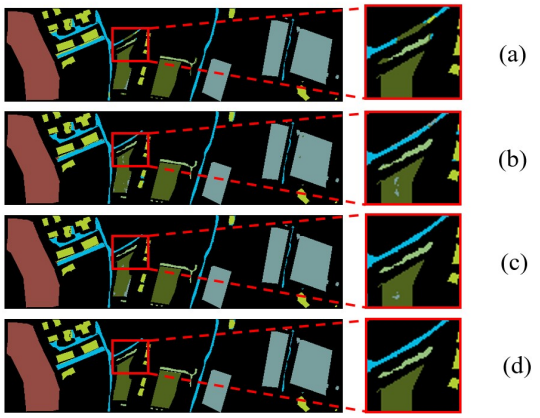


Fig. 13. Visualization results of the ablation study on the Trento dataset. (a) The supervised network. (b) The network with ordinary CL. (c) The network with the SPBS strategy and ordinary CL. (d) SemiBaCon.

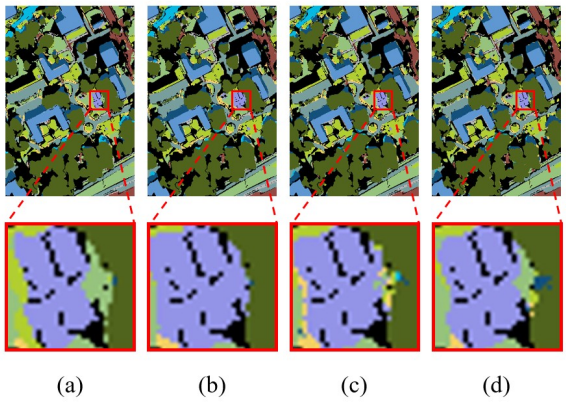


Fig. 14. Visualization results of the ablation study on the MUUFL dataset. (a) The supervised network. (b) The network with ordinary CL. (c) The network with the SPBS strategy and ordinary CL. (d) SemiBaCon.

strategy enables categories with fewer samples to participate in the semi-supervised learning process.

D. Ablation Studies

To verify the effectiveness of the proposed components, we conduct ablation studies on the balanced sampling (BS), CL, and pseudo-labels (PS) components. The ablation results are presented in Table VII, where the symbols ✓ and × under BS, CL, and PS indicate whether the corresponding component is enabled. The first row of Table VII corresponds to a purely supervised learning baseline. The second row reflects the performance of a semi-supervised framework using random sampling without CL. The third row includes the balanced sampling strategy, while the last row reports the results of the complete SemiBaCon framework. The performance improvements across the rows validate the effectiveness of each proposed module. The visualization results of the ablation studies are shown in Fig. 12 to Fig. 14. In Fig. 12, it can be observed that the purely supervised model misclassifies parts of the Industrial Area as the Commercial Area. After introducing the contrastive learning module, such misclassifications are significantly reduced, indicating that contrastive learning enhances the discriminability between classes. Furthermore, when the SPBS strategy is integrated, the classification accuracy of the minority class Industrial Area is further improved. This enhancement can be attributed to the balanced sampling strategy, which increases the representation of minority classes in the construction of contrastive pairs, thereby strengthening their feature representations in the embedding space. In Fig. 13, the introduction of the SPBS strategy leads to an improvement in the classification accuracy of the minority class Ground, indicating that balanced sampling effectively enhances the representation of underrepresented classes during training. Building on this, the complete SemiBaCon framework further reduces the intra-class variance of both the Ground and Roads categories. In Fig. 14, the accurate classification of the minority class, Cloth Panels, further validates the effectiveness of the SPBS strategy and the PLCL module.

Beyond this, we investigate the impact of the positive sample selection strategy in the CL. The conventional approaches typically construct positive pairs by applying single-modal data augmentation techniques, such as noise injection or rotation. In contrast, the proposed SemiBaCon framework considers multi-modal patches with the same location as positive pairs, thereby effectively leveraging the complementary characteristics of the HSI and LiDAR data. As illustrated in Fig. 15, our strategy results in more discriminative feature representations and leads to improved classification performance.

