Cross-Modal Graph Knowledge Representation and Distillation Learning for Land Cover Classification

Wenzhen Wang, Fang Liu, *Member, IEEE*, Wenzhi Liao, *Senior Member, IEEE*, and Liang Xiao, *Member, IEEE*

***Abstract*— Complementary multimodal remote sensing (RS) data often lead to more robust and accurate classification performance. However, not all modal data can be available at the time of inference due to imaging conditions. To mitigate this issue, cross-modal knowledge distillation becomes an effective method, as it can leverage the complementary characteristics of multimodal data to guide cross-modal classification in cases with missing data. Therefore, this article examines the shortcomings of traditional convolutional neural network (CNN) cross-modal distillation methods in land cover classification: 1) insufficient knowledge representation and 2) unstable knowledge transfer. Moreover, a novel cross-modal graph knowledge representation and distillation learning (CGKR-DL) framework is proposed to enhance land cover classification performance. The proposed CGKR-DL designs a single-stream joint feature learning network with CNN and the graph convolutional network (GCN) to effectively construct the remote topology of data based on the strong correlation between land objects, thus enhancing the knowledge representation ability of the network. In addition, a multigranularity graph distillation method is proposed to compensate for the inability of traditional CNN distillation in handling graph-structured information, where a feature distillation module based on graph discrimination (FD-GDM) is designed for stable graph feature distillation. We evaluate CGKR-DL on three publicly available multimodal RS datasets [hyperspectral (HS)-light detection and ranging (LiDAR), HS-synthetic aperture radar (SAR), and HS-SAR-digital surface model (DSM)] and achieve a significant improvement in comparison with several state-of-the-art methods.**

***Index Terms*— Cross-modal classification, graph knowledge representation, multigranularity graph distillation, multimodal remote sensing (RS) data.**

I. Introduction

LAND cover classification identifies target attributes based on the unique spatial–spectral information of land cover in remote sensing (RS) images and assigns semantic class labels to each pixel in the image. It can then provide fundamental geographic data support for map drawing, environmental monitoring, disaster relief, military reconnaissance, and other fields [1]. After the successful development of various sensors, RS data can reflect land cover information from different modalities and perspectives. For example, hyperspectral (HS) images reflect spectral information, light detection and ranging (LiDAR) images describe elevation information, synthetic aperture radar (SAR) images provide surface structure information, and optical images capture color and texture information. Multimodal learning utilizes the complementary characteristics of different modal RS data to further improve land cover classification accuracy. However, during the acquisition of multimodal RS data, cloud and fog obscuration is inevitable, as illustrated by the partial cloud coverage in patch3's HS data in Fig. 1. The presence of clouds and fog scatters and absorbs most of the sunlight, leading to a decrease in the signal-to-noise ratio and impaired spatial and spectral resolutions of some modal RS data, which seriously affects their usability. Traditional multimodal learning methods inherently require all modal data to be available during both training and testing phases. Consequently, missing modal data present challenges for fine-grained land cover classification based on multimodal learning.

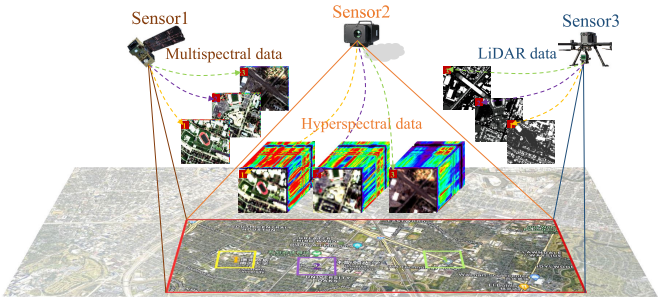

Fig. 1. Schematic of different sensors acquiring multimodal RS data in the same area. The red box indicates the acquired area, patches 1–3 indicate the part of the area used for visualization, and the three sensors are used to acquire RS data of different modalities separately, where the HS data acquired by the cloud-obscured part are not available.

The problem mentioned above is defined as a cross-modal land cover classification problem in [2]. It is highlighted that the learning paradigm of privileged information [3], which uses additional information to train predictive models, can obtain the best approximate representation during testing with missing data. Knowledge distillation, an effective

Manuscript received 13 July 2023; accepted 18 August 2023. Date of publication 22 August 2023; date of current version 31 August 2023. This work was supported in part by the Jiangsu Geological Bureau Research Project under Grant 2023KY11, in part by the Postgraduate Research & Practice Innovation Program of Jiangsu Province under Grant KYCX23_0489, and in part by the Open Research Fund in 2021 of the Jiangsu Key Laboratory of Spectral Imaging and Intelligent Sense under Grant JSGP202101 and Grant JSGP202204. *(Corresponding author: Liang Xiao.)*

Wenzhen Wang and Fang Liu are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: wangwz@njust.edu.cn; liufang_cs@njust.edu.cn).

Wenzhi Liao is with the Flemish Institute for Technological Research (VITO), 3920 Flanders, Belgium, and also with Department of Telecommunications and information processing, Faculty of Engineering and Architecture, Ghent University, 9000 Ghent, Belgium (e-mail: wenzhi.liao@flandersmake.be).

Liang Xiao is with the School of Computer Science and Engineering and the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: xiaoliang@mail.njust.edu.cn).

Digital Object Identifier 10.1109/TGRS.2023.3307604

strategy for learning privileged information, leverages available multimodal data to learn teacher knowledge to guide unimodal student learning. It allows for classification effects that approximate the collaborative action of multimodal data based solely on unimodal data. Among them, sufficient mining and stable transfer of knowledge are considered key challenges. Li et al. [4] design a unique feature extraction module for multimodal representation of optical and SAR images, and adopt a multilevel adaptive distillation approach to transfer teacher knowledge to students. Lv et al. [5] use a cross-modal fusion network to capture complementary fusion features of modalities and further transfer the fused knowledge through knowledge distillation to enhance the feature discrimination capability of specific networks. In addition, Liu et al. [6] employ deep neural networks to acquire modal knowledge and propose a cross-modal knowledge distillation framework to improve multispectral (MS) scene classification performance. Although the aforementioned methods achieve promising experimental results, they all rely on traditional convolutional neural network (CNN) cross-modal distillation methods, which, due to their inherent characteristics, still have two pressing issues that need to be addressed.

A. Insufficient Knowledge Representation

Traditional CNN cross-modal distillation methods utilize CNN and its variants for RS data feature representation. These methods primarily focus on local information mining and cannot effectively construct remote topological structures from a global perspective, which results in limited feature representation capability. Su et al. [7] attempt to improve network performance by constantly deepening the network, but this approach substantially increases computational complexity while delivering high performance. Given the strong correlations between land objects, graph convolutional networks (GCNs) are progressively used to mine rich topological structure information between objects, achieving information aggregation and transmission between land objects through graph convolution operations. Since GCN overly relies on noisy initial graphs, it is often combined with CNN for feature extraction in land cover classification. It typically takes the form of weighted feature fusion classification after dual-stream network feature extraction. Yan et al. [8] introduce GCN into the CNN distillation framework to explore long-term relationships between objects and global states, obtaining a more comprehensive knowledge representation. In the literature [9], [10], [11], CNN and GCN branches are separately employed for extracting land cover information and modeling the correlation between neighboring land covers, with their fusion features enhancing discriminative capabilities. However, we believe that the fusion of dual-stream features can easily lead to overmining of data, causing information redundancy, and the issue of GCN's noisy initial graphs is not effectively addressed.

B. Unstable Knowledge Transfer

Traditional CNN cross-modal distillation methods employ soft label distillation [12] or a combination of soft label and intermediate feature distillation [13] to transfer teacher knowledge to students. Among these, intermediate feature distillation often uses a predefined Euclidean distance metric as a common means for students to learn from teacher knowledge. In addition, when graph representations have been obtained, traditional CNN distillation methods only consider the feature similarity of data while ignoring the topological structure information of the data graph. Tung and Mori [14] guide student networks to learn feature representation relationships similar to teacher networks by matching the similarity matrices of student and teacher graphs; Yang et al. [15] design a local structure-preserving module to measure the topological structure relationship between teacher and student graphs, transferring topological knowledge through minimizing distances; and Zhou et al. [16] guide student network to learn from teacher network by minimizing the mutual information between teacher and student graphs. As a result, the knowledge transferred by graph distillation methods is not just simple data knowledge but graph knowledge containing both data information and topological structure information. However, due to the existence of modality differences between multimodal and unimodal data, there is a significant performance gap between teacher and student networks, and the aforementioned methods that use distance minimization metrics are prone to instability during the knowledge distillation process.

To address the aforementioned issues, we propose a novel framework called cross-modal graph knowledge representation and distillation learning (CGKR-DL), in which a CGKR-DL procedure is adopted to promote land cover classification performance. The proposed CGKR-DL designs a single-stream CNN-GCN joint feature learning network for graph knowledge representation, which utilizes the message-passing strategy of GCN on the basis of CNN to effectively construct the remote topology of data. Within this, a relationship learning module based on position embedding (PE-RLM) and a channel-gated aggregation module (CGAM) are specifically designed to enhance the feature representation capability of the network. These modules are intended to improve the graph construction method and information aggregation ability. Furthermore, a multigranularity graph distillation method is employed to transfer the graph knowledge of available multimodal data to unimodal data with missing data. Here, the feature distillation module based on graph discrimination (FD-GDM) is developed to guide the stable transfer of intermediate features. The main contributions of this article can be summarized as follows.

1) A single-stream CNN-GCN joint feature learning network is designed to tackle the issue of insufficient knowledge representation in traditional CNN cross-modal distillation methods. It utilizes the message-passing strategy of GCN on the basis of CNN to effectively construct the remote topology of data and enhance the feature representation capability of the network.
2) The designed PE-RLM and CGAM, respectively, alleviate the problem that the patches have different labels with similar features in the graph construction of traditional GCNs and improve the aggregation ability and

transmission mode of information in the graph encoding, further learning discriminative feature representations.

3) A multigranularity graph distillation method is introduced to address the problem of traditional CNN distillation that cannot handle graph structural information. Within this method, the FD-GDM is developed to alleviate the instability caused by the minimization of distance in the transfer of teacher knowledge to students in traditional feature distillation.

The rest of this article is organized as follows. Section II illustrates the related work. The proposed framework in detail is described in Section III. Section IV shows the experimental results and analysis. Finally, conclusion is given in Section V.

II. Related Work

A. Multimodal Land Cover Classification

Compared to single data source HS images, different modalities of RS data targeting the same area can reveal various attributes and features of land objects. By utilizing the differences and complementarities between them in specific circumstances, land cover classification accuracy can be further improved. On the one hand, due to the low spatial resolution of HS images, pixel confusion may occur. Spatial–spectral fusion methods that employ high-resolution panchromatic and MS images help compensate for the lack of spatial detail in HS images, thereby alleviating the mutual constraints between spatial and spectral resolutions to a certain extent. For instance, Gao et al. [17] design a transformer-based cross-scale hybrid fusion model that combines HS and MS images; Guan and Lam [18] develop a multilevel dual attention-guided fusion network for HS and panchromatic sharpening, effectively merging high-resolution panchromatic images with low-resolution HS images. On the other hand, different modalities of RS data typically contain more diverse information about land objects, so designing effective fusion strategies can enhance classification accuracy even with limited samples. For example, Gao et al. [19] adopt an adversarial complementary learning strategy to extract complementary information from multimodal data, improving the accuracy of multimodal RS image classification. Furthermore, Wang et al. [20] fully exploit a large amount of unlabeled data to learn discriminative feature representations so that efficient classification of multimodal RS images can be achieved by self-supervised learning.

In addition, HS/panchromatic/MS images are all imaged by capturing the reflected information from objects, which lack penetration capabilities. When imaging is affected by surrounding clouds and fog, the spectral features of land objects become blurred, leading to errors in determining object categories. In contrast, SAR and LiDAR images possess stronger cloud and fog penetration capabilities. Hu et al. [21] improve land cover classification performance influenced by imaging conditions by fusing SAR data and HS data obtained under thin clouds. Luo et al. [22] utilize a decision-level fusion of LiDAR data in cloudy regions and fused data in noncloudy areas to enhance classification performance. Although the aforementioned multimodal learning methods can mitigate the impact of cloud and fog obstructions, they all require that all modal data be available during training and testing. When severe cloud and fog obstructions cause local dark spots in the image and some modal data become unavailable, multimodal methods are no longer applicable.

B. Cross-Modal Land Cover Classification

Cross-modal land cover classification has emerged as a trend in RS intelligent analysis tasks due to its advantages of complementary input and increased output accuracy, often being utilized for handling issues related to missing modal data. Modality hallucination-based methods are regarded as one of the effective approaches, as they recover information from missing data using existing modal data, thus mitigating the negative impact caused by missing modal data. Current research employs mapping relationships between available modalities or generative models to recover missing data or learn its distribution, enabling the use of hallucination data to simulate multimodal collaborative classification during the testing phase. In [23], missing HS data are recovered by enhancing the MS image spectrum, while Hong et al. [24] utilize partially overlapping spectral information between HS and MS images for modality mapping relationship learning. Beyond specific modality constraints, Garcia et al. [25], [26] sequentially adopt regression and adversarial generation methods to create hallucination data for missing modalities; Pande et al. [27] propose an adversarial training-driven method to learn feature representations of missing modalities; and Wei et al. [28] design a modality-shared hallucination network to reconstruct comprehensive modality-shared features. However, such methods typically necessitate obtaining hallucination data before performing classification tasks, and the quality of hallucination data significantly impacts classification accuracy. Knowledge distillation aims to transfer the extensive knowledge acquired by a large, high-performance teacher network to a smaller student network, allowing it to achieve superior performance while reducing computational overhead. It can usually acquire the optimal approximate representation during training when testing with missing data and has become the mainstream method for cross-modal classification tasks [4], [5], [6]. Nonetheless, due to the presence of modality differences among data, traditional distillation methods may be insufficient for efficient knowledge learning and transfer, which consequently limits the enhancement of cross-modal classification performance.

