remote sensing

MDPI

Article

Remote Sensing LiDAR and Hyperspectral Classification with Multi-Scale Graph Encoder–Decoder Network

Fang Wang [1], Xingqian Du [2], Weiguang Zhang [1], Liang Nie [1], Hu Wang [1,3,4,5], Shun Zhou [1] and Jun Ma [6,*]

[1] School of Optoelectronic Engineering, Xi'an Technological University, Xi'an 710021, China; wangfang86@xatu.edu.cn (F.W.); zhangweiguang@xatu.edu.cn (W.Z.); nieliang@xatu.edu.cn (L.N.); wanghu@opt.ac.cn (H.W.); zhoushun@xatu.edu.cn (S.Z.)
[2] China Academy of Space Technology (Xi'an), Xi'an 710100, China; duxingqian16@mails.ucas.edu.cn
[3] Xi'an Institute of Optics and Precision Mechanics of CAS, Xi'an 710119, China
[4] University of Chinese Academy of Sciences, Beijing 100049, China
[5] Xi'an Space Sensor Optical Technology Engineering Research Center, Xi'an 710119, China
[6] Institute for Interdisciplinary and Innovation Research, Xi'an Technological University, Xi'an 710021, China
* Correspondence: majun@xatu.edu.cn

Abstract: The rapid development of sensor technology has made multi-modal remote sensing data valuable for land cover classification due to its diverse and complementary information. Many feature extraction methods for multi-modal data, combining light detection and ranging (LiDAR) and hyperspectral imaging (HSI), have recognized the importance of incorporating multiple spatial scales. However, effectively capturing both long-range global correlations and short-range local features simultaneously on different scales remains a challenge, particularly in large-scale, complex ground scenes. To address this limitation, we propose a multi-scale graph encoder–decoder network (MGEN) for multi-modal data classification. The MGEN adopts a graph model that maintains global sample correlations to fuse multi-scale features, enabling simultaneous extraction of local and global information. The graph encoder maps multi-modal data from different scales to the graph space and completes feature extraction in the graph space. The graph decoder maps the features of multiple scales back to the original data space and completes multi-scale feature fusion and classification. Experimental results on three HSI-LiDAR datasets demonstrate that the proposed MGEN achieves considerable classification accuracies and outperforms state-of-the-art methods.

Keywords: LiDAR; hyperspectral image; multi-modal data fusion; graph model; deep learning

Citation: Wang, F.; Du, X.; Zhang, W.; Nie, L.; Wang, H.; Zhou, S.; Ma, J. Remote Sensing LiDAR and Hyperspectral Classification with Multi-Scale Graph Encoder–Decoder Network. *Remote Sens.* **2024**, *16*, 3912. https://doi.org/10.3390/rs16203912

Academic Editors: Ganchao Liu, Yaxiong Chen, Qingyu Li and Cong Zhang

Received: 10 September 2024
Revised: 15 October 2024
Accepted: 18 October 2024
Published: 21 October 2024

1. Introduction

Remote sensing image classification plays a significant role in fields like land cover monitoring and forest management [1,2]. The hyperspectral image (HSI) stands out for its exceptional ability to differentiate materials based on the unique spectral signatures [3,4]. However, in real-world applications, different objects may display similar or even identical spectral signatures due to environmental influences under certain conditions [5]. Conversely, the same type of object may exhibit varying spectral characteristics due to noise, temperature fluctuations, and other factors [6] during the imaging process [7]. As a result, relying solely on HSI for analysis poses certain limitations when confronted with complex ground details in large-scale scenes [8].

Facing the limitation of single-modality data, integrating multi-modal images has demonstrated significantly greater advantages across various research fields [9–11]. For instance, light detection and ranging (LiDAR) data [12], acquired by emitting laser beams and analyzing the reflected signals to determine precise distances, provides valuable altitude information over large areas through sensor scanning, which derives various tasks such as 3D instance segmentation [13] and point cloud segmentation [14]. Combining

LiDAR with HSI data collected from the same area is beneficial to partly overcome the limitations of single-modal information, substantially enhancing land cover classification accuracy [15,16]. Unlike natural images, remote sensing images often contain more intricate information, with variations in the positions of similar land cover features, making it crucial to extract global information from multi-modal remote sensing data [17]. In recent years, numerous scholars have explored this challenge using a variety of deep learning models. These models can be generally categorized into three types, including convolutional neural network (CNN)-based methods, transformer-based methods, and graph-based methods [18–21].