Another noteworthy design is that we exclude patches with the same pseudo-label from the set of negative samples, but intentionally do not consider them as positive samples. This is because samples with the same pseudo-label may share common features but also introduce noise. We conduct

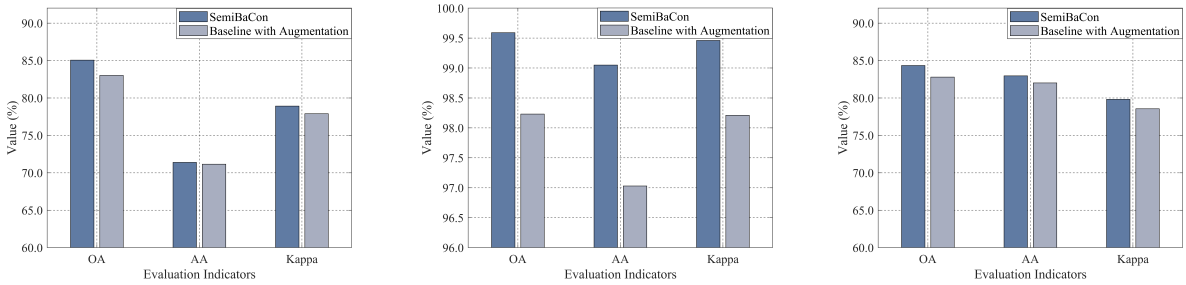


Fig. 15. Classification performance with different CL strategies. (a) Augsburg. (b) Trento. (c) MUUFL.

TABLE VIII
PERFORMANCE COMPARISON OF CONTRASTIVE PAIR SELECTION STRATEGIES ON AUGSBURG, TRENTO, AND MUUFL DATASETS

Strategy	Augsburg			Trento			MUUFL		
	OA (%)	AA (%)	Kappa (×100)	OA (%)	AA (%)	Kappa (×100)	OA (%)	AA (%)	Kappa (×100)
Standard CL	84.52±00.82	70.89±01.05	78.45±01.49	99.12±00.13	98.42±00.12	98.87±00.15	82.79±00.30	82.19±00.38	78.91±00.39
PLCL-Include	84.47±00.74	70.79±01.03	78.36±01.47	99.21±00.11	98.45±00.15	98.99±00.14	83.48±00.29	82.57±00.31	78.93±00.36
PLCL-Omit (Ours)	**85.06±00.84**	**71.39±01.02**	**78.91±01.53**	**99.59±00.11**	**99.05±00.13**	**99.46±00.12**	**84.34±00.31**	**82.96±00.39**	**79.82±00.37**

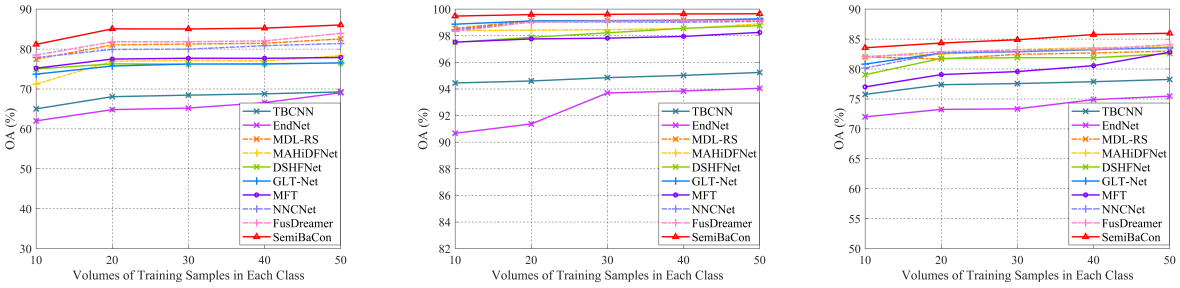


Fig. 16. Classification performance with different volumes of training samples. (a) Augsburg. (b) Trento. (c) MUUFL.

ablation experiments to validate this design, as shown in Table VIII. The first row represents the standard CL strategy, where patches from the same location but different modalities are used as positive pairs. In the second row, patches with the same pseudo-label are excluded from the negative set and simultaneously treated as positive pairs. In the third row, these patches are also excluded from the negative set but are not included in the positive set. The experimental results demonstrate the effectiveness of our design.

E. Performance with Different Training Sample Volumes

To further validate the robustness and effectiveness of the proposed method, we investigate the classification accuracy of the model when the sample size is smaller. Specifically, we randomly select a fixed percentage of samples from each category in the training set for training, while keeping the test set unchanged. The number of training samples is increased from 10 to 50. Fig. 16 shows the varying classification performance of the proposed SemiBaCon model and other comparison methods, described by the average OA over ten independent experiments. From Fig. 16, it can be seen that as the number of training samples decreases, the classification accuracies of all methods decrease, indicating that the models rely on sufficient sample sizes for training. It is worth noting that our model always achieves optimal performance when the number of training samples is the same, which proves the robustness of the framework.