C. GCN in RS Image Classification

GCNs aggregate and transfer neighboring object information via graph convolution operations, which are frequently employed to capture the rich topological structure information present in RS data. In this regard, constructing a meaningful graph based on RS data characteristics is an area worthy of investigation. Nodes, node features, and adjacency matrices comprise the three fundamental elements of graph construction. RS data graph construction methods can be categorized into three types according to their acquisition approach. One is to consider each sample as a node and directly use individual spectral information as node features, and its

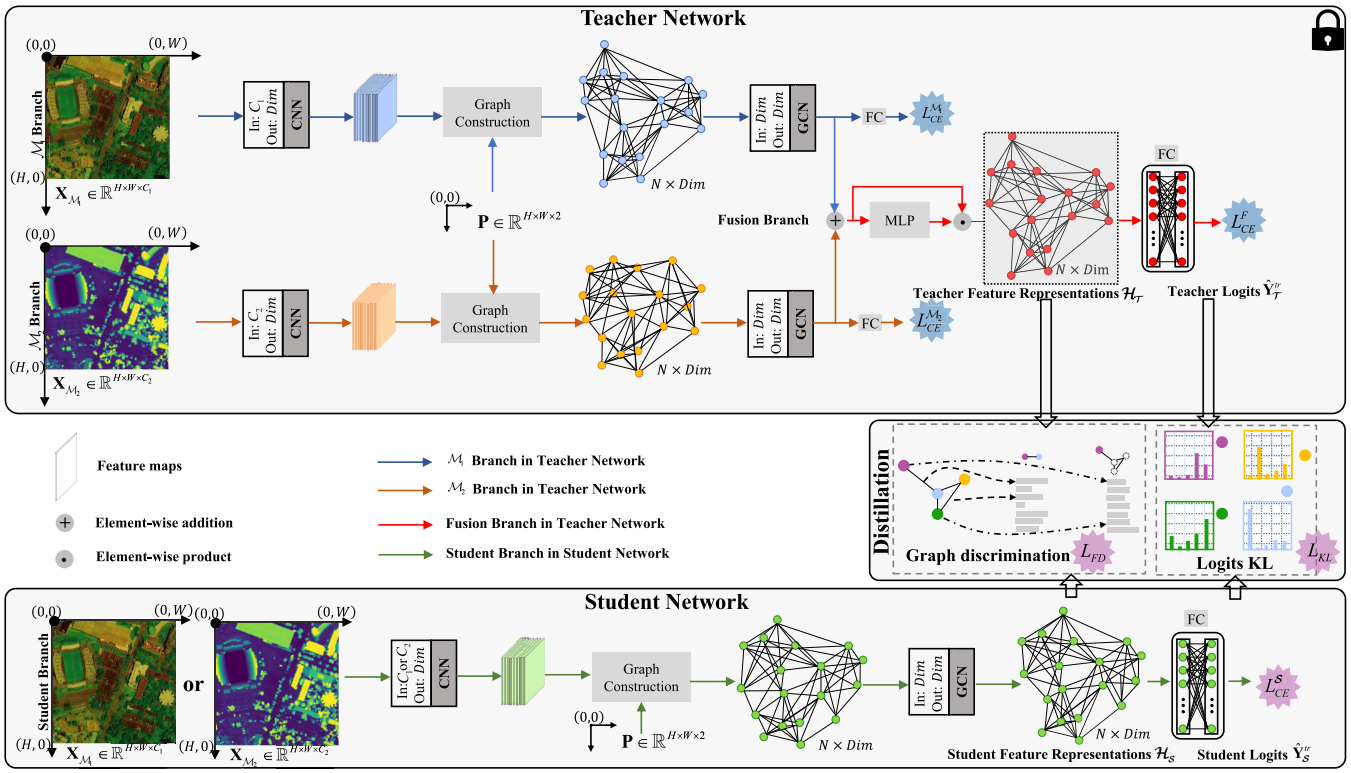

Fig. 2. Overview of the proposed CGKR-DL framework. It consists of a teacher network, a student network, and a distillation network. In particular, the teacher network designs its own land cover classification branch for each modality, and a fusion branch is used to integrate the information from multimodal RS data. Three branches are trained collaboratively, and the output of the fusion branch is used as transferable knowledge for guiding the student network. In addition, to highlight the guiding role of the teacher network, the student network structure directly uses the same modality branch as the teacher network. As for the distillation network, a multigranularity graph distillation method is adopted, which combines soft-label distillation and graph feature distillation.

adjacency matrix is calculated by the similarity relationships between nodes [29]. Although this method is appropriate for HS data with ample spatial–spectral information, it may result in wasted spatial data. The second is to identify nodes and their features via superpixel segmentation and determine adjacency matrices based on connections between superpixel blocks [30]. Nonetheless, the adjacency matrix in this way cannot be dynamically updated, making it heavily dependent on initial data correlation. The third method differs from the node feature representation in the first one but extracts the features from each sample patch using convolution or other techniques [31]. This method might inevitably face issues with different patches having similar labels. Each method has its own set of strengths and weaknesses, and is typically chosen to cater to specific requirements. Once the graph construction is completed, controlling the aggregation and transfer of neighboring node information is vital for obtaining distinctive features. From the initial method of summing neighboring node features and applying degree matrix weighting [32] and learning a neighboring aggregation function for node representation [33] to the currently prevalent method of learning the importance of neighboring nodes on themselves to regulate neighboring feature embedding [34], attention mechanisms have increasingly been implemented for adaptively selecting node embedding features.

III. Proposed Method

Given a set of multimodal data $\{\mathbf{X}_{\mathcal{M}_1} \in \mathbb{R}^{H\times W\times C_1}, \mathbf{X}_{\mathcal{M}_2} \in \mathbb{R}^{H\times W\times C_2}$, and $\mathbf{Y} \in \mathbb{R}^{H\times W}\}$, where $\mathcal{M}_1$ and $\mathcal{M}_2$ denote two aligned modalities, i.e., the data in the corresponding position in the two modalities describe the same area and share the same class label, H and W represent the height and width, respectively, and C_1 and C_2 refer to the number of channels in the respective modalities. The number of trainable labeled samples for the two modalities can be expressed as N_1, N_2 and $N_1 > N_2$ when we discard part of the labeled data of $\mathcal{M}_2$ to simulate the missing data problem. To build a cross-modal classification task with data missing, the paired labeled data of two modalities are used to train a collaborative classification network, and the remaining labeled data of $\mathcal{M}_1$ are used for testing. The training and testing data can be represented as $\{\mathbf{X}^{\text{tr}}_{\mathcal{M}_1} \in \mathbb{R}^{N_2\times C_1}, \mathbf{X}^{\text{tr}}_{\mathcal{M}_2} \in \mathbb{R}^{N_2\times C_2}, \mathbf{Y}^{\text{tr}} \in \mathbb{R}^{N_2}\}$ and $\{\mathbf{X}^{\text{te}}_{\mathcal{M}_1} \in \mathbb{R}^{(N_1-N_2)\times C_1}, \mathbf{Y}^{\text{te}} \in \mathbb{R}^{N_1-N_2}\}$, respectively. Our goal is to tackle the dual challenges of inadequate knowledge representation and unstable knowledge transfer found in traditional CNN distillation by devising a novel knowledge distillation technique, so as to solve the problem of cross-modal classification with missing data. To accomplish this, a CGKR-DL framework shown in Fig. 2 is proposed, which incorporates a single-stream CNN-GCN joint feature learning network specifically designed for efficient graph knowledge representation, as well as a multigranularity graph distillation method for effective graph knowledge transfer.

A. CNN-GCN Joint Feature Learning

The designed CNN-GCN joint feature learning network is shown in Fig. 3. The network takes patches of s-neighborhood

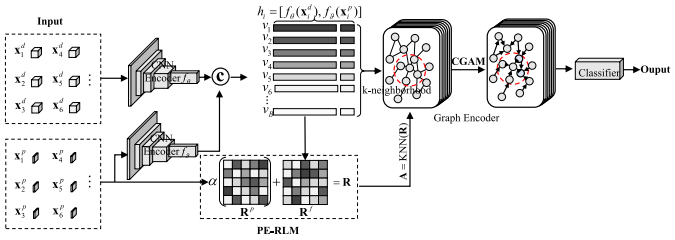

Fig. 3. Illustration of the proposed CNN-GCN joint feature learning network, in which the PE-RLM and the CGAM are designed to enhance the feature representation capability of the network.

size around the data as input, extracts features by CNN for graph construction and graph coding, and, finally, outputs the class prediction results. In particular, the position representation of the whole image $\boldsymbol{P} \in \mathbb{R}^{H\times W\times 2}$ is obtained by means of horizontal and vertical coordinate calibrations, first. It is worth noting that the position information is fixed after calibration. Then, the data and position of the modality are sampled simultaneously, and for a small patch, its data and position can be represented as $\mathbf{x}^d \in \mathbb{R}^{(2s+1)\times(2s+1)\times C_1}$ and $\mathbf{x}^p \in \mathbb{R}^{(2s+1)\times(2s+1)\times 2}$. After that, they are fed into CNN encoders f_θ, f_ϑ to encode data and position features for subsequent graph construction, respectively.

1) Graph Construction With PE-RLM: Reviewing the definition of GNNs, a $\mathcal{M}_1$ graph $\boldsymbol{\mathcal{G}} = (\boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{E}})$ can be defined by the node and adjacency matrix, where $\boldsymbol{\mathcal{V}} = \{v_1, v_2, \ldots, v_N\}$, $\boldsymbol{\mathcal{E}}$, and $|\boldsymbol{\mathcal{V}}| = N_2$ denote the node set, the edge set, and the total number of nodes in the graph, respectively. In this article, each patch is treated as a graph node, and the similarity distance between each node and its neighboring nodes is used to calculate the adjacency matrix, which represents the topological structure relationship between the nodes in the graph. It is generally determined by the feature similarity distance between its own node and its neighboring nodes, and for the feature similarity, the distance between nodes i and j can be expressed as

$$\mathbf{R}^f_{i,j} = \exp\left(-\frac{\left\| f_\theta\left(\mathbf{x}^d_i\right) - f_\theta\left(\mathbf{x}^d_j\right)\right\|^2}{2\tau^2}\right). \tag{1}$$

$\mathbf{R}^f \in \mathbb{R}^{|\boldsymbol{\mathcal{V}}|\times|\boldsymbol{\mathcal{V}}|}$ denotes the feature relationship matrix, and τ is a penalty parameter. In this article, considering that the label of a patch is determined by the central image element label, it is easy to have the problem that the patches have different labels with similar features. At this point, the closer the patches are, the more likely they have the same class labels, which helps to alleviate the problem. The position distance between patches is calculated using the coordinate distance of the central element position, and then, the position relationship matrix $\mathbf{R}^p \in \mathbb{R}^{|\boldsymbol{\mathcal{V}}|\times|\boldsymbol{\mathcal{V}}|}$ is given as follows:

$$\mathbf{R}^p_{i,j} = \exp\left(-\left\|\mathbf{x}^p_{i_s} - \mathbf{x}^p_{j_s}\right\|^2\right). \tag{2}$$

$\mathbf{x}^p_{i_s}$ represents the position coordinate of the ith patch at the central image element. Consequently, the proposed PE-RLM employs a unified approach that incorporates both feature representation and positional distance to improve the relationship discernment capabilities among similar nodes. Following this, the initial adjacency matrix between nodes is determined using the K-nearest neighbor (KNN) method

$$\mathbf{A} = \mathrm{KNN}(\mathbf{R}), \quad \text{where } \mathbf{R} = \mathbf{R}^f + \alpha\mathbf{R}^p \tag{3}$$

where $\mathbf{A}$ represents the adjacency matrix with weights, which is dynamically updated and makes the obtained graph structure continuously optimized as the feature learning capability of CNN is enhanced. α is the scale parameter used to adjust the assistant capacity of position information.

To somewhat reduce the oversmoothing problem in the process of graph convolution, the data feature learned from CNN cascade position embedding information is taken together as the initial feature representation of the node $\boldsymbol{\mathcal{H}} = \{h_1, h_2, \ldots, h_{|\boldsymbol{\mathcal{V}}|}\} \in \mathbb{R}^{|\boldsymbol{\mathcal{V}}|\times\dim}$, where $h_i = [f_\theta(\mathbf{x}^d_i), f_\vartheta(\mathbf{x}^p_i)]$, dim represents the node feature dimension, and $[\cdot,\cdot]$ represents the cascade operation. At this point, the graph construction process is successfully completed.

2) Graph Convolution With CGAM: A graph convolution operation is performed on the constructed graph to capture the rich topological structure information by aggregating the information of neighboring nodes, hereby enhancing the learned feature representation capabilities

$$h^{(l+1)}_i = \mathrm{UPDAT}E^{(l)}\left(h^{(l)}_i, \mathrm{AGG}^{(l)}_{j\in\mathcal{N}(i)}\left(h^{(l)}_j\right)\right) \tag{4}$$

where $h^{(l+1)}_i = \mathrm{BN}(\mathrm{GC}(h^{(l)}_i))$ denotes the node embedding features of node v_i at the $l+1$th layer via the graph convolution layer GC and the normalization layer BN; in particular, when $l = 0$, $h^{(0)}_i = [f_\theta(\mathbf{x}^d_i), f_\vartheta(\mathbf{x}^p_i)]$. AGG and UPDATE are the node feature aggregation function and the message-passing update function, respectively, and $\mathcal{N}(i)$ denotes the neighborhood node set of node v_i.

Nonetheless, effectively managing the aggregation and transmission of neighboring node information is essential for acquiring distinct features. It has been demonstrated that the way of aggregating all the neighboring node information without selection makes the aggregation operation unnecessary. With the development of graph convolution, the attention weight is mostly used to selectively feature aggregate the information of neighboring nodes in order to emphasize the contribution of different neighboring nodes to their own nodes. For multichannel HS data with rich spectral information and strong discriminative power, it is no longer appropriate to assign the same attention weight to the channel features of nodes. A multilayer perceptron (MLP) is used to learn the channel feature correlation representation of neighboring node v_j and its own node v_i, and to maximize the receptive field by filtering out anomalous information and absorbing similar features

$$e^l_{i,j} = \mathbf{MLP}\left(A^l h^l_i + B^l h^l_j\right) \tag{5}$$

where $e_{ij} \in \mathbb{R}^{1\times\dim}$ represents the channel feature correlation of nodes v_i and v_j; **MLP** consists of a fully connected layer, a BN layer, and a sigmoid layer; and A^l and B^l are learnable parameters. Then, the information transfer from the channel features of neighboring nodes to their own nodes is controlled by gating to enhance the class feature representation, and

finally, the information of their own nodes is added to prevent information loss. Therefore, the graph convolution process of the proposed method can be expressed as follows:

$$h_i^{l+1} = \sigma\left(A^l h_i^l + \sum_{j\in\mathcal{N}(i)} e_{i,j}^l \odot B^l h_j^l\right). \tag{6}$$

σ denotes the activation function, and $\odot$ means elementwise product. With this, the process for a single layer of graph convolution is successfully completed.