In the CNN-based methods, various strategies have been developed to capture global information from multi-modal data. For example, Song et al. [22] introduced a two-stream deep CNN in their hashing-based deep metric learning method to separately extract and then fuse spectral-spatial features, and enhanced classification accuracy with a unique loss function that incorporates both semantic and metric losses. Wang et al. [23] optimized traditional CNN by pyramid convolutions with different kernel sizes to extract features at different scales and subsequently used effective feature recalibration modules to enhance the multi-scale spatial-spectral features. Zhao et al. [24] proposed a hierarchical random walk network, which utilized the predicted distribution of dual-tunnel CNN to serve as a global prior on the fusion of HSI and LiDAR data and employed a pixel-wise affinity branch as pixel priors to enforce spatial consistency in the deeper layers of networks. Feng et al. [25] designed a dynamic scale hierarchical fusion network based on CNN, which dynamically selected and integrated features across scales to address the high-dimensional problem of multiscale features. It used spatial attention for shallow feature fusion and modal attention for deep fusion to improve classification accuracy. Wang et al. [26] introduced a nearest neighbor-based contrastive learning network (NNCNet) by introducing self-supervised contrastive learning and a bilinear attention fusion module to CNN-based joint classification for interpreting ground objects at a more precise level. Xue et al. [27] augmented the attentional feature fusion module on the basis of CNN, and a global average pooling layer was designed to enhance the representation of global information in features. Gao et al. [28] proposed an adaptive multiscale spatial–spectral enhancement network (AMSSE-Net) based on CNN to fuse features from HSI and LiDAR data to improve classification performance. With the property that the convolution kernel shares the feature channels within the group, the involution operator was introduced in the network to enhance the correlation of spectral dimensions. Besides, dynamically assigned weights were utilized to guide the selection of the optimal model, which is determined by the joint loss across the three feature fusion methods (maximum, adaptive addition and concat). Mohla et al. [29] devised deeper networks for multimodality features extraction, incorporating two spatial attention modules and one modality attention module. With a higher number of network layers, deeper convolutional layers can obtain larger receptive fields, corresponding to larger regions of the original image, achieving extensive feature perception. However, blindly stacking numerous convolutional layers may increase network depth and training difficulty, leading to a higher risk of overfitting.

Besides CNN, the transformer has attracted significant attention in computer vision due to its remarkable ability to model global relationships among samples in the visual domain [30,31]. In deep networks based on transformers, images often need to be serialized and input to the network in the form of image block sequences. In Ref. [32], a cross-modal enhanced CNN and transformer module was incorporated into a dual-branch feature fusion network to enhance interactive information from multi-source data both locally and globally, thereby enabling the robust integration of diverse features. Zhao et al. [33] proposed a novel dual-branch method combining a hierarchical CNN and a transformer network to enhance multisource data fusion and improve classification accuracy, with a cross-token attention fusion encoder that leverages CNN's spatial extraction capabilities and the transformer's long-range dependency modeling. Song et al. [34] designed a height information-guided hierarchical fusion-and-separation network, in which the transformer and CNN were introduced in the dual-structure feature encoders to encode the spectral