F. Computational Complexity Analysis

To comprehensively evaluate model performance, the computational complexity must also be considered. Specifically, the number of trainable parameters and computational cost are effective indicators for measuring the spatial and temporal complexity of the network, respectively. Fig. 17 (a)-(c) illustrate the OA, parameters and computational costs of the competitive networks on the three datasets. The horizontal coordinate represents the number of trainable parameters, the vertical coordinate represents the OA, and the bubble diameter represents the computational cost in million floating-point operations (MFLOPS). The following conclusions can

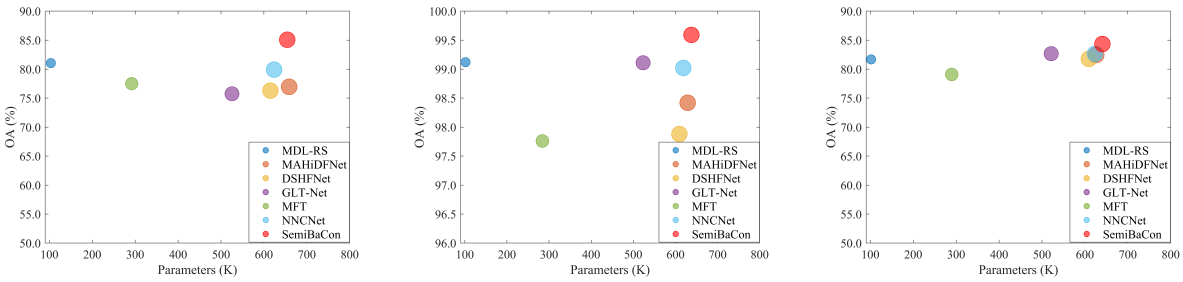


Fig. 17. Comparison of the networks on parameters, computational cost, and OA. (a) Augsburg. (b) Trento. (c) MUUFL.

be drawn from Fig. 17. The MFT has fewer parameters and computational requirements, but its accuracy is relatively low. The DSHFNet has a larger number of parameters and higher computational cost, with an accuracy that is somewhat improved but still below that of the SemiBaCon. The parameters and computational cost of the SemiBaCon are similar to those of the MAHiDFNet, but it achieves better accuracy. Overall, the SemiBaCon enhances accuracy while keeping the increase in parameters and computational cost within an acceptable range.

V. Conclusion

In this paper, a novel SemiBaCon framework is designed for multi-modal RS image classification using HSI and LiDAR data. The proposed network addresses the key challenges in the existing methods, including the heavy reliance on large-scale labeled data and class imbalance among unlabeled samples. By introducing the SPBS strategy, the framework ensures a more equitable distribution of unlabeled samples across classes. The dual-stream encoder enhances the capability of multi-modal feature learning by constructing positive pairs from patches at the same spatial location across different modalities. In addition, the introduction of the pseudo-labels in the PLCL module enhances intra-class feature compactness and helps reduce intra-class misclassification. The extensive experiments conducted on three public HSI-LiDAR datasets demonstrate that the proposed SemiBaCon achieves state-of-the-art classification performance, particularly under limited supervision, highlighting its effectiveness and potential for practical RS applications. Future work will explore the integration of active learning strategies and adaptive pseudo-label refinement to further improve model robustness and scalability.

References

[1] X. Wang, J. Zhu, Y. Feng, and L. Wang, "MS2CANet: Multiscale spatial–spectral cross-modal attention network for hyperspectral image and LiDAR classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024.

[2] B. Zhang, Y. Chen, S. Xiong, and X. Lu, "Hyperspectral image classification via cascaded spatial cross-attention network," *IEEE Transactions on Image Processing*, vol. 34, pp. 899–913, 2025.

[3] P. Wang, Z. He, B. Huang, M. Dalla Mura, H. Leung, and J. Chanussot, "VOGTNet: Variational optimization-guided two-stage network for multispectral and panchromatic image fusion," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 5, pp. 9268–9282, 2025.

[4] D. Hong, B. Zhang, H. Li, Y. Li, J. Yao, C. Li, M. Werner, J. Chanussot, A. Zipf, and X. X. Zhu, "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sensing of Environment*, vol. 299, p. 113856, 2023.