3) Graph Knowledge Representation Learning: The graph node features $\widetilde{\mathcal{H}}$ obtained from the graph convolution are then fed into a fully connected layer for class prediction, and the cross-entropy classification loss can be calculated

$$\hat{\mathbf{Y}}^{\text{tr}} = \text{FC}(\widetilde{\mathcal{H}}), \quad L_{\text{CE}} = \text{CE}(\hat{\mathbf{Y}}^{\text{tr}}, \mathbf{Y}^{\text{tr}}) \tag{7}$$

where $\hat{\mathbf{Y}}^{\text{tr}}$ and L_{CE} represent the predicted labels and training losses, respectively, FC is a two-layer fully connected network, and CE denotes the cross-entropy function. In our framework, multimodal data $\{\mathbf{X}_{\mathcal{M}_1}^{\text{tr}}, \mathbf{X}_{\mathcal{M}_2}^{\text{tr}}, \mathbf{Y}^{\text{tr}}\}$ possessing complementary characteristics serve as a teacher, with each modality independently employing the CNN-GCN joint feature learning network for graph knowledge representation learning. Initially, the graph convolution features of $\mathcal{M}_1$ and $\mathcal{M}_2$ branches are represented as $\mathcal{H}_{\mathcal{T}}^{\mathcal{M}_1}$ and $\mathcal{H}_{\mathcal{T}}^{\mathcal{M}_2}$, respectively. These features are then combined into teacher knowledge via a fusion branch to guide the student network. Notably, given the complementarity and redundancy among different modalities, **MLP** is used to adaptively learn the attention weights for each node. Subsequently, the fused teacher features $\mathcal{H}_{\mathcal{T}} \in \mathbb{R}^{|\mathcal{V}|\times\text{dim}}$ are obtained by weighting and integrating the contributions from individual nodes

$$\mathcal{H}_{\mathcal{T}} = \mathbf{MLP}\left(\mathcal{H}_{\mathcal{T}}^F\right) \odot \mathcal{H}_{\mathcal{T}}^F \tag{8}$$

where $\mathcal{H}_{\mathcal{T}}^F = \mathcal{H}_{\mathcal{T}}^{\mathcal{M}_1} + \mathcal{H}_{\mathcal{T}}^{\mathcal{M}_2}$.

Following (7), the predicted labels $\hat{\mathbf{Y}}_{\mathcal{M}_1}^{\text{tr}}$,$\hat{\mathbf{Y}}_{\mathcal{M}_2}^{\text{tr}}$, and $\hat{\mathbf{Y}}_{\mathcal{T}}^{\text{tr}}$ for $\mathcal{M}_1$ and $\mathcal{M}_2$, and fusion branches can be calculated individually, as well as the respective training losses $L_{\text{CE}}^{\mathcal{M}_1}$, $L_{\text{CE}}^{\mathcal{M}_2}$, and L_{CE}^F. Subsequently, an adaptive weighting approach is employed to balance the training process across the three branches, ensuring efficient network training. The overall loss of the teacher network can be expressed as

$$L^{\mathcal{T}}(\Theta_{\mathcal{T}}) = \sum_{\zeta=\mathcal{M}_1,\mathcal{M}_2,F} \lambda^{\zeta} L_{\text{CE}}^{\zeta}. \tag{9}$$

$\Theta_{\mathcal{T}}$ is the trainable parameter of the teacher network $\mathcal{T}$, $\lambda^{\zeta} = L_{\text{CE}}^{\zeta}/\sum_{\zeta} L_{\text{CE}}^{\zeta}$, $\zeta = \mathcal{M}_1, \mathcal{M}_2$, and F denotes the adaptive weight parameter. Finally, the teacher knowledge with complementary characteristics is extracted from the pretrained teacher network, denoted as $\mathcal{H}_{\mathcal{T}}$ and $\hat{\mathbf{Y}}_{\mathcal{T}}^{\text{tr}}$.

B. Multigranularity Graph Distillation

In our proposed framework, efficiently and stably transferring complementary teacher knowledge to the student network under data missing is a vital factor to consider in the design of distillation networks. Soft label distillation and intermediate feature distillation are continuously researched and developed as common approaches. Soft label distillation supplies probability distribution information on discrete categories during the initial phase of student network training and has proven its significant guiding role. Concurrently, the intermediate feature distillation approach enables the student network to acquire feature representations akin to those of the teacher network, fostering a stable knowledge-learning process. As such, we continue to adopt a multigranularity graph distillation method that merges soft label distillation with intermediate graph feature distillation, guiding the student network in learning from the teacher network. The distinction lies in proposing a tailored multigranularity graph knowledge distillation network for the characteristics of the graphs presented in this article. In this regard, FD-GDM is designed to mitigate the instability in transferring teacher knowledge to students caused by distance minimization. The loss function of the multigranularity graph distillation network (MGDN) can be articulated as

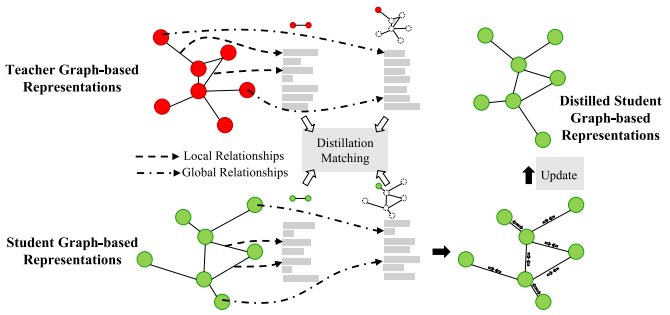

Fig. 4. Proposed FD-GDM. The local relationship representation between nodes in the graph and the global relationship representation between nodes and their own graphs are used to distill and match the student network and the teacher network, respectively, so as to guide the student network to learn the knowledge extracted from the teacher network.

$$L_{\text{KD}} = L_{\text{FD}} + L_{\text{KL}} \tag{10}$$

where L_{FD} and L_{KL} represent the feature distillation loss and the soft-label loss, respectively.

1) Feature Distillation Based on Graph Discrimination: Inspired by generative adversarial learning, the discriminator is designed exclusively to discriminate the feature representation of the teacher network and the student network, using the student network as the generator. Notably, the discriminator in our design will give more inclusiveness to the distillation process by no longer requiring the absolute distance between teacher and student node pairs to be equal. As shown in Fig. 4, the proposed method guides the student network to learn from the teacher network in terms of both data-based node representation and structure-based graph representation, respectively.

Formally, corresponding to the teacher knowledge $\mathcal{H}_{\mathcal{T}} \in \mathbb{R}^{|\mathcal{V}|\times\text{dim}}$ provided by the teacher network, the student network takes unimodal data as input and outputs the node representation $\mathcal{H}_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{V}|\times\text{dim}}$ of the student graph using joint CNN-GCN feature learning. If the average of all node representation in a graph is used as the graph representation $\mathcal{Z}$ of that graph, the graph representation of the teacher and student can be denoted as $\mathcal{Z}_{\mathcal{T}} = (1/|\mathcal{V}|)\sum_{i\in\mathcal{V}} h_i^{\mathcal{T}}$, $h_i^{\mathcal{T}} \in \mathcal{H}_{\mathcal{T}}$ and

$\mathcal{Z}_{\mathcal{S}} = (1/|\mathcal{V}|)\sum_{i\in\mathcal{V}} h_i^{\mathcal{S}}$, and $h_i^{\mathcal{S}} \in \mathcal{H}_{\mathcal{S}}$, respectively. Since the parameters of the pretrained teacher network have been frozen, the feature representation relationship between any node pair in the teacher graph (local relationships $\{h_i^{\mathcal{T}}, h_j^{\mathcal{T}}\}$ and between any node, and the whole graph (global relationships $\{h_i^{\mathcal{T}}, \mathcal{Z}_{\mathcal{T}}\}$) are fixed. i and j are any two nodes from the same graph. Based on this idea, the relationship is characterized in the form of a vector inner product. A node discriminator $\mathcal{D}_{\varphi}^{n}$ and a graph discriminator $\mathcal{D}_{\psi}^{g}$ are designed to distill and match the local and global relationships in the teacher and student graphs, respectively,

$$\mathcal{D}_{\varphi}^{n}(h_i, h_j) = \langle h_i, \mathbf{\Lambda}_{\varphi} h_j \rangle \tag{11}$$

$$\mathcal{D}_{\psi}^{g}(h_i, \mathcal{Z}) = \langle h_i, \mathbf{\Lambda}_{\psi} \mathcal{Z} \rangle \tag{12}$$

where $\langle \cdot, \cdot \rangle$ represent the inner product of vectors, $\mathbf{\Lambda}_{\varphi}$ and $\mathbf{\Lambda}_{\psi}$ denote the learnable diagonal matrices, and φ and ψ are the network parameters of the discriminators. The matching result is expressed in the form of a binary value, as shown in (13). In local relationship discrimination, the node discriminator discriminates the node representation from the teacher graph as real and conversely from the student graph as fake

$$L_N(\Theta_{\mathcal{T}}, \Theta_{\mathcal{S}}, \varphi) = \frac{1}{|\mathcal{V}|} \sum_{(i,j)\in\mathcal{V}} \left(\log P\left(\text{Real} \middle| \mathcal{D}_{\varphi}^{n}\left(h_i^{\mathcal{T}}, h_j^{\mathcal{T}}\right)\right) + \log P\left(\text{Fake} \middle| \mathcal{D}_{\varphi}^{n}\left(h_i^{\mathcal{S}}, h_j^{\mathcal{S}}\right)\right)\right) \tag{13}$$

where $\Theta_{\mathcal{S}}$ is the trainable parameters of the student network $\mathcal{S}$. On the other hand, for global relationship discrimination, the graph discriminator discriminates node representation from the same graph with graph representation as real and vice versa as fake

$$\begin{aligned} L_G(\Theta_{\mathcal{T}}, \Theta_{\mathcal{S}}, \psi) = \frac{1}{2|\mathcal{V}|} \sum_{i\in\mathcal{V}} \Bigg(&\log P\left(\text{Real} \middle| \mathcal{D}_{\psi}^{g}\left(\{h_i^{\mathcal{T}}, \mathcal{Z}_{\mathcal{T}}\}\right)\right) \\ &+ \log P\left(\text{Fake} \middle| \mathcal{D}_{\psi}^{g}\left(\{h_i^{\mathcal{S}}, \mathcal{Z}_{\mathcal{T}}\}\right)\right) \\ &+ \log P\left(\text{Real} \middle| \mathcal{D}_{\psi}^{g}\left(\{h_i^{\mathcal{S}}, \mathcal{Z}_{\mathcal{S}}\}\right)\right) \\ &+ \log P\left(\text{Fake} \middle| \mathcal{D}_{\psi}^{g}\left(\{h_i^{\mathcal{T}}, \mathcal{Z}_{\mathcal{S}}\}\right)\right)\Bigg). \end{aligned} \tag{14}$$

On the other hand, the student network, as a generator, is encouraged to generate node representation and graph representation that approximate the teacher network and thus fool the discriminator so that the adversarial training process for a two-player minimax game can be expressed as

$$\min_{\Theta_{\mathcal{S}}} \max_{\varphi,\psi} L_{\text{FD}}(\Theta_{\mathcal{T}}, \Theta_{\mathcal{S}}, \varphi, \psi) \tag{15}$$

where $L_{\text{FD}} = L_N + L_G$.

2) Soft-Label Distillation: The Kullback–Leibler (KL) distance is a measure of the relative difference between two probability distributions in the same event space and is used to calculate the soft-label distillation loss. Given the predictions of the teacher network obtained from pretraining $\hat{\mathbf{Y}}_{\mathcal{T}}^{\text{tr}}$ and the predictions generated by the student network $\hat{\mathbf{Y}}_{\mathcal{S}}^{\text{tr}}$, the soft-label distillation loss can be calculated as follows:

$$L_{\text{KL}}(\Theta_{\mathcal{T}}, \Theta_{\mathcal{S}}) = t^2 \mathbf{KL}\left(\hat{\mathbf{Y}}_{\mathcal{T}}^{\text{tr}}/t, \hat{\mathbf{Y}}_{\mathcal{S}}^{\text{tr}}/t\right) \tag{16}$$

Algorithm 1 Training of the Proposed CGKR-DL

Input: $\{\mathbf{X}_{\mathcal{M}_1}^{tr}, \mathbf{X}_{\mathcal{M}_2}^{tr}, \mathbf{Y}^{tr}\}$, *Epoches*, $\mathcal{T}$ with $\Theta_{\mathcal{T}}$ and $\mathcal{S}$ with $\Theta_{\mathcal{S}}$.
Output: The well-trained $\mathcal{S}$ with $\Theta_{\mathcal{S}}{}^{*}$.

```
Initialize Θ_T, Θ_S;
Token the position information and sample randomly;
for epoch in Epoches do
    // Teacher Graph Knowledge Representation;
    Perform the CNN-GCN joint feature representation
      on X^tr_M1, X^tr_M2, respectively;
    Calculate the fused teacher features by Eq.(8);
    Predict labels and calculate teacher loss by Eq.(9);
    Update Θ_T with the gradient descent by Eq.(17);
end
for epoch in Epoches do
    // Multi-Granularity Graph Distillation;
    Freeze teacher parameters Θ_T*;
    Extract teacher knowledge H_T, Ŷ^tr_T =
      T_{Θ_T*}(X^tr_M1, X^tr_M2); Z_T = (1/|V|) Σ_{i∈V} h_i^T;
    H_S, Ŷ^tr_S = S(X^tr_M1); Z_S = (1/|V|) Σ_{i∈V} h_i^S;
    Calculate the distillation loss L_KD by Eq.(10);
    Calculate the student classification loss L_CE^S;
    Update Θ_S with the gradient descent by Eq.(18);
end
```

where t is a hyperparameter indicating the temperature, which is used to soften the label; the higher the value, the more obvious the softening effect.