and spatial information, while deformable convolution-based modules were employed in feature fusion-and-separation blocks for modality-shared and modality-specific feature extraction. Yang et al. [35] selected HSI bands based on LiDAR data by introducing a cross-attention mechanism from the transformer architecture to reduce the redundancy of HSI and improve the classification accuracy. Zhang et al. [36] proposed a local information interaction transformer (LIIT) model to capture and fuse multi-modal data dynamically. A dual-branch transformer was designed in LIIT to fully extract the sequence features, and a local-based multisource feature interactor was developed to coordinate local spatial features with the global-based transformer. Ni et al. [37] introduced a multiscale head selection transformer (MHST) network to capture nonredundant features across multiple dimensions of data. An adaptive global feature extraction module was designed in MHST, which leveraged head selection pooling within the transformer to dynamically reduce token redundancy. Sun et al. [38] introduced a morphological feature enhancement module and a transformer-based deep dilated convolution module in the encoder enabling efficient integration of data features. Feng et al. [39] proposed a spectral-spatial-elevation fusion Transformer framework (S2EFT) adopting the Patch as the input of the transformer for taking full advantage of sequence data and spatial features. Additionally, Zhao et al. [40] used a CNN with residual connections to extract features from multi-modal data, then the features were serialized to execute further feature learning by integrating the transformer with Fourier transform, ultimately predicting the categories of land objects for classification tasks. While the transformer-based classification methods excel at extracting global features, images need to be serialized into image blocks to accommodate the structural characteristics of the transformer, resulting in the capture of global information to be converted into the extraction of associations between image blocks, which will potentially lead to insufficient extraction of local information within each image block.

Compared to CNNs, graph models inherently possess an advantage in modeling and extracting global information. In a topological graph, any two nodes can be connected by associating node features to establish edges, and the relationship between features among nodes is characterized by the weights of edges, thus overcoming the limitations of the two-dimensional structure of images. For example, Feng et al. [41] developed a multi-branch dual-channel graph convolutional network to remove the redundant information for integrating spectral–elevation–spatial features. Cai et al. [42] constructed an undirected weighted graph with modality-specific tokens in their multimodal fusion network to address the problem of long-distance dependencies. Wan et al. [43] segmented HSI into regions, with each segmented region serving as a node, to establish a complete topological graph by associating regions with each other. After that, they designed a dynamic graph convolutional network for feature learning, continuously updating the connection relationships between edges during training. Cai et al. [44] employed principal component analysis to reduce the HSI dimensionality. The principal components were input into CNN for feature extraction, and the features were taken as a series of graph nodes to establish a topological graph. Then the graph CNNs were subsequently employed for feature extraction, with cross-attention added to guide the features. Sha et al. [45] adopted graph attention networks for feature extraction from the topological graph constructed with HSI, ensuring that the features included both spatial and spectral characteristics. The aforementioned graph-based methods are effective in modeling global information, yet they face two main challenges. Firstly, these methods are still restricted to feature extraction within a single modality, without considering the fusion of global information from multiple modalities. Secondly, these methods still have limited capability in capturing local information. For instance, Sha et al. [45] directly converted pixels in images into the nodes of the topological graph, disregarding the local spatial information in the original images, while Wan et al. [43] used superpixel segmentation to retain certain spatial information which was highly sensitive to the scale of segmentation.

To address the mentioned issues, a multi-scale graph encoder–decoder network (MGEN) is proposed for multi-modal data classification. The MGEN is capable of ex-

tracting multi-modal image information at multiple scales, achieving local-global information fusion and robust feature representation. Specifically, MGEN consists of a graph encoder, graph feature extraction module and graph decoder, each module comprises three hierarchical levels of information.

In the graph encoder, unsupervised region segmentation is carried out on the images of two modalities through the segmentation algorithm, dividing spatial regions according to the semantic information of the images, and aggregating pixels from the original images to generate a series of superpixels. The superpixels are then transformed into the graph space, generating nodes and edges of the topological graph based on superpixel features and adjacency relationships, and the topological graphs with variety scales are generated by controlling the number of regions in superpixel segmentation. In the graph feature extraction module, different network branches are adopted to extract features from multi-scale topological graphs, incorporating the multiscale short- and long-range graph convolutional network (MSLGCN) [46] to extract features. In the graph decoder, the features are mapped from graph nodes back to the original pixels, and feature alignment is performed at the pixel scale, followed by multi-scale feature fusion. Finally, a classifier constructed from fully connected networks predicts pixel categories to obtain a category map.

The main contributions of this work are summarized as follows:

1. A multi-scale graph encoder–decoder network is proposed for the classification of remote sensing multi-modal data. This network is able to extract features from the graph space with multiple scales, achieving multi-level fusion of multi-modal global and local information.
2. Graph encoder and graph decoder are proposed for extracting modality-independent multi-scale features, while simultaneously measuring the direct similarity between short-range and long-range samples in multi-modal images to enhance feature discriminability.
3. Experiments on remote sensing multi-modal datasets are conducted, revealing that the proposed MGEN achieves comparable performance with state-of-the-art methods.