[5] X. Ma, X. Zhang, M.-O. Pun, and M. Liu, "A multilevel multimodal fusion Transformer for remote sensing semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024, Art. no. 5403215.

[6] Y. Zhang, S. Yan, L. Zhang, and B. Du, "Fast projected fuzzy clustering with anchor guidance for multimodal remote sensing imagery," *IEEE Transactions on Image Processing*, vol. 33, pp. 4640–4653, 2024.

[7] Y. Gu, Q. Wang, X. Jia, and J. A. Benediktsson, "A novel MKL model of integrating LiDAR data and MSI for urban area classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 10, pp. 5312–5326, 2015.

[8] W. Liao, A. Pižurica, R. Bellens, S. Gautama, and W. Philips, "Generalized graph-based fusion of hyperspectral and LiDAR data using morphological features," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 3, pp. 552–556, 2015.

[9] Q. Wang, J. Huang, S. Wang, Z. Zhang, T. Shen, and Y. Gu, "Community structure guided network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–15, 2025, Art. no. 4404115.

[10] B. Xi, Y. Zhang, J. Li, T. Zheng, X. Zhao, H. Xu, C. Xue, Y. Li, and J. Chanussot, "MCTGCL: Mixed CNN–Transformer for mars hyperspectral image classification with graph contrastive learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–14, 2025, Art. no. 5503214.

[11] J. Peng, L. Li, and Y. Y. Tang, "Maximum likelihood estimation-based joint sparse representation for the classification of hyperspectral remote sensing images," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 6, pp. 1790–1802, 2019.

[12] L. Song, Z. Feng, S. Yang, X. Zhang, and L. Jiao, "Discrepant bidirectional interaction fusion network for hyperspectral and LiDAR data classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[13] Y. He, B. Xi, G. Li, T. Zheng, Y. Li, C. Xue, and J. Chanussot, "Multilevel attention dynamic-scale network for HSI and LiDAR data fusion classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–16, 2024, Art. no. 5529916.

[14] P. Duan, S. Hu, X. Kang, and S. Li, "Shadow removal of hyperspectral remote sensing images with multiexposure fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022, Art. no. 5537211.

[15] P. Duan, T. Shan, X. Kang, and S. Li, "Spectral super-resolution in frequency domain," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–11, 2024.

[16] D. Hong, L. Gao, R. Hang, B. Zhang, and J. Chanussot, "Deep encoder–decoder networks for classification of hyperspectral and LiDAR data," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[17] Z. Zhang, L. Huang, Q. Wang, L. Jiang, Y. Qi, S. Wang, T. Shen, B.-H. Tang, and Y. Gu, "UAV hyperspectral remote sensing image classification: A systematic review," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 3099–3124, 2025.

[18] B. Xi, M. Cai, J. Li, Z. Wang, S. Feng, Y. Li, and J. Chanussot, "HyLiOSR: Staged progressive learning for joint open-set recognition

of hyperspectral and LiDAR data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–14, 2025, Art. no. 5506714.

[19] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 4340–4354, 2021.

[20] R. Hang, Z. Li, P. Ghamisi, D. Hong, G. Xia, and Q. Liu, "Classification of hyperspectral and LiDAR data using coupled CNNs," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 7, pp. 4939–4950, 2020.

[21] X. Xu, W. Li, Q. Ran, Q. Du, L. Gao, and B. Zhang, "Multisource remote sensing data classification based on convolutional neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 937–949, 2018.

[22] X. Wang, Y. Feng, R. Song, Z. Mu, and C. Song, "Multi-attentive hierarchical dense fusion net for fusion classification of hyperspectral and LiDAR data," *Information Fusion*, vol. 82, pp. 1–18, 2022.

[23] Y. Feng, L. Song, L. Wang, and X. Wang, "DSHFNet: Dynamic scale hierarchical fusion network based on multiattention for hyperspectral image and LiDAR data classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023, Art. no. 5522514.

[24] K. Ding, T. Lu, W. Fu, S. Li, and F. Ma, "Global–local Transformer network for HSI and LiDAR data joint classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022, Art. no. 5541213.

[25] S. K. Roy, A. Deria, D. Hong, B. Rasti, A. Plaza, and J. Chanussot, "Multimodal fusion Transformer for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–20, 2023, Art. no. 5515620.

[26] Q. Hu, Z. Shen, Z. Sha, and W. Tan, "Multiloss adversarial attacks for multimodal remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024, Art. no. 4600813.