C. Cross-Modal Training and Testing

1) Training: The proposed framework is dedicated to solving the problem of cross-modal RS image classification in the presence of data missing while enhancing the feature representation ability of the network. Following the general paradigm of knowledge distillation, a teacher network is first trained in the training phase of the network using the paired multimodal data $\{\mathbf{X}_{\mathcal{M}_1}^{\text{tr}}, \mathbf{X}_{\mathcal{M}_2}^{\text{tr}}, \mathbf{Y}^{\text{tr}}\}$ with complementary characteristics, and its network parameters are obtained and frozen by minimizing the following equation:

$$\Theta_{\mathcal{T}}{}^{*} = \arg\min_{\Theta_{\mathcal{T}}} L^{\mathcal{T}}(\Theta_{\mathcal{T}}). \tag{17}$$

Next, multimodal data $\{\mathbf{X}_{\mathcal{M}_1}^{\text{tr}}, \mathbf{X}_{\mathcal{M}_2}^{\text{tr}}\}$ are used as input to the teacher network with frozen network parameters $\Theta_{\mathcal{T}}{}^{*}$, and unimodal data $\{\mathbf{X}_{\mathcal{M}_1}^{\text{tr}}, \mathbf{Y}^{\text{tr}}\}$ are used as input to the student network. The student network jointly trains its own model based on the ground truth and the knowledge extracted by the teacher network

$$\Theta_{\mathcal{S}}{}^{*} = \arg\min_{\Theta_{\mathcal{S}}} \max_{\varphi,\psi} \left(\eta L_{\text{CE}}^{\mathcal{S}}(\Theta_{\mathcal{S}}) + L_{\text{KD}}\left(\Theta_{\mathcal{T}}^{*}, \Theta_{\mathcal{S}}, \varphi, \psi\right)\right) \tag{18}$$

where $\Theta_{\mathcal{S}}{}^{*}$ is the network parameters obtained from the final learned of the student network $\mathcal{S}$, $L_{\text{CE}}^{\mathcal{S}} = \text{CE}(\hat{\mathbf{Y}}_{\mathcal{S}}^{\text{tr}}, \mathbf{Y}^{\text{tr}})$ is the cross-entropy loss of the student network, and η is the weight hyperparameter of the loss. The training procedure of the proposed framework is available in Algorithm 1.

2) Testing: The proposed framework performs cross-modal classification on unimodal testing data $\mathbf{X}^{\text{te}}_{\mathcal{M}_1}$ using only the trained student network $\boldsymbol{\mathcal{S}}$ with $\Theta_{\boldsymbol{\mathcal{S}}^*}$ during the testing phase

$$\hat{\mathbf{Y}}^{\text{te}} = \boldsymbol{\mathcal{S}}_{\Theta_{\boldsymbol{\mathcal{S}}^*}}\left(\mathbf{X}^{\text{te}}_{\mathcal{M}_1}\right) \tag{19}$$

in which $\hat{\mathbf{Y}}^{\text{te}}$ denotes the predicted label of the testing data. Thanks to the guidance of the teacher, the classification performance of the student network is improved.

IV. Experiments

A. Experimental Dataset

In this article, three different publicly available heterogeneous RS datasets[1] are selected for training and testing for the cross-modal land cover classification task, including the HS-LiDAR MUUFL Gulfport dataset (MUUFL), the HS-SAR Berlin dataset (Berlin), and the HS-SAR-digital surface model (DSM) Augsburg dataset (Augsburg). LiDAR data or DSM data, which provide ground elevation information, and SAR data, which provide information on the Earth's surface structure, are combined around HS data, respectively, to further improve the fine identification of land cover. In order to simulate the data missing problem, only the unimodal data are retained for testing in the testing phase. Furthermore, existing methods are mostly used to deal with the classification problem of two modalities. In this article, three datasets are formed around HS data into four multimodal RS data for experiments.

1) HS-LiDAR MUUFL Gulfport Dataset: The MUUFL Gulfport dataset describes scenes from the University of Southern Mississippi Gulf Park campus, Long Beach, MS, USA. It consists of two data sources, including an HS image and a LiDAR image. Among them, the HS data is collected by the Compact Airborne Spectrographic Imager (CASI-1500) sensor, which is a scene with 325 × 220 pixels, 64 spectral bands, a wavelength range of 367.7∼1043.4 nm, and a spectral resolution of 9.5∼9.6 nm. LiDAR data are acquired by the Optech Airborne Laser Terrain Mapper (ALTM) sensor relying on a laser with a wavelength of 1064 nm and aligned. The dataset covers 11 classes, and the information about the training and testing data is shown in Fig. 5 and Table I.

2) HS-SAR Berlin Dataset: The Berlin dataset describes scenes from the city of Berlin and its countryside surroundings. It consists of two data sources, including the HS image and the SAR image. Among them, the simulated Environmental Mapping and Analysis Programme (EnMAP) data synthesized based on the HS mapper (HyMap) HS data graphically describes the Berlin urban and its rural neighboring area at 30 m GSD, which is a scene with 797 × 220 pixels, 244 spectral bands, and a wavelength range of 0.4∼2.5 μm. The SAR data of the corresponding area are preprocessed with the Sentinel-1 double-pol single-look complex (SLC) product downloaded from ESA obtained with a size of 1723 × 476 pixels. Finally, nearest neighbor interpolation is applied to HS to obtain data matching the SAR size for experiments. The dataset covers eight classes, and the information about the training and testing data is shown in Fig. 5 and Table II.

[1]https://drive.google.com/file/d/1dLJJrNJpQoQeDHybs37iGxmrSU6aP2xv/view?usp=sharing

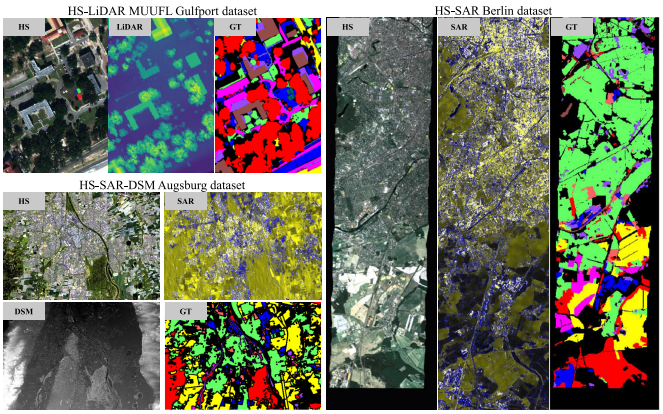

Fig. 5. False-color visualization of three multimodal RS benchmark datasets: the HS-LiDAR MUUFL Gulfport dataset, the HS-SAR Berlin dataset, and the HS-SAR-DSM Augsburg dataset.

TABLE I
LAND COVER CLASSES AND NUMBERS OF SAMPLES IN THE MUUFL DATASET

Class No.	Class Name	Training	Testing
1	Trees	150	23096
2	Mostly Grass	150	4120
3	Mixed Ground Surface	150	6732
4	Dirt and Sand	150	1676
5	Road	150	6537
6	Water	150	316
7	Building Shadow	150	2083
8	Buildings	150	6090
9	Sidewalk	150	1235
10	Yellow Curb	100	83
11	Cloth Panels	100	169
	Total	1550	52137

TABLE II
LAND COVER CLASSES AND NUMBERS OF SAMPLES IN THE BERLIN DATASET

Class No.	Class Name	Training	Testing
1	Forest	443	54511
2	Residential Area	423	268219
3	Industrial Area	499	19067
4	Low Plants	376	58906
5	Soil	331	17095
6	Allotment	280	13025
7	Commercial Area	298	24526
8	Water	170	6502
	Total	2820	461851

3) HS-SAR-DSM Augsburg Dataset: The Augsburg dataset describes scenes from the city of Augsburg, Germany. It consists of three data sources, including the HS image, the SAR image, and the DSM image, which are acquired by the HySpex sensor, the Sentinel-1 sensor, and the DLR-3 K system, respectively. The scene pixels are 332 × 485 pixels, and there are 180 spectral bands in the wavelength range of 0.4∼2.5 μm for HS data, four spectral bands for SAR data, and one spectral band for DSM data. The dataset covers seven classes, and the information about the training and testing data is shown in Fig. 5 and Table III.

B. Experimental Setting

1) Implementation Details: To reduce the high computational load caused by large graph computation, a minibatch

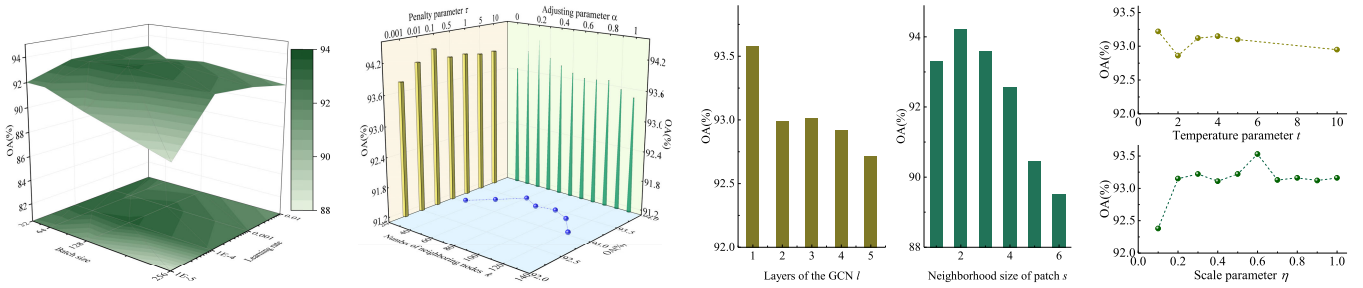

Fig. 6. Classification results of several sets of hyperparameters on the MUUFL dataset with different value settings. (a) Learning rate and batch size. (b) Penalty parameter τ, adjusting parameter α, and number of neighboring nodes k. (c) Layers of the GCN l and neighborhood size of patch s. (d) Temperature parameter t and scale parameter η.

TABLE III
LAND COVER CLASSES AND NUMBERS OF SAMPLES IN THE AUGSBURG DATASET

Class No.	Class Name	Training	Testing
1	Forest	146	13361
2	Residential Area	264	30065
3	Industrial Area	21	3830
4	Low Plants	248	26609
5	Allotment	52	523
6	Commercial Area	7	1638
7	Water	23	1507
	Total	761	77533

procedure is used to obtain a set of subgraphs for computation $\mathbb{G} = \{\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i) | i = 1, \ldots, [|\mathcal{V}|/B]\}$. Here, $[|\mathcal{V}|/B]$ denotes a downward rounding operation in steps of batch size B. Moreover, the residual learning and attention mechanisms in the HResNetAM [35] model are used in the CNN encoder in this article, in which the convolution kernel of each encoder block is set to decide whether the encoder block extracts spatial features or spectral features, and the feature output dimension is 48. Specifically, the space-spectral features extracted from multichannel data are fused and output using an attention mechanism. To verify the effectiveness of the proposed method, this article evaluates the performance of the multimodal classification network and the cross-modal distillation network, respectively. Specifically, the performance of the proposed multimodal classification method is further evaluated from three different perspectives. MGFN [29] and ACGC [16] methods are selected to compare different graph adjacency matrix construction ways; GCN [32], GAT [33], and Gated-GCN [34] methods are selected to compare different graph convolution aggregation manners. It is worth noting that, in order to fairly compare and validate the performance of each of the proposed modules, the previously mentioned comparison methods only replace the current modules used for comparison in the multimodal teacher network proposed in this article. Then, five state-of-the-art multimodal RS collaborative classification methods (CoupledCNNs [36], AsyFFNet [37], DFINet [38], CALC [39], and MFT [40]) are chosen to validate the proposed multimodal classification method in terms of overall classification performance. For cross-modal distillation performance, various comparisons are made from the vanilla soft-label distillation method (KD [12]) and intermediate feature distillation method (Hint [13]), to generative distillation methods (MSGD [25] and ADMD [26]), and to the latest graph-based distillation methods (LSP [15] and HKD [16]), respectively.

Ten random divisions of the training and testing sets are performed, and their average performance is taken as the final performance of the model. Here, overall accuracy (OA), average accuracy (AA), and kappa coefficients (κ) are used to evaluate the performance of all models.

2) Hyperparameter Analysis: This article conducts hyperparametric analysis experiments on the proposed model using the available datasets and shows the hyperparametric analysis process on the MUUFL dataset in this subsection. The learning rate and the batch size determine the weight update of the model and are the most important parameters affecting the convergence of the model. In this article, the grid search method is used to select the learning rate $= \{0.01, 0.001, 0.0001, 0.00001\}$ and the batch size $= \{32, 64, 128, 256\}$. Fig. 6(a) shows the results of OA for multimodal classification accuracy with different value settings on the MUUFL dataset, where learning rate $= 0.0001$ and batch size $= 128$ are selected. During the construction of the adjacency matrix, there are several important hyperparameters worth analyzing. The first one is the penalty parameter τ in (1). When τ is relatively small (large), the magnitude of the feature similarity distance function diminishes (increases), leading to reduced (expanded) distances between data points. By adjusting this parameter, the model can accurately capture the differences among the data. The second hyperparameter is α in (3), which is used to adjust the auxiliary capabilities of positional information. The number of nearest neighbors is also one of the important factors affecting the quality of aggregated and updated nodes. We randomly select several sets of values based on the determined batch size to test their classification performance. According to the experimental results shown in Fig. 6(b), τ is set to 0.1, α is set to 0.2, and k is set to 60. In addition, Fig. 6(c) demonstrates other hyperparameters that impact classification accuracy, such as setting the layers of the GCN l to 1 and setting the neighborhood size of patch s to 2.