The remaining parts of the paper are organized as follows. Section 2 describes the proposed network in detail. The experimental results and analyses are shown in Section 3. Then, the effects of parameters in the network are discussed in Section 4. Finally, Section 5 summarizes some concluding remarks.

2. Methodology

This section details the proposed MGEN. Section 2.1 introduces the network architecture of MGEN, Section 2.2 introduces the graph encoder, Section 2.3 details the multi-scale graph features extraction, and Section 2.4 describes the graph decoder.

2.1. Network Architecture

The framework of the proposed MGEN in this paper is shown as Figure 1. The network performs feature extraction within three different scales, each of which involves the fusion process of HSI and LiDAR data. During the feature extraction process at each scale, images of HSI and LiDAR are mapped into graph space with a variety of scale transformations. Then, graph convolutional networks are employed to extract features from the encoded topological graphs, and a graph decoder is adopted to reconstruct the features consistent with the original data size. Finally, the reconstructed multi-scale features are fused to generate multi-modal features which are subsequently inputted into a classifier to derive a class map.

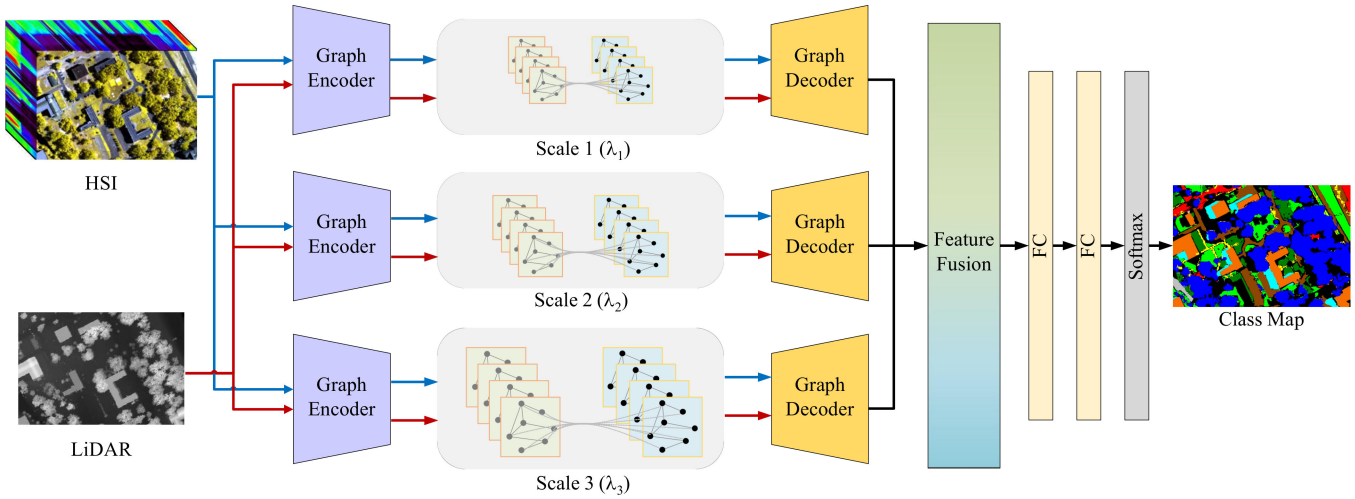

Figure 1. Overall framework of the proposed MGEN for multi-modal data classification.

2.2. Graph Encoder

The structure of the graph encoder in the MGEN is illustrated in Figure 2. The HSI and LiDAR images are represented as $\mathbf{X}^{\mathrm{H}} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times C}$ and $\mathbf{X}^{\mathrm{L}} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times 1}$, respectively, where H and W denote the pixel height and width of the images, and C represents the number of spectral bands in the HSI. To preserve the spatial neighborhood information of the original images, unlike some existing methods that directly construct topological graphs from images, we perform unsupervised image segmentation to divide images into multiple adjacent regions, and each region serves as a node in the topological graph. Considering the excellent performance of simple linear iterative clustering (SLIC) [47] in unsupervised segmentation, SLIC is adopted here to perform region segmentation on HSI and LiDAR images to generate superpixels. The number of superpixels, denoted as n, can be expressed as follows:

$$n = \left\lceil \frac{\mathrm{H}\times\mathrm{W}}{\lambda} \right\rceil \tag{1}$$

where λ is the scale parameter for controlling the number of superpixels. The λ in three scales are denoted with λ_1, λ_2, and λ_3. Let $S = \{S_i\}_{i=1}^{n}$ denote the set of all superpixels in the image, where $S_i = \left\{\mathbf{x}_j^i\right\}_{j=1}^{N_i}$ is the ith superpixel, $\mathbf{x}_j^i$ is the jth original pixel in S_i, and N_i is the number of original pixels contained in the superpixel. Relation matrices $\mathbf{Q}^{\mathrm{H}} \in \mathbb{R}^{n\times C}$ and $\mathbf{Q}^{\mathrm{L}} \in \mathbb{R}^{1\times C}$ are accordingly generated by SLIC, where each element in $\mathbf{Q}^{\mathrm{H}}$ and $\mathbf{Q}^{\mathrm{L}}$ records the assignment of each pixel in the image to a certain suprepixel. Specifically, each row in $\mathbf{Q}^{\mathrm{H}}$ corresponds to a origin HSI pixel, each column corresponds to a superpixel generated by SLIC. When the ith pixel is assigned to the jth superpixel, $\mathbf{Q}_{ij}^{\mathrm{H}} = 1$, otherwise $\mathbf{Q}_{ij}^{\mathrm{H}} = 0$. The elements in $\mathbf{Q}^{\mathrm{L}}$ are determined similarly.

Then, we transform all superpixels into nodes of topological graph, each node corresponds to a superpixel S_i. To complete the construction of the topological graph, it is necessary to obtain the feature matrices $\mathbf{V}^{\mathrm{H}}$ and $\mathbf{V}^{\mathrm{L}}$ for HSI and LiDAR, respectively, as well as the adjacency matrices $\mathbf{A}^{\mathrm{H}}$ and $\mathbf{A}^{\mathrm{L}}$ representing the connectivity between nodes.

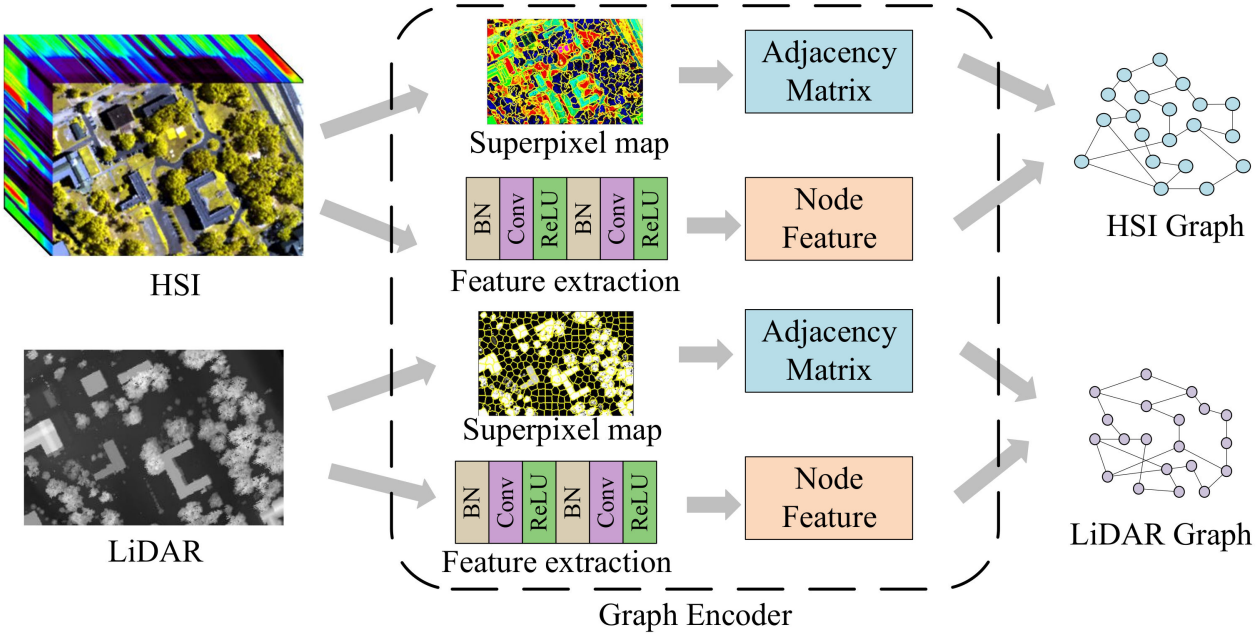

Figure 2. Structure of the graph encoder.