[27] Z. Xu, W. Jiang, and J. Geng, "Texture-aware causal feature extraction network for multimodal remote sensing data classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–12, 2024, Art. no. 5103512.

[28] T. Lu, Y. Fang, W. Fu, K. Ding, and X. Kang, "Dual-stream class-adaptive network for semi-supervised hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–11, 2024, Art. no. 5507511.

[29] D. Xiu, Z. Pan, Y. Wu, and Y. Hu, "MAGE: Multisource attention network with discriminative graph and informative entities for classification of hyperspectral and LiDAR data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022, Art. no. 5539714.

[30] Y. Zhang and S. Yan, "Semi-supervised active learning image classification method based on tri-training algorithm," in *2020 IEEE International Conference on Artificial Intelligence and Information Systems (ICAIIS)*, 2020, pp. 206–210.

[31] J. Qu, L. Zhang, W. Dong, N. Li, and Y. Li, "Shared-private decoupling-based multilevel feature alignment semisupervised learning for HSI and LiDAR classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024, Art. no. 5537314.

[32] J. Li, Y. Ma, R. Song, B. Xi, D. Hong, and Q. Du, "A triplet semisupervised deep network for fusion classification of hyperspectral and LiDAR data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022, Art. no. 5540513.

[33] B. Xi, Y. Zhang, J. Li, Y. Li, Z. Li, and J. Chanussot, "CTF-SSCL: CNN-Transformer for few-shot hyperspectral image classification assisted by semisupervised contrastive learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–17, 2024, Art. no. 5532617.

[34] B. Xi, Y. Zhang, J. Li, Y. Huang, Y. Li, Z. Li, and J. Chanussot, "Transductive few-shot learning with enhanced spectral-spatial embedding for hyperspectral image classification," *IEEE Transactions on Image Processing*, vol. 34, pp. 854–868, 2025.

[35] D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, N. Yokoya, H. Li, P. Ghamisi, X. Jia, A. Plaza, P. Gamba, J. A. Benediktsson, and J. Chanussot, "SpectralGPT: Spectral remote sensing foundation model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5227–5244, 2024.

[36] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," 2020. [Online]. Available: https://arxiv.org/abs/1911.05722

[37] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020. [Online]. Available: https://arxiv.org/abs/2002.05709

[38] J. Wang, W. Song, H. Chen, J. Ren, and H. Zhao, "FusDreamer: Label-efficient remote sensing world model for multimodal data classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–14, 2025, Art. no. 5702314.

[39] W. Cai, X. Ning, G. Zhou, X. Bai, Y. Jiang, W. Li, and P. Qian, "A novel hyperspectral image classification model using bole convolution with three-direction attention mechanism: Small sample and unbalanced learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2023, Art. no. 5500917.

[40] Y.-S. Liu, Z.-L. Wang, Z. Yang, L.-X. Xu, Y. Chen, B. Ai, and Y. Su, "Data augmentation and feature fusion to improve graph contrastive learning for recommender systems," in *2024 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)*, 2024, pp. 1–6.

[41] J. Qiu, B. Chen, D. Song, and W. Wang, "Semi-supervised specific emitter identification based on contrastive learning and data augmentation," *IEEE Transactions on Aerospace and Electronic Systems*, pp. 1–18, 2025.

[42] S. Jia, X. Zhou, S. Jiang, and R. He, "Collaborative contrastive learning for hyperspectral and LiDAR classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023, Art. no. 5507714.

[43] S. Xu, X. Ding, Y. Zhang, Z. Zhang, H. Gao, and B. Zhang, "Dual-feature attention-based contrastive prototypical clustering for multimodal remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024, Art. no. 5539913.

[44] Y. Zhang, K. Liu, Y. Dong, K. Wu, and X. Hu, "Semisupervised classification based on SLIC segmentation for hyperspectral image," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 8, pp. 1440–1444, 2020.

[45] M. Van den Bergh, X. Boix, G. Roig, and L. V. Gool, "SEEDS: Superpixels extracted via energy-driven sampling," in *European Conference on Computer Vision*, vol. 111, 2015, pp. 298–314.

[46] J. Li, Y. Liu, R. Song, W. Liu, Y. Li, and Q. Du, "HyperMLP: Superpixel prior and feature aggregated perceptron networks for hyperspectral and LiDAR hybrid classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024, Art. no. 5505614.