To ensure the best performance of the teacher network, the distillation network in this article uses the same superparameter settings as described previously. In addition, the temperature parameter t and the scale parameter η in (18) are analyzed separately by taking random values. The results

TABLE IV
CLASS-SPECIFIC CLASSIFICATION ACCURACY (%) ON THE MUUFL DATASET BY CONSIDERING HS AND LIDAR DATA

Methods \ Class No.	1	2	3	4	5	6	7	8	9	10	11	OA(%)	AA(%)	κ
MGFN[29]	96.82 ±0.42	68.79 ±1.91	79.19 ±2.39	82.05 ±4.75	95.21 ±1.49	70.53 ±7.89	63.08 ±5.63	94.23 ±1.84	55.99 ±3.86	31.28 ±8.97	83.01 ±10.26	87.39 ±0.55	74.56 ±1.43	0.8338 ±0.0070
ACGC[16]	98.34 ±0.37	83.33 ±2.79	88.66 ±1.29	89.60 ±3.51	95.94 ±0.60	85.62 ±3.45	76.56 ±5.72	97.62 ±0.63	67.39 ±5.51	20.25 ±3.83	67.58 ±6.32	92.40 ±0.32	79.17 ±0.91	0.8995 ±0.0042
GCN[32]	80.55 ±3.44	49.77 ±7.00	62.26 ±6.44	47.01 ±10.99	80.34 ±5.47	34.15 ±22.71	52.80 ±10.37	81.61 ±9.21	19.83 ±5.65	1.74 ±1.24	45.17 ±26.44	71.85 ±2.28	50.48 ±4.90	0.6105 ±0.0322
GAT[33]	86.02 ±3.74	60.50 ±9.08	60.11 ±6.74	70.99 ±16.02	77.88 ±6.45	57.87 ±24.26	63.19 ±13.63	91.05 ±4.27	27.10 ±17.90	4.54 ±5.82	45.76 ±32.18	76.78 ±0.89	58.64 ±3.00	0.6810 ±0.0150
Gated-GCN[34]	96.64 ±0.93	85.89 ±1.53	89.79 ±1.99	91.44 ±2.11	96.61 ±1.64	88.38 ±8.82	84.69 ±4.43	95.21 ±2.33	**70.77** **±4.81**	21.53 ±8.80	66.64 ±10.94	93.05 ±0.42	80.69 ±1.69	0.9071 ±0.0055
CoupledCNNs[36]	97.81 ±0.99	79.56 ±3.56	88.37 ±2.39	84.87 ±4.52	96.02 ±0.98	74.95 ±11.63	79.59 ±6.27	97.71 ±0.68	69.73 ±9.15	13.00 ±2.71	60.34 ±7.12	91.33 ±0.74	76.54 ±2.03	0.8853 ±0.0092
AsyFFNet[37]	98.41 ±0.37	84.49 ±3.84	88.62 ±2.26	85.88 ±3.12	94.94 ±1.03	71.20 ±10.54	81.26 ±4.48	96.82 ±0.78	51.02 ±5.09	8.94 ±1.36	58.14 ±5.22	90.57 ±0.55	74.52 ±1.35	0.8760 ±0.0072
DFINet[38]	**98.73** **±0.21**	85.35 ±3.51	87.76 ±3.11	85.69 ±5.20	95.93 ±1.02	67.04 ±11.66	78.75 ±3.69	**98.14** **±0.39**	60.84 ±4.29	14.34 ±2.76	61.46 ±6.87	91.78 ±0.75	75.82 ±1.67	0.8917 ±0.0097
CACL[39]	97.65 ±0.44	84.87 ±5.18	**91.82** **±2.63**	86.69 ±3.67	95.04 ±1.91	60.15 ±9.55	80.25 ±3.88	96.92 ±0.77	54.77 ±6.58	8.64 ±2.19	52.56 ±8.52	90.48 ±0.86	73.58 ±1.58	0.8744 ±0.0110
MFT[40]	93.69 ±1.13	83.69 ±3.12	86.03 ±3.05	**93.48** **±1.61**	87.79 ±1.52	**99.27** **±1.09**	**90.45** **±4.61**	93.90 ±1.41	68.23 ±7.21	**85.06** **±5.78**	**98.88** **±1.17**	90.49 ±0.56	**89.13** **±1.02**	0.8745 ±0.0072
OURS	98.49 ±0.33	**86.33** **±2.70**	90.79 ±1.39	91.51 ±2.05	**97.01** **±0.64**	86.58 ±4.97	77.94 ±5.20	98.06 ±0.62	68.69 ±4.50	22.45 ±5.97	70.55 ±5.70	**93.40** **±0.55**	80.77 ±1.30	**0.9127** **±0.0071**

are shown in Fig. 6(d), it can be found that the temperature parameter does not work at this point, and the probability distributions in the teacher and student models are relatively small in difference. In addition, it is found that distillation loss dominates the training loss of the student network, and the model performs better when the scale parameter of classification loss is taken as 0.6. The same way of handling hyperparameters is adopted for the other datasets, and the proposed optimal setting values are used for all the compared methods.

C. Compared With Multimodal Classification Methods

The proposed multimodal classification method uses the CNN-GCN joint feature learning network to obtain discriminative graph knowledge representation. Among them, the designed PE-RLM and CGAM modules provide a useful tool to enhance feature representation. Therefore, to more comprehensively verify the effectiveness of the proposed method, not only the comparison with the state-of-the-art multimodal collaborative classification methods but also the comparative analysis is performed separately for different modules in the proposed teacher network. The multimodal classification results on the four datasets are available in Tables IV–VII, and the corresponding classification maps are presented in Figs. 7–10.

1) Compared With Graph Construction Methods: Each sample corresponds to a graph node in the graph construction process, the adjacent structural relationship between nodes is represented by the adjacency matrix, and the KNN method is often used to select such structural relationship. Two more common graph construction methods are selected to compare with the proposed PE-RLM method. The MGFN method uses the data similarity relationship between nodes to represent the structural relationship between nodes in the graph and determines whether the nodes are adjacent according to the Euclidean distance, where the original information of each data is directly used as the initial information of the nodes; the ACGC method takes the convolutional features of the data for the initial node representation, predicts the node class relationship, and determines the nodes with similar class probability in the prediction result as the neighboring relationship.

Comparing the classification results on the four datasets, it is found that the proposed method performs the best. The MGFN method is limited in classification accuracy due to the node representation using the original information of the data, which brings more noise; while the ACGC method uses the feature representation of the data to reduce the influence of noise and constructs a dynamic graph structure based on the predicted relationships. It can be seen that the accuracy of the ACGC method is significantly improved compared with the MGFN method. However, as mentioned by us about the problem that different classes of patches have similar feature representations, it is not enough to rely on feature similarity relationships alone. Our proposed method picks neighboring nodes by adding position information and jointly with feature similarity. Compared to the performance of the MGFN method and the ACGC method on OA in four groups of experiments, the highest improvement reached 10.3% and 1.58%, respectively.

2) Compared With Graph Aggregation Methods: The graph aggregation operation updates nodes by aggregating information from neighboring nodes, thus enhancing the feature representation and obtaining the more discriminative feature. Three classical graph aggregation methods are selected to compare with the proposed CGAM. The GCN method uses the summation of neighboring node features followed by degree matrix weighting to merge the neighboring nodes directly

TABLE V
CLASS-SPECIFIC CLASSIFICATION ACCURACY (%) ON THE BERLIN DATASET BY CONSIDERING HS AND SAR DATA

Methods \ Class No.	1	2	3	4	5	6	7	8	OA(%)	AA(%)	κ
MGFN[29]	77.86 ±4.83	93.26 ±1.13	53.45 ±8.92	86.62 ±2.09	75.47 ±7.07	38.31 ±10.78	33.01 ±4.22	58.68 ±7.74	78.85 ±1.36	64.58 ±2.41	0.6868 ±0.0158
ACGC[16]	82.38 ±2.75	95.11 ±1.22	79.23 ±4.49	**94.53 ±1.76**	90.14 ±3.76	49.08 ±6.25	61.15 ±5.85	67.12 ±6.42	87.57 ±0.81	77.34 ±1.75	0.8081 ±0.0111
GCN[32]	52.52 ±6.59	85.28 ±1.96	40.37 ±16.90	55.79 ±11.59	49.87 ±21.70	13.54 ±5.01	21.69 ±7.40	7.50 ±7.06	62.30 ±5.53	40.82 ±5.87	0.4373 ±0.0537
GAT[33]	70.82 ±5.77	85.94 ±2.39	69.68 ±8.56	73.03 ±6.89	79.89 ±10.41	31.47 ±7.81	41.62 ±5.53	48.02 ±13.48	76.00 ±1.88	62.56 ±3.67	0.6124 ±0.0227
Gated-GCN[34]	85.97 ±4.12	94.83 ±0.97	**84.98 ±5.16**	90.29 ±3.64	88.62 ±3.65	52.97 ±9.02	63.95 ±5.39	64.55 ±4.40	88.97 ±0.71	78.27 ±1.60	0.8271 ±0.0093
CoupledCNNs[36]	83.68 ±3.43	89.81 ±1.41	65.21 ±7.09	94.10 ±1.91	84.14 ±4.38	**81.17 ±12.80**	**67.39 ±5.33**	61.19 ±9.94	86.21 ±1.10	78.34 ±2.68	0.7760 ±0.0180
AsyFFNet[37]	85.97 ±3.37	95.82 ±0.65	81.02 ±4.66	94.26 ±1.31	87.90 ±4.22	44.42 ±8.25	44.67 ±5.00	67.57 ±7.28	85.67 ±1.12	75.20 ±1.75	0.7839 ±0.0144
DFINet[38]	83.57 ±4.76	95.70 ±1.30	75.61 ±7.40	91.46 ±2.08	88.31 ±2.96	42.86 ±4.05	52.37 ±5.39	75.90 ±6.16	86.02 ±0.87	75.72 ±2.12	78.78 ±0.0101
CACL[39]	88.37 ±3.21	91.99 ±1.40	71.69 ±6.19	91.89 ±1.65	81.99 ±5.86	64.10 ±6.02	57.50 ±3.57	79.99 ±7.30	87.64 ±0.43	78.44 ±1.42	0.8014 ±0.0062
MFT[40]	**89.11 ±1.50**	90.98 ±1.61	82.60 ±7.49	90.99 ±2.53	**97.03 ±1.01**	74.92 ±9.73	57.64 ±5.89	**87.45 ±5.30**	88.36 ±0.47	**83.84 ±2.13**	0.8169 ±0.0067
OURS	87.22 ±1.60	**95.90 ±0.75**	79.22 ±6.72	93.28 ±1.05	88.68 ±4.01	54.57 ±4.69	64.68 ±5.19	64.65 ±5.71	**89.15 ±0.60**	78.52 ±1.17	**0.8319 ±0.0090**

TABLE VI
CLASS-SPECIFIC CLASSIFICATION ACCURACY (%) ON THE AUGSBURG DATASET BY CONSIDERING HS AND SAR DATA

Methods \ Class No.	1	2	3	4	5	6	7	OA(%)	AA(%)	κ
MGFN[29]	93.80 ±0.92	86.47 ±1.22	64.86 ±2.56	97.42 ±0.61	25.09 ±3.42	26.14 ±4.53	70.10 ±10.68	89.44 ±0.37	66.27 ±1.76	0.8461 ±0.0057
ACGC[16]	97.00 ±0.61	94.61 ±0.88	67.94 ±2.06	**99.29 ±0.21**	85.44 ±7.90	49.98 ±15.15	78.91 ±7.37	94.60 ±0.21	81.88 ±2.71	0.9223 ±0.0030
GCN[32]	86.25 ±2.79	83.65 ±1.26	57.40 ±3.01	88.68 ±1.89	22.78 ±4.21	27.54 ±13.03	37.22 ±12.71	83.98 ±0.81	57.65 ±2.89	0.7653 ±0.0115
GAT[33]	93.73 ±1.35	92.86 ±1.00	58.28 ±6.68	95.58 ±1.62	75.97 ±12.20	32.19 ±7.40	54.02 ±13.83	90.43 ±0.95	71.80 ±3.21	0.8626 ±0.0134
Gated-GCN[34]	97.30 ±0.61	95.19 ±0.67	70.38 ±2.59	99.03 ±0.22	**87.70 ±3.69**	48.53 ±7.90	77.59 ±8.62	94.77 ±0.43	82.24 ±2.54	0.9248 ±0.0061
CoupledCNNs[36]	96.12 ±0.70	96.12 ±0.58	65.53 ±2.18	98.43 ±0.42	92.72 ±4.65	42.54 ±7.73	71.67 ±7.22	94.32 ±0.13	80.45 ±1.45	0.9185 ±0.0020
AsyFFNet[37]	95.32 ±0.99	94.91 ±1.13	67.95 ±2.10	97.76 ±0.69	46.58 ±8.36	34.14 ±42.56	**80.21 ±14.68**	93.43 ±0.37	73.84 ±4.92	0.9055 ±0.0054
DFINet[38]	96.36 ±1.06	95.08 ±0.58	71.93 ±1.55	98.62 ±0.35	64.52 ±4.84	**77.23 ±11.28**	73.49 ±6.17	94.50 ±0.23	82.46 ±1.98	0.9208 ±0.0034
CACL[39]	96.12 ±0.78	95.72 ±0.64	70.69 ±2.86	97.08 ±0.52	81.76 ±7.55	59.31 ±12.10	76.89 ±6.79	94.28 ±2.14	82.51 ±1.49	0.9176 ±0.0031
MFT[40]	97.39 ±0.74	**96.59 ±0.50**	**77.94 ±7.52**	96.87 ±0.45	85.76 ±7.47	15.05 ±5.36	55.11 ±3.40	93.30 ±0.34	74.96 ±1.63	0.9037 ±0.0051
OURS	**97.88 ±0.45**	95.82 ±0.72	72.04 ±1.49	99.15 ±0.22	84.67 ±5.15	51.82 ±6.34	76.51 ±6.44	**95.22 ±0.26**	**82.56 ±1.33**	**0.9312 ±0.0037**

without selection. The GAT method takes an attentional approach to learn the importance of the neighboring nodes to its own nodes and integrates the neighboring information selectively. The Gated-GCN method further adopts a relational gate to select similar neighboring nodes to aggregate with its own nodes.

Comparing the classification results on the four datasets reveals that the proposed method performs the best followed by the Gated-GCN method. This is because they retain the information of their own nodes while selectively absorbing the information of neighboring nodes compared to the GAT method, which ensures the discriminability of features. In spite of the fact that the GCN method aggregates the most information, it does not mean that its discriminability is the strongest. Furthermore, the proposed method learns the channel correlation between neighboring nodes with its own nodes and uses channel gating to control the information transfer of channel features from neighboring nodes to its own nodes, so as to enhance the class feature representation capability. In comparison to the second-best performing Gated-GCN method, the OA is improved by 0.35%, 0.18%, 0.45%, and 0.43% on each of the four experiments.