2.2.1. Node Feature Extraction

The original data from HSI and LiDAR may contain redundant information and noise, which negatively influence feature extraction. To mitigate these influence factors, we preprocess the pixels by a 1×1 CNN. Then, according to the results of superpixel segmentation, pixel information is further integrated into node features of the topological graph. Specifically, the output of the lth convolutional layer in the node feature network structure is as follows:

$$\mathbf{X}^{(l)} = \sigma\left(\mathbf{W}^{(l)} * \mathrm{BN}\left(\mathbf{X}^{(l-1)}\right) + \mathbf{b}^{(l)}\right) \tag{2}$$

where $*$ denotes the convolution operator, $\mathbf{X}^{(l-1)}$ is the input to the lth layer, $\mathrm{BN}(\cdot)$ is batch normalization operation, $\sigma(\cdot)$ is the activation function, and $\mathbf{W}^{(l)}$ and $\mathbf{b}^{(l)}$, respectively, denote the learnable parameters and biases of this convolutional layer. Since the convolutional kernel size is 1×1, the size of the network output remains the same as the input after the convolution operation. The $\mathbf{X}^{(l)} = \{\widetilde{\mathbf{x}}_i\}$ from Equation (2) is still a pixel-level feature, where $\widetilde{\mathbf{x}}_i$ represents the feature of the ith pixel rather than the node-level features of the topological graph. To achieve transformation of the feature levels while preserving the spatial information of the original image, feature aggregation based on superpixels is required. If $\mathbf{V}_k$ represents the feature of the kth node, the feature aggregation method is shown as follows:

$$\mathbf{V}_k = \frac{1}{N_k} \sum_{j=1}^{N_k} \widetilde{\mathbf{x}}_j \tag{3}$$

where N_k represents the number of pixels in the superpixel corresponding to node $\mathbf{V}_k$. The feature aggregation method with average value helps alleviate the impact of outlier pixels with inaccurate segmentation. Combining all node feature vectors yields the node feature matrix $\mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_k, \cdots, \mathbf{V}_m]$, where m is the number of nodes. Thereby, node feature matrices $\mathbf{V}^{\mathrm{H}}$ and $\mathbf{V}^{\mathrm{L}}$ for HSI and LiDAR data can be obtained through the process above.

2.2.2. Adjacency Matrices

After defining the node features $\mathbf{V}$, it is also necessary to establish the connectivity of edges according to the relationship between nodes, generating adjacency matrix $\mathbf{A}$. To

maximize the preservation of spatial information contained in the original image, we utilize the adjacency between superpixels to determine $\mathbf{A}$. Generating an adjacency matrix based on spectral similarity is a common way for hyperspectral image; however, calculating vertex-wise similarity imposes a significant computational burden. Therefore, the proposed MGEN constructs the adjacency matrix using spatial relationships within the image. In other words, the edge between every two spatially adjacent superpixels is weighted to 1, while the weight of all other edges is set to 0, indicating no edge adjacency. Thus, the adjacency matrix can be expressed as follows:

$$\mathbf{A}_{ij} = \begin{cases} \mathbf{1}, & \text{if } S_i \text{ and } S_j \text{ are adjacent} \\ \mathbf{0}, & \text{otherwise.} \end{cases} \tag{4}$$

where $\mathbf{A}_{ij}$ is the element located at (i, j) in the adjacency matrix $\mathbf{A}$. With the method above, we can obtain the adjacency matrices $\mathbf{A}^{\mathrm{H}}$ and $\mathbf{A}^{\mathrm{L}}$ for HSI and LiDAR data, respectively.