[47] O. T. Nartey, K. Sarpong, D. Addo, Y. Rao, and Z. Qin, "PiCovS: Pixel-level with covariance pooling feature and superpixel-level feature fusion for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–20, 2023, Art. no. 4409320.

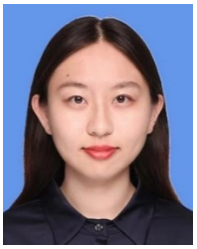
[48] Y. Sheng and L. Xiao, "Manifold augmentation based self-supervised contrastive learning for few-shot remote sensing scene classification," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, 2022, pp. 2239–2242.

[49] M. Wang, F. Gao, J. Dong, H.-C. Li, and Q. Du, "Nearest neighbor-based contrastive learning for hyperspectral and LiDAR data classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023, Art. no. 5501816.

[50] R. Guan, Z. Li, X. Li, and C. Tang, "Pixel-superpixel contrastive learning and pseudo-label correction for hyperspectral image clustering," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 6795–6799.

[51] H. Chen, J. Ru, H. Long, J. He, T. Chen, and W. Deng, "Semi-supervised adaptive pseudo-label feature learning for hyperspectral image classification in internet of things," *IEEE Internet of Things Journal*, vol. 11, no. 19, pp. 30 754–30 768, 2024.

[52] M. S. Kotzagiannidis and C.-B. Schönlieb, "Semi-supervised superpixel-based multi-feature graph learning for hyperspectral image data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022, Art. no. 4703612.



Yufei He received her bachelor's degree in Automation from Beijing University of Science and Technology in 2020. She is currently pursuing a Ph.D. degree in Computer Application Technology at the National Space Science Center, Chinese Academy of Sciences.

Her research interests focus on the joint processing of deep learning, hyperspectral images, and LiDAR data.

Bobo Xi (Member, IEEE) received the B.E. degree in information engineering and Ph.D. degree in information and communication engineering from Xidian University, Xi'an, China, in 2017 and 2022, respectively.

He is currently an Associate Professor with the State Key Laboratory of Integrated Services Networks, Xidian University. He has published over thirty papers in refereed journals, including the IEEE Transactions on Image Processing, the IEEE Transactions on Neural Networks and Learning Systems, and the IEEE Transactions on Geoscience and Remote Sensing. His research interests include hyperspectral image processing, machine learning, and deep learning.

Guocheng Li received his bachelor's degree in electronic information engineering from North China University of Technology in 2020. He is currently pursuing a Ph.D. in Computer Application Technology at the National Space Science Center, Chinese Academy of Sciences.

His research interests are focused on deep learning and hardware acceleration.

Tie Zheng received his bachelor's degree from Northeast Agricultural University and later completed his Ph.D. at the University of the Chinese Academy of Sciences in 2017 and 2023, respectively.

He is currently employed at the National Space Science Center, Chinese Academy of Sciences, with major research on space signal processing and remote sensing techniques.

Yunsong Li (Member, IEEE) received the M.S. degree in telecommunication and information systems and the Ph.D. degree in signal and information processing from Xidian University, Xi'an, China, in 1999 and 2002, respectively.

In 1999, he joined the School of Telecommunications Engineering, Xidian University, where he is currently a Professor. He is also the Director of the State Key Laboratory of Integrated Service Networks, Image Coding and Processing Center. His research interests include image and video processing, hyperspectral image (HSI) processing, and high-performance computing.

Changbin Xue (Member, IEEE) received the M.S. degree in Harbin Institute of Technology and the Ph.D. degree in Beijing Institute of Technology, Beijing, China, in 1997 and 2017, respectively.

He has been engaged in the design and development of complex space electronic information systems and space science exploration payload systems for more than 20 years. He served as the chief designer of the Chang'e-4 payload system and the chief designer of the space science pilot project "Shijian-10 Return Science Experiment Satellite". He has published more than 30 academic papers and more than 10 authorized patents. His research interests include full process design and simulation technology of deep space exploration scientific payload, machine learning, and deep learning.

Ming Shen (Senior Member, IEEE) received his M.Sc. and Ph.D. degrees from the University of Chinese Academy of Sciences, China, and Aalborg University, Denmark, in 2005 and 2010, respectively. He is currently an Associate Professor in the Department of Electronic Systems at Aalborg University. He is the head of the AI RF Sensor Group, and his research interests include applied machine learning in communication systems, satellite communication, and healthcare.