TABLE VII
CLASS-SPECIFIC CLASSIFICATION ACCURACY (%) ON THE AUGSBURG DATASET BY CONSIDERING HS AND DSM DATA

Methods \ Class No.	1	2	3	4	5	6	7	OA(%)	AA(%)	κ
MGFN[29]	91.22 ±0.92	88.00 ±1.18	64.47 ±1.66	94.96 ±1.11	20.98 ±1.62	25.94 ±2.88	67.01 ±12.57	88.13 ±0.49	64.66 ±2.10	0.8287 ±0.0071
ACGC[16]	96.61 ±0.99	95.17 ±0.99	69.30 ±2.14	98.85 ±0.36	89.40 ±6.01	58.42 ±19.62	76.99 ±6.91	94.69 ±0.15	83.53 ±3.01	0.9236 ±0.0023
GCN[32]	84.49 ±2.24	81.02 ±2.05	52.96 ±6.28	79.94 ±2.29	20.90 ±6.68	23.09 ±18.61	38.12 ±9.71	79.31 ±1.46	54.36 ±2.17	0.6964 ±0.0214
GAT[33]	93.59 ±1.32	91.84 ±1.77	56.14 ±4.74	94.87 ±1.95	65.17 ±13.20	33.96 ±6.14	55.33 ±9.25	89.64 ±1.03	70.13 ±2.83	0.8512 ±0.0147
Gated-GCN[34]	97.44 ±0.46	95.36 ±0.60	73.81 ±2.27	98.71 ±0.38	85.39 ±3.20	53.26 ±7.79	75.61 ±6.44	94.96 ±0.19	82.80 ±1.55	0.9274 ±0.0027
CoupledCNNs[36]	96.57 ±0.69	95.57 ±0.95	67.35 ±1.65	98.69 ±0.32	80.26 ±12.24	36.53 ±5.81	74.06 ±8.61	94.37 ±0.38	78.43 ±1.85	0.9190 ±0.0055
AsyFFNet[37]	95.19 ±1.26	95.64 ±0.53	70.23 ±2.41	97.97 ±0.29	51.57 ±11.64	27.29 ±34.60	**80.27 ±12.80**	93.92 ±0.40	74.02 ±4.05	0.9127 ±0.0056
DFINet[38]	96.96 ±0.58	95.21 ±0.51	72.63 ±0.93	98.30 ±0.44	73.36 ±5.03	68.14 ±12.61	73.58 ±6.69	94.67 ±0.18	82.59 ±2.07	0.9232 ±0.0026
CACL[39]	96.68 ±0.83	95.43 ±0.57	71.43 ±1.21	97.53 ±0.53	**90.63 ±3.29**	**58.76 ±12.37**	78.03 ±9.21	94.50 ±0.20	84.07 ±2.09	0.9207 ±0.0029
MFT[40]	**98.15 ±0.95**	**96.98 ±0.52**	82.60 ±4.48	96.97 ±0.32	88.45 ±5.17	9.56 ±5.87	53.25 ±3.41	93.71 ±0.23	75.14 ±1.16	0.9098 ±0.0033
OURS	97.72 ±0.41	95.83 ±0.35	**74.42 ±1.83**	**98.98 ±0.17**	86.13 ±3.27	58.00 ±7.14	78.63 ±4.57	**95.39 ±0.17**	**84.25 ±0.97**	**0.9336 ±0.0024**

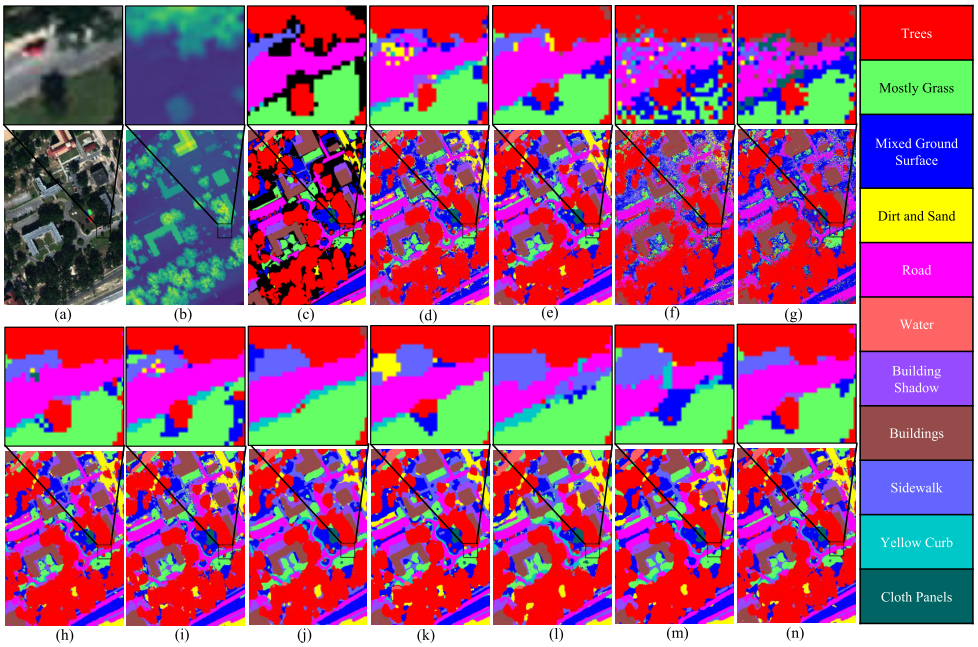

Fig. 7. MUUFL. (a) HS. (b) LiDAR. (c) Ground truth. (d)–(n) Classification maps for different multimodal classification methods. (d) MGFN. (e) ACGC. (f) GCN. (g) GAT. (h) Gated-GCN. (i) CoupledCNNs. (j) AsyFFNet. (k) DFINet. (l) CACL. (m) MFT. (n) OURS.

3) Compared With Collaborative Classification Methods: The quality of the extracted merged information from multimodal RS data is very important, and it will be used as teacher knowledge to guide student network learning. Five state-of-the-art multimodal RS collaborative classification methods are selected to compare with the proposed teacher network. Of these, CNN-based methods (CoupledCNNs, AsyFFNet, DFINet, and CALC) are the latest work using active radar data (SAR and LiDAR) to mitigate the effects of clouds, while the MFT method is the latest transformer method based on position information embedding, and they are considered to be in sharp contrast to the proposed work.

It is observed that the CNN-based methods achieve a certain level of classification performance in several sets of experiments. The MFT method is a transformer-based method that additionally captures spectral sequence information but fails to achieve the best classification performance due to its failure to describe local semantic information as well. Nevertheless, our method achieves the best performance in several experiments, with an improvement of 1.62%, 0.79%, 0.72%, and 0.72% on

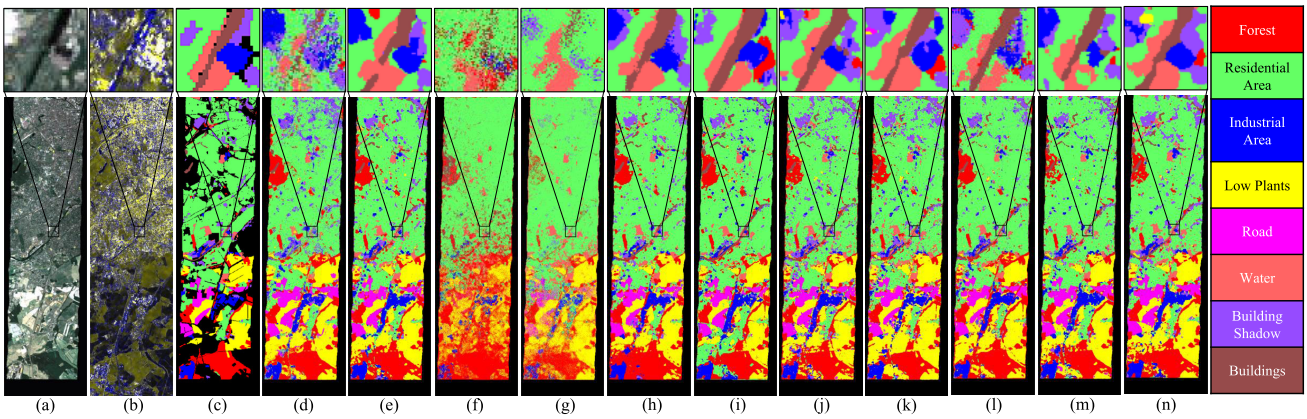

Fig. 8. Berlin. (a) HS. (b) SAR. (c) Ground truth. (d)–(n) Classification maps for different multimodal classification methods. (d) MGFN. (e) ACGC. (f) GCN. (g) GAT. (h) Gated-GCN. (i) CoupledCNNs. (j) AsyFFNet. (k) DFINet. (l) CACL. (m) MFT. (n) OURS.

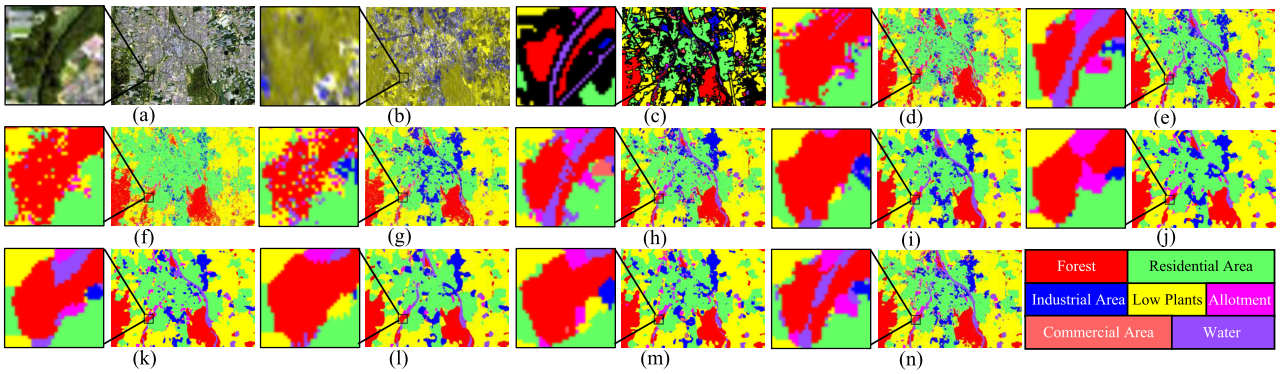

Fig. 9. Augsburg. (a) HS. (b) SAR. (c) Ground truth. (d)–(n) Classification maps for different multimodal classification methods. (d) MGFN. (e) ACGC. (f) GCN. (g) GAT. (h) Gated-GCN. (i) CoupledCNNs. (j) AsyFFNet. (k) DFINet. (l) CACL. (m) MFT. (n) OURS.

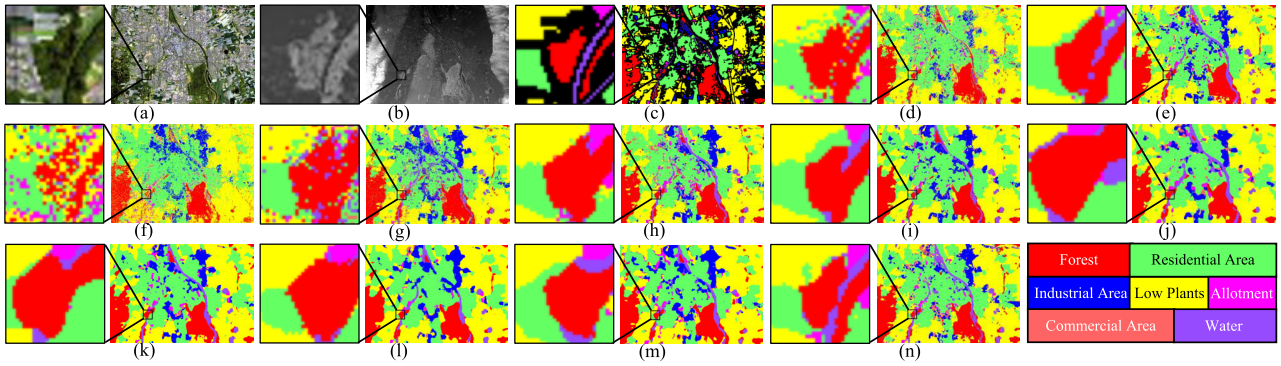

Fig. 10. Augsburg. (a) HS. (b) DSM. (c) Ground truth. (d)–(n) Classification maps for different multimodal classification methods. (d) MGFN. (e) ACGC. (f) GCN. (g) GAT. (h) Gated-GCN. (i) CoupledCNNs. (j) AsyFFNet. (k) DFINet. (l) CACL. (m) MFT. (n) OURS.

OA compared to the subbest method, which is attributed to the additional position and topology information learned.

Combining Tables IV–VII and Figs. 7–10, we analyze the specific classification performance of each class in each dataset. In the comparative experiments involving graph construction and aggregation, methods such as the MGFN, GCN, and GAT generally yield lower classification accuracy on the four datasets. They also manifest rather blurred classification boundaries in the highlighted classification maps. However, both the ACGC method and the GatedGCN method exhibit competitive classification performance relative to our proposed method, even outpacing state-of-the-art collaborative classification methods. Yet, from the perspective of experimental settings, all these methods merely replace corresponding modules within the proposed framework. Their final classification results are solely used for comparison analysis with the proposed method's corresponding module performances. Nevertheless, our proposed method exhibits superior recognition effects for the complex "Water" class in classification maps, as shown in Fig. 10(e), (h), and (n). The ultimate performance of the method demonstrates a competitive edge over several advanced collaborative classification methods. It should be noted in particular that the MFT method achieves a surprising performance on the AA of the MUUFL and

TABLE VIII

CROSS-MODAL DISTILLATION RESULTS ON THE MUUFL, BERLIN, AND AUGSBURG DATASETS. (∗) DENOTES ANOTHER MODALITY OUTSIDE THE HS AT EACH SET OF EXPERIMENTS, AND JOINTS DENOTES COMBINED MULTIMODAL