2.3. Multi-Scale Graph Features Extraction

To extract multi-modal data features from different scales, a multi-branch graph convolutional network is adopted for multi-scale feature extraction after graph encoding. For each branch, the same depth network is used for feature extraction. The process of feature extraction is elaborated below.

2.3.1. Graph Convolution

After obtaining the node feature matrix $\mathbf{V}$ and adjacency matrix $\mathbf{A}$ of the topological graph, graph neural networks can be employed for feature extraction. We can obtain the degree matrix $\mathbf{D}$ based on the adjacency matrix $\mathbf{A}$. The degree matrix $\mathbf{D}$ is a diagonal matrix, and its elements on the diagonal can be calculated as $\mathbf{D}_{ii} = \sum_j \mathbf{A}_{ij}$. Thereon, the Laplacian matrix of the topological graph is derived as follows:

$$\mathbf{L} = \mathbf{D} - \mathbf{A} \tag{5}$$

The Laplacian matrix contains crucial information in the topological graph, thus allowing the transformation from graph analysis problems into Laplacian matrix analysis problems. The Laplacian matrix is normalized as $\mathbf{L}_{norm} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$ [48], which is a real symmetric positive semidefinite matrix and can be factorized as follows:

$$\mathbf{L}_{norm} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \tag{6}$$

where $\mathbf{U} \in \mathbb{R}^{n \times n}$ and $\mathbf{\Lambda} \in \mathbb{R}^{n \times n}$ represent the matrix of eigenvectors and the diagonal matrix of eigenvalues, respectively. All eigenvectors are mutually orthogonal, and the elements λ_i on the diagonal of $\mathbf{\Lambda}$ are the eigenvalues. So the convolutional operations on the graph can be defined as follows:

$$\mathbf{V}^{(l+1)} = \mathbf{U}\mathbf{g}_\theta\mathbf{U}^T\mathbf{V}^{(l)} \tag{7}$$

where $\mathbf{V}^{(l)}$ is the input node features, $\mathbf{V}^{(l+1)}$ is the output node features, and diagonal matrix $\mathbf{g}_\theta$ represents the parameters to be learned. To reduce computational complexity during convolution $\mathbf{g}_\theta$ can be considered as a function of $\mathbf{\Lambda}$ [49], and the Kth order Chebyshev polynomial $T_K(\cdot)$ is used to approximate $\mathbf{g}_\theta$ as follows:

$$\mathbf{g}_\theta(\mathbf{\Lambda}) \approx \sum_{k=0}^{K} \theta_k T_K\left(\widetilde{\mathbf{\Lambda}}\right) \tag{8}$$

where $\widetilde{\mathbf{\Lambda}} = \frac{2\mathbf{\Lambda}}{\lambda_{\max}} - \mathbf{I}$. Generally, it is enough to take $K = 1$. As $T_0\left(\widetilde{\mathbf{\Lambda}}\right) = 1$, $T_1\left(\widetilde{\mathbf{\Lambda}}\right) = \widetilde{\mathbf{\Lambda}}$, it could be the following:

$$\mathbf{g}_\theta(\mathbf{\Lambda}) = \theta_0 + \theta_1\widetilde{\mathbf{\Lambda}} \tag{9}$$

If $\lambda_{\max} \approx 2$, $\theta = \theta_0 = -\theta_1$, the convolutional operations in Equation (7) are redefined as follows:

$$\mathbf{V}^{(l+1)} = \mathbf{U}\left(\theta_0 + \theta_1\widetilde{\mathbf{\Lambda}}\right)\mathbf{U}^T\mathbf{V}^{(l)} = \theta\left(\mathbf{I} + \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}\right)\mathbf{V}^{(l)} \tag{10}$$

2.3.2. Long-Range and Short-Range Attention Graph Convolution Module

Considering the large sizes of the multi-modal images will make it difficult to simultaneously take into account the relationships between distant nodes and nearby nodes, we adopt here the MSLGCN proposed by Zhu et al. [46] to cope with the adjacency information over different ranges. An attention mechanism is incorporated into the graph convolution process by MSLGCN to separately extract relations of long-range nodes and short-range nodes in a targeted manner. In the network, the convolution operation consists of two sub-modules to extract local and global features, each of which employs different adjacency matrices for the relations of long-range nodes and short-range nodes to achieve distinct feature learnings.