#	Methods	Test Modalities	HS-LiDAR MUUFL			HS-SAR Berlin			HS-SAR Augsburg			HS-DSM Augsburg		
			OA(%)	AA(%)	κ	OA(%)	AA(%)	κ	OA(%)	AA(%)	κ	OA(%)	AA(%)	κ
1	KD[12]	HS	92.92	78.90	0.9064	88.48	77.68	0.8263	95.34	83.08	0.9331	95.30	84.96	0.9324
2	Hint[13]	HS	93.00	80.49	0.9070	88.96	78.94	0.8240	95.39	82.85	0.9338	95.38	83.49	0.9336
3	MSGD[25]	HS	93.17	79.97	0.9094	89.13	77.76	0.8318	95.45	**85.81**	0.9346	95.33	84.73	0.9328
4	ADMD[26]	HS	93.25	80.49	0.9103	89.56	79.19	0.8396	95.48	85.74	0.9348	95.43	**85.88**	0.9344
5	LSP[15]	HS	93.04	80.12	0.9076	89.28	77.56	0.8360	95.45	83.98	0.9347	95.45	85.00	0.9345
6	HKD[16]	HS	93.27	80.46	0.9104	89.52	78.72	0.8401	95.52	85.30	0.9355	95.50	84.23	0.9351
7	KD[12]	(∗)	70.68	51.09	0.6312	69.09	58.06	0.5085	87.09	65.52	0.8111	82.95	57.76	0.7481
8	Hint[13]	(∗)	70.92	52.41	0.6230	69.18	58.35	0.5094	87.28	64.65	0.8153	83.00	59.48	0.7488
9	MSGD[25]	(∗)	70.30	51.50	0.6196	72.11	60.08	0.5495	87.85	66.73	0.8214	82.66	59.63	0.7423
10	ADMD[26]	(∗)	71.62	51.52	0.6344	72.19	54.66	0.5329	88.11	66.85	0.8244	83.49	51.37	0.7541
11	LSP[15]	(∗)	71.40	52.00	0.6349	69.25	57.93	0.5103	87.55	63.17	0.8176	83.27	59.66	0.7521
12	HKD[16]	(∗)	70.98	51.28	0.6341	72.07	55.03	0.5470	88.04	67.09	0.8245	83.43	54.85	0.7542
13	OURS(Teacher)	*Joints*	*94.30*	*83.61*	*0.9241*	*90.32*	*80.30*	*0.8499*	*95.57*	*83.91*	*0.9365*	*95.78*	*85.51*	*0.9393*
14	OURS(Student)	HS	*92.89*	*79.31*	*0.9058*	*88.05*	*75.69*	*0.8161*	*95.23*	*83.05*	*0.9314*	*95.23*	*83.05*	*0.9314*
15		(∗)	*68.75*	*49.03*	*0.6056*	*68.84*	*57.43*	*0.5040*	*86.47*	*65.97*	*0.8020*	*82.54*	*59.24*	*0.7415*
16	OURS(Distillation)	HS	**93.72**	**80.50**	**0.9165**	**89.90**	**79.45**	**0.8459**	**95.61**	85.36	**0.9370**	**95.52**	85.02	**0.9354**
17		(∗)	**72.23**	**52.51**	**0.6486**	**72.89**	**60.30**	**0.5712**	**88.86**	**67.57**	**0.8376**	**84.11**	**60.45**	**0.7660**

Berlin datasets, especially on class 10 in the MUUFL dataset, although, at this point, they are still inferior to the proposed method on OA and κ. However, it is interesting that the MFT method again performs much worse than the proposed method on the Augsburg dataset, especially on class 6, where none of the classification results is very good. Observing the training data on these anomalous classes, it is not difficult to attribute this phenomenon to the data-hungry nature of the transformers-based method. In comparison with the unstable performance of the MFT method, our method still achieves a classification performance closer to that of ground truth among all the compared methods.

D. Compared With Cross-Modal Distillation Methods

In the process of cross-modal knowledge distillation, a multigranularity graph distillation method is proposed to perform intermediate feature distillation and soft-label distillation, respectively, and the extracted teacher knowledge is leveraged to guide the learning of the unimodal student network. In particular, the FD-GDM is developed to stabilize the distillation process of intermediate features. Thus, the vanilla distillation method is not only selected but also compared with the traditional generative distillation methods. In addition, the latest graph-based distillation methods are selected for the proposed graph distillation module for comparison with it.

The KD method, as the pioneer of knowledge distillation, has the output of the teacher network as a soft label giving the correlation between classes. It can be used as a regular term to constrain the distribution of parameters in the student network. The Hint method, on the other hand, incorporates feature map distillation to avoid overfitting. Both of them once served as classical comparison methods among existing knowledge distillation methods. The MSGD and ADMD methods are typical generative distillation methods that use minimized feature distance and a learnable generative adversarial network to generate discriminative features of missing data, respectively, thus simulating the hallucination data of missing data for the multimodal fusion classification model. While the graph-based distillation method is used to transfer the transferable knowledge from the teacher network to the student network in the distillation process in the form of graphs, the focus is on the representation of graph knowledge. The LSP approach utilizes a local structural preserving module to represent graph knowledge locally, enabling compact distillation learning by minimizing the local structure distance between the teacher and the student. Instead, the HKD method uses graph knowledge as overall knowledge for distillation and applies a mutual information metric to measure the amount of overall knowledge learned by the student network from the teacher network.

Table VIII shows the cross-modal distillation classification results on the four datasets, where the teacher network uses the best pretrained classification model of the proposed multimodal fusion network in ten experiments and the student network represents the unimodal network without teacher knowledge guidance. Comparing the classification results can summarize the following findings.

1) Due to the rich information of HS images, their classification performance under a single modality is close to that of the teacher, and the cross-modal classification performance improvement on this modality is limited regardless of the proposed method or the compared method, especially on the Augsburg dataset. In contrast, for the modality besides the HS image in the dataset, the classification performance is much less than that of the teacher, and the cross-modal classification performance improvement on this modality is obvious, reaching a maximum of 4.05% improvement.
2) The first two rows in Fig. 11 show the classification results of three types of distillation methods (vanilla methods, generative methods, and graph-based methods) on four datasets in the form of box plots. Among them, vanilla distillation methods improve the classification performance of student networks less, and generative distillation methods and graph-based distillation methods perform better and are relatively stable.

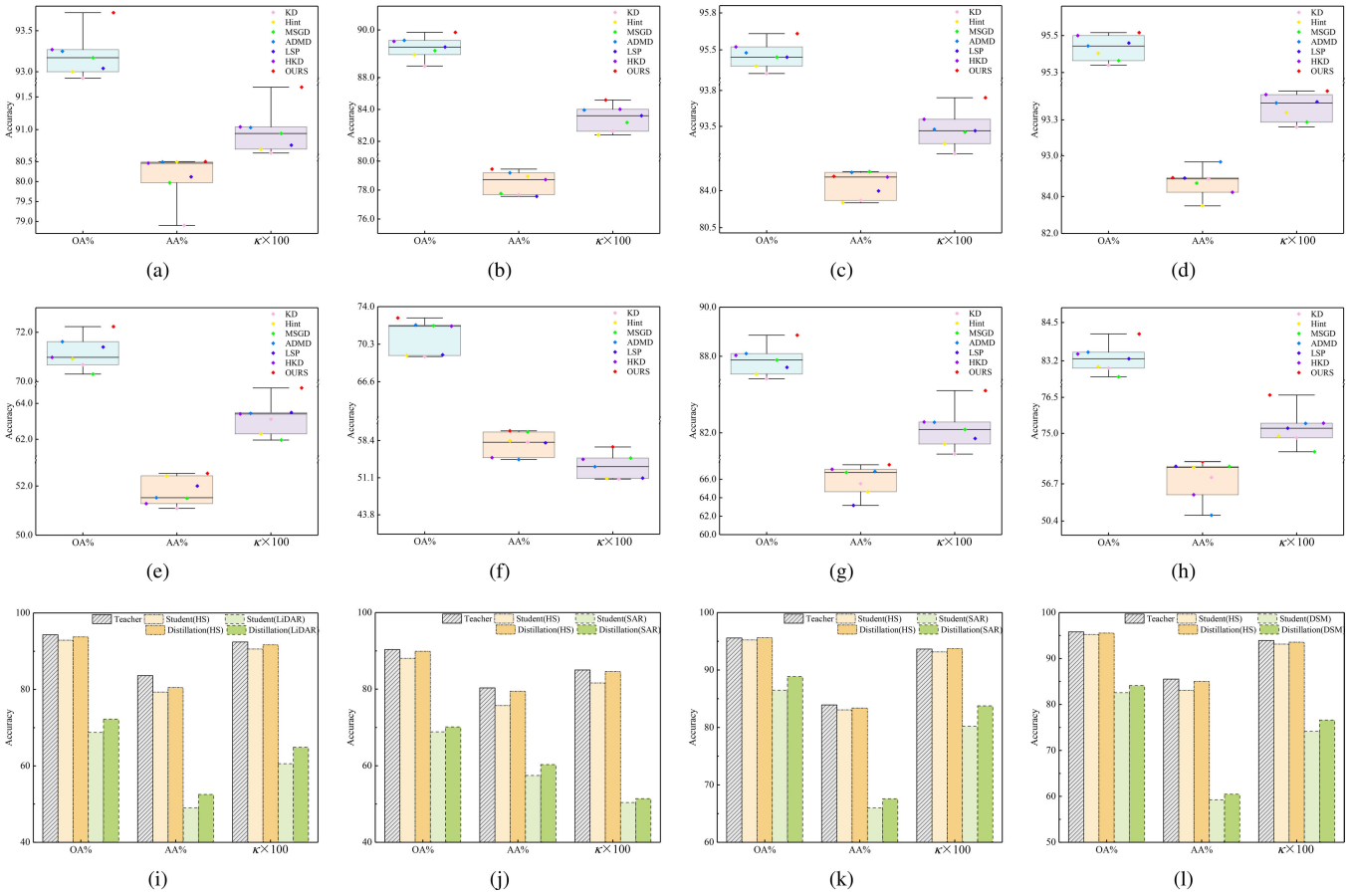

Fig. 11. Visualization results of cross-modal distillation classification under different modalities. (a)–(d) HS modality of the HS-LiDAR MUUFL dataset, the HS-SAR Berlin dataset, the HS-SAR Augsburg dataset, and the HS-DSM Augsburg dataset. (e) LiDAR modality of the HS-LiDAR MUUFL dataset. (f) SAR modality of the HS-SAR Berlin dataset. (g) SAR modality of the HS-SAR Augsburg dataset. (h) DSM modality of the HS-DSM Augsburg dataset. (i)–(l) Comparison results of the proposed method on four datasets for teacher, student, and distillation networks, respectively.

However, regardless of the distillation performance based on which modality, the proposed method has some competitive power.

3) The good distillation performance of each modality on the four datasets is visualized in the last row of the bar chart in Fig. 11 by comparing the classification results of the teacher and student networks.

E. Ablation Study

To verify the effectiveness of each of the proposed modules in the CGKR-DL framework in this article, the ablation experiments are designed in the collaborative classification and cross-modal distillation networks, respectively. The classification results are shown in Tables IX and X. Fig. 12 shows the effect of each module more intuitively.

Taking the MUUFL dataset as an example, #1 in Table IX indicates the proposed method without graph convolution operation, i.e., the classification network using only CNN method, #2 indicates the proposed method with CNN-GCN joint feature learning but without CGAM module for node information aggregation but with ordinary gated aggregation, and #3 indicates the proposed method. It is observed that the accuracy of classification using only CNN is generally low, and with the addition of the PE-RLM module and CGAM module, the classification accuracy is gradually improved. Furthermore, the proposed CNN-GCN joint feature learning approach achieves the largest improvement in classification accuracy compared to the CNN-only approach, with a maximum improvement of 1.63% on OA among four sets of experiments.

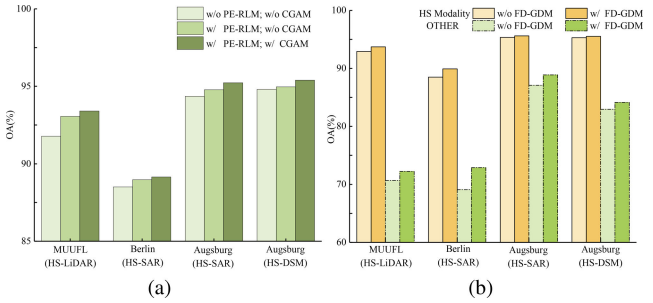

Fig. 12. Comparison of classification results for different module roles. (a) PE-RLM and CGAM. (b) FD-GDM.

The cross-modal distillation ablation experiments shown in Table X further validate the effectiveness of the proposed FD-GDMs. Among them, #1 and #3 represent the soft-label distillation experiments of the student network in HS and LiDAR modalities, respectively, while #2 and #4 represent the multigranularity distillation experiments of the student

TABLE IX
ABLATION EXPERIMENT COLLABORATIVE CLASSIFICATION RESULTS ON THE MUUFL, BERLIN, AND AUGSBURG DATASETS

#	Datasets	PE-RLM	CGAM	Metrice		
				OA(%)	AA(%)	κ
1	MUUFL (HS-LiDAR)			91.77 ±0.33	78.25 ±0.73	0.8914 ±0.0043
2		✓		93.05 ±0.42	80.69 ±1.69	0.9071 ±0.0055
3		✓	✓	**93.40 ±0.55**	**80.77 ±1.30**	**0.9127 ±0.0071**
4	Berlin (HS-SAR)			88.50 ±0.67	77.32 ±1.74	0.8215 ±0.0084
5		✓		88.97 ±0.71	78.27 ±1.60	0.8271 ±0.0093
6		✓	✓	**89.15 ±0.60**	**78.52 ±1.17**	**0.8319 ±0.0090**
7	Augsburg (HS-SAR)			94.35 ±0.38	80.01 ±3.63	0.9185 ±0.0056
8		✓		94.77 ±0.43	82.24 ±2.54	0.9248 ±0.0061
9		✓	✓	**95.22 ±0.26**	**82.56 ±1.33**	**0.9312 ±0.0037**
10	Augsburg (HS-DSM)			94.80 ±0.22	82.08 ±3.77	0.9250 ±0.0031
11		✓		94.96 ±0.19	82.80 ±1.55	0.9274 ±0.0027
12		✓	✓	**95.39 ±0.17**	**84.25 ±0.97**	**0.9336 ±0.0024**

TABLE X
ABLATION EXPERIMENT CROSS-MODAL DISTILLATION RESULTS ON THE MUUFL, BERLIN, AND AUGSBURG DATASETS

#	Datasets	Test Modalities	KD	FD-GDM	Metrice		
					OA(%)	AA(%)	κ
1	MUUFL (HS-LiDAR)	HS	✓		92.92	78.90	0.9064
2		HS	✓	✓	**93.72**	**80.50**	**0.9165**
3		LiDAR	✓		70.68	51.09	0.6312
4		LiDAR	✓	✓	**72.23**	**52.51**	**0.6486**
5	Berlin (HS-SAR)	HS	✓		88.48	77.68	0.8263
6		HS	✓	✓	**89.90**	**79.45**	**0.8459**
7		SAR	✓		69.09	58.06	0.5085
8		SAR	✓	✓	**72.89**	**60.30**	**0.5712**
9	Augsburg (HS-SAR)	HS	✓		95.34	83.08	0.9331
10		HS	✓	✓	**95.61**	**83.36**	**0.9370**
11		SAR	✓		87.09	65.52	0.8111
12		SAR	✓	✓	**88.86**	**67.57**	**0.8376**
13	Augsburg (HS-DSM)	HS	✓		95.30	84.96	0.9324
14		HS	✓	✓	**95.52**	**85.02**	**0.9354**
15		DSM	✓		82.95	57.76	0.7481
16		DSM	✓	✓	**84.11**	**60.45**	**0.7660**

network in HS and LiDAR modalities, respectively, i.e., the proposed FD-GDM module is added to the soft-label distillation. The addition of the feature distillation process further improves the classification accuracy of the student network, especially its distillation OA accuracy can best reach a 3.8% improvement on the modalities with weaker classification performance.

F. Visualization Analysis

To visually demonstrate the advantages of the proposed CNN-GCN joint feature learning approach, CNN and CNN-GCN joint classification performances based on the proposed method are visualized for the MUUFL dataset in a teacher knowledge extraction network, respectively. Fig. 13 shows the visualization results using the t-SNE method, and it can be found that the data features obtained using only the CNN method are more densely distributed and the interclass distance is smaller. It is especially the part marked by the red box that shows poor discriminability between classes, while the proposed CNN-GCN joint feature learning approach benefits from the enhancement of structure and position information to obtain clearer class boundaries.

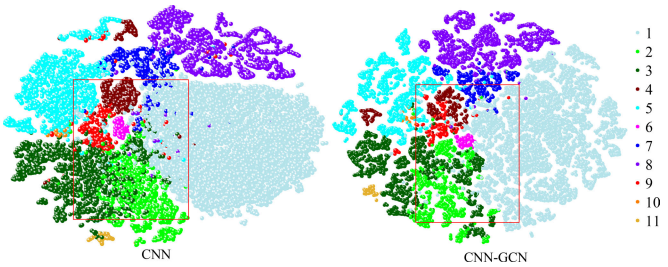

Fig. 13. Visualization results of features based on the t-SNE method.

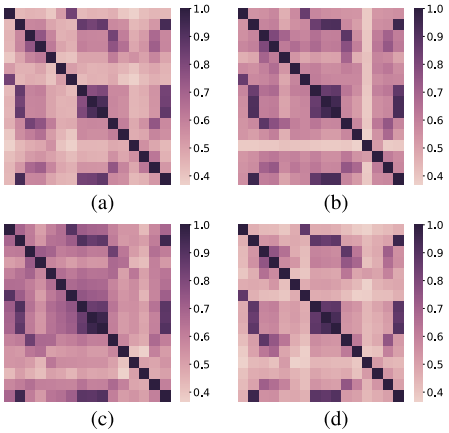

Fig. 14. Heat map visualization results of four networks. (a) Teacher. (b) Student. (c) HKD. (d) OURS.

Moreover, the similarity relationship between the feature pairs corresponding to the teacher network and the student network in cross-modal distillation is visualized and analyzed. We design the visualization experiments by first selecting 16 samples randomly on the MUUFL dataset, which are fed into the trained teacher network ("Teacher"), the student network ("Student"), the global knowledge-based graph distillation network (HKD), and the proposed MGDN (OURS), respectively. The similarity relationship between the samples is represented by calculating the cosine similarity of the extracted features, and their results are presented as a heat map in Fig. 14. Each block in the figure indicates the similarity between samples, and the darker the color, the higher the similarity. Comparing "Teacher" and "Student" reveals that there is some difference in the characterization of the relationship between "Teacher" and "Student" for the same sample pair, which implies that the method of distillation relying only on the local relationship of the sample pair is not optimal, and it prompts the inclusion of the global relationship. While the overall structure between sample pairs exhibits some similarity in the relationship, the HKD method is based on this global

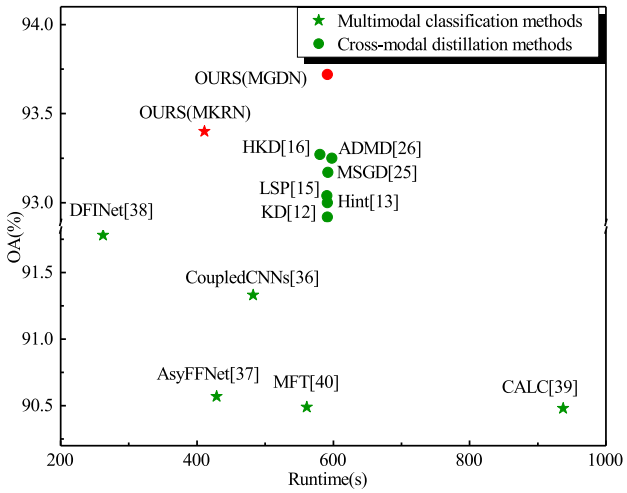

Fig. 15. Results of different models in terms of runtime and training accuracy.

relationship for distillation, and unfortunately, its visualization results do not perform well on local sample pairs. Comparing our proposed method with the HKD method, the proposed method combines local and global structural relationships for graph distillation, and its visualization result is much closer to that of "Teacher."

G. Model Runtime Evaluation

In this section, an evaluation of the model is performed on the MUUFL dataset in conjunction with the model's runtime and training accuracy. In Fig. 15, the proposed multimodal knowledge representation network (MKRN) is compared with the state-of-the-art collaborative classification network and performs a comparison between the proposed MGDN and its comparison methods, respectively. It can be found that the proposed method performs very well among all the compared methods, taking into account both runtime and training accuracy. Furthermore, the small difference in runtime between the different distillation methods implies that the improved distillation process does not impose a significant time cost while improving the classification ability of the student network.

V. Conclusion

In this article, a CGKR-DL framework is proposed, which cooperates with the complementary characteristics of multimodal data to guide cross-modal RS image classification under data missing. In the teacher network, a joint CNN-GCN feature learning network is used to perform graph knowledge representation, where the designed PE-RLM and CGAM modules further enhance the knowledge representation ability of the network. In the student network, the graph knowledge extracted from the teacher network is used to guide the student network learning in a multigranularity graph distillation method, in which the developed FD-GDM module stabilizes the knowledge transfer process. In cross-modal RS image classification tasks under data missing, it is considered equally important to mine multimodal information and collaborative manner of complementary characteristics. Unfortunately, the proposed method does not fully investigate the latter, which will be one of the future research directions.

References

[1] H. Su, F. Shao, Y. Gao, H. Zhang, W. Sun, and Q. Du, "Probabilistic collaborative representation based ensemble learning for classification of wetland hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5509517.

[2] M. Zhang, X. Zhao, W. Li, and Y. Zhang, "Multi-source remote sensing data cross scene classification based on multi-graph matching," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Kuala Lumpur, Malaysia, Jul. 2022, pp. 827–830.

[3] W. Lee, J. Lee, D. Kim, and B. Ham, "Learning with privileged information for efficient image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, vol. 12369. Cham, Switzerland: Springer, Nov. 2020, pp. 465–482.

[4] X. Li, L. Lei, C. Zhang, and G. Kuang, "Dense adaptive grouping distillation network for multimodal land cover classification with privileged modality," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4411114.

[5] Y. Lv, W. Xiong, X. Zhang, and Y. Cui, "Fusion-based correlation learning model for cross-modal remote sensing image retrieval," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[6] H. Liu, Y. Qu, and L. Zhang, "Multispectral scene classification via cross-modal knowledge distillation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5409912.

[7] H. Su, Y. Hu, H. Lu, W. Sun, and Q. Du, "Diversity-driven multikernel collaborative representation ensemble for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 2861–2876, 2022.

[8] K. Yan, M. Zhou, L. Liu, C. Xie, and D. Hong, "When pansharpening meets graph convolution network and knowledge distillation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5408915.

[9] Y. Dong, Q. Liu, B. Du, and L. Zhang, "Weighted feature fusion of convolutional neural network and graph attention network for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 31, pp. 1559–1572, 2022.

[10] K. Xu, H. Huang, and P. Deng, "CNN-GCN joint network for remote sensing scene classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Brussels, Belgium, Jul. 2021, pp. 4920–4923.

[11] Q. Liu, L. Xiao, J. Yang, and Z. Wei, "CNN-enhanced graph convolutional network with pixel- and superpixel-level feature fusion for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8657–8671, Oct. 2021.

[12] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *Comput. Sci.*, vol. 14, no. 7, pp. 38–39, 2015.

[13] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015, pp. 1–12.

[14] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 1365–1374.

[15] Y. Yang, J. Qiu, M. Song, D. Tao, and X. Wang, "Distilling knowledge from graph convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 7072–7081.

[16] S. Zhou et al., "Distilling holistic knowledge with graph neural networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 10367–10376.

[17] Y. Gao, M. Zhang, J. Wang, and W. Li, "Cross-scale mixing attention for multisource remote sensing data fusion and classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5507815.

[18] P. Guan and E. Y. Lam, "Multistage dual-attention guided fusion network for hyperspectral pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5515214.

[19] Y. Gao, M. Zhang, W. Li, X. Song, X. Jiang, and Y. Ma, "Adversarial complementary learning for multisource remote sensing classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5505613.

[20] M. Wang, F. Gao, J. Dong, H.-C. Li, and Q. Du, "Nearest neighbor-based contrastive learning for hyperspectral and LiDAR data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5501816.

[21] J. Hu, L. Mou, A. Schmitt, and X. X. Zhu, "FusioNet: A two-stream convolutional neural network for urban scene classification using PolSAR and hyperspectral data," in *Proc. Joint Urban Remote Sens. Event (JURSE)*, Dubai, United Arab Emirates, Mar. 2017, pp. 1–4.

[22] R. Luo, W. Liao, H. Zhang, Y. Pi, and W. Philips, "Classification of cloudy hyperspectral image and LiDAR data based on feature fusion and decision fusion," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Beijing, China, Jul. 2016, pp. 2518–2521.

[23] X. Sun, L. Zhang, H. Yang, T. Wu, Y. Cen, and Y. Guo, "Enhancement of spectral resolution for remotely sensed multispectral image," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 5, pp. 2198–2211, May 2015.

[24] D. Hong, J. Chanussot, N. Yokoya, J. Kang, and X. X. Zhu, "Learning-shared cross-modality representation using multispectral-LiDAR and hyperspectral data," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1470–1474, Aug. 2020.

[25] N. C. Garcia, P. Morerio, and V. Murino, "Modality distillation with multiple stream networks for action recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 106–121.

[26] N. C. Garcia, P. Morerio, and V. Murino, "Learning with privileged information via adversarial discriminative modality distillation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2581–2593, Oct. 2020.

[27] S. Pande, A. Banerjee, S. Kumar, B. Banerjee, and S. Chaudhuri, "An adversarial approach to discriminative modality distillation for remote sensing image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Seoul, South Korea, Oct. 2019, pp. 4571–4580.

[28] S. Wei, Y. Luo, X. Ma, P. Ren, and C. Luo, "MSH-Net: Modality-shared hallucination with joint adaptation distillation for remote sensing image classification using missing modalities," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4402615.

[29] X. Du, X. Zheng, X. Lu, and A. A. Doudkin, "Multisource remote sensing data classification with graph fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10062–10072, Dec. 2021.

[30] Y. Li, Y. Chong, S. Pan, and Y. Ding, "First-order smoothing-based deep graph network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5504716.

[31] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2021, doi: 10.1109/TGRS.2020.3015157.

[32] Z. Liu and J. Zhou, *Introduction to Graph Neural Networks*. San Rafael, CA, USA: Morgan & Claypool, 2020.

[33] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," 2017, *arXiv:1710.10903*.

[34] X. Bresson and T. Laurent, "Residual gated graph ConvNets," in *Proc. Int. Conf. Learn. Represent.*, New Orleans, LA, USA, Feb. 2018, pp. 1–20.

[35] Z. Xue, X. Yu, B. Liu, X. Tan, and X. Wei, "HResNetAM: Hierarchical residual network with attention mechanism for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3566–3580, 2021.

[36] R. Hang, Z. Li, P. Ghamisi, D. Hong, G. Xia, and Q. Liu, "Classification of hyperspectral and LiDAR data using coupled CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4939–4950, Jul. 2020.

[37] W. Li, Y. Gao, M. Zhang, R. Tao, and Q. Du, "Asymmetric feature fusion network for hyperspectral and SAR image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Feb. 18, 2022, doi: 10.1109/TNNLS.2022.3149394.

[38] Y. Gao et al., "Hyperspectral and multispectral classification for coastal wetland using depthwise feature interaction network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5512615.

[39] T. Lu, K. Ding, W. Fu, S. Li, and A. Guo, "Coupled adversarial learning for fusion classification of hyperspectral and LiDAR data," *Inf. Fusion*, vol. 93, pp. 118–131, May 2023.

[40] S. K. Roy, A. Deria, D. Hong, B. Rasti, A. Plaza, and J. Chanussot, "Multimodal fusion transformer for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5515620.

Wenzhen Wang received the M.S. degree from the School of Electric and Electronic Engineering, Shanghai University of Engineering Science, Shanghai, China, in 2021. She is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China.

Her fields of interest include deep learning and remote sensing image processing.

Fang Liu (Member, IEEE) received the B.S. degree in information and computing science from Henan University, Kaifeng, China, in 2012, and the Ph.D. degree in intelligent information processing from Xidian University, Xi'an, China, in 2018.

She is currently an Associate Professor with the Nanjing University of Science and Technology, Nanjing, China. Her research interests include deep learning, object detection, polarimetric synthetic aperture radar (SAR) image classification, and change detection.

Wenzhi Liao (Senior Member, IEEE) received the Ph.D. degree in computer science engineering from Ghent University, Ghent, Belgium, in 2012.

From 2012 to 2019, he worked at Ghent University first as a Post-Doctoral Researcher and then as a Research Fellow with the Research Foundation Flanders (FWO), Vlaanderen, Belgium. Since 2020, he has been a Data Scientist with the Flemish Institute for Technological Research (VITO), Flanders, Belgium, and a Professor with Ghent University. In particular, he has coordinated several national and international projects, and successfully applied his developed methods in the fields of optical hyperspectral image restoration and interpretation, data fusion and classification of multimodal remote sensing imagery for Earth observation, food sorting, and precision agriculture. His research interests include image processing and interpretation (ranging from satellite remote sensing to microscopy), multisensor data fusion, pattern recognition, and artificial intelligence (AI) for material recycling.

Liang Xiao (Member, IEEE) received the B.S. degree in applied mathematics and the Ph.D. degree in computer science from the Nanjing University of Science and Technology (NJUST), Nanjing, China, in 1999 and 2004, respectively.

He has served as the Second Director of the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, NJUST, where he is currently a Professor with the School of Computer Science. From 2009 to 2010, he was a Post-Doctoral Fellow with the Rensselaer Polytechnic Institute, Troy, NY, USA. Since 2014, he has been the Deputy Director of the Jiangsu Key Laboratory of Spectral Imaging Intelligent Perception, Nanjing. He has published more than 100 international journal articles, including IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING. His main research interests include inverse problems in image processing, computer vision and image understanding, pattern recognition, and remote sensing.