An attention matrix $\mathbf{M}$ is defined to characterize the range relations between nodes in the topological graph. If the feature of the ith node is represented as $\mathbf{V}(i)$, the element located at (i, j) in the attention matrix $\mathbf{M}$ is as follows:

$$\mathbf{M}(i,j) = \frac{1}{1 + \mathrm{e}^{-\mathrm{FC}(\mathbf{V}(i))\cdot\mathrm{FC}(\mathbf{V}(j))}} \tag{11}$$

where FC is a fully connected network, $\cdot$ is matrix multiplication. The values of $\mathbf{M}(i,j)$ indicate the similarity between the ith and jth nodes. Based on the attention matrix $\mathbf{M}$, the adjacency matrices $\widetilde{\mathbf{A}}_l$ for representing long-range node relations and $\widetilde{\mathbf{A}}_s$ for short-range node relations can be expressed as follows:

$$\begin{aligned} \widetilde{\mathbf{A}}_l(i,j) &= \begin{cases} \mathbf{M}(i,j), & i \neq j \\ \mathbf{M}(i,j) + 1, & i = j \end{cases} \\ \widetilde{\mathbf{A}}_s(i,j) &= \begin{cases} \mathbf{M}(i,j)\cdot(\mathbf{A}(i,j) + \mathbf{I}(i,j)), & i \neq j \\ \mathbf{M}(i,j)\cdot(\mathbf{A}(i,j) + \mathbf{I}(i,j)) + 1, & i = j \end{cases} \end{aligned} \tag{12}$$

Note that self-connection operation is added to the adjacency matrix by $\mathbf{A} + \mathbf{I}$ here to preserve the own features of the node in convolution. Correspondingly, the self-connection degree matrix $\widetilde{\mathbf{D}}$ has $\widetilde{\mathbf{D}}(i,i) = \sum_j(\mathbf{A}(i,j) + \mathbf{I}(i,j))$. From the definition of adjacency matrix above, we obtain the attention weights with long-range and short-range as $\mathbf{C}_l$ and $\mathbf{C}_s$:

$$\begin{aligned} \mathbf{C}_l &= \widetilde{\mathbf{D}}^{-1/2}\widetilde{\mathbf{A}}_l\widetilde{\mathbf{D}}^{-1/2} + \mathbf{I} \\ \mathbf{C}_s &= \widetilde{\mathbf{D}}^{-1/2}\widetilde{\mathbf{A}}_s\widetilde{\mathbf{D}}^{-1/2} + \mathbf{I} \end{aligned} \tag{13}$$

According to the analysis above and combining the results of long-range and short-range feature extraction, the propagation rule for the final graph convolutional network layer is as follows:

$$\mathbf{V}^{(l+1)} = \mathrm{FC}\left[\sigma\left(\mathbf{C}_l\mathrm{FC}\left(\mathbf{V}^{(l)}\right)\right), \sigma\left(\mathbf{C}_s\mathrm{FC}\left(\mathbf{V}^{(l)}\right)\right)\right] \tag{14}$$

where $\mathrm{FC}(\cdot)$ denotes a fully connected layer, $[\cdot,\cdot]$ represents feature concatenation operation, and $\sigma(\cdot)$ is the activation function.

The network consists of branches in three scales, each of which yields features $\mathbf{V}_{\mathrm{H}}^{(l+1)}$ and $\mathbf{V}_{\mathrm{L}}^{(l+1)}$ for both HSI and LiDAR images.

2.4. Graph Decoder

After completing the multi-scale feature extraction, fusing the features of every scale is demanded. Furthermore, aiming at the final class map, the features are mapped back

to the pixel level to obtain pixel-level features and the class information of each pixel. Therefore, we propose a graph decoder to achieve multi-scale multi-modal feature fusion and pixel-level feature mapping.

Here, we let $\mathbf{V}_{\mathrm{H1}}^{(l+1)}, \mathbf{V}_{\mathrm{H2}}^{(l+1)}, \mathbf{V}_{\mathrm{H3}}^{(l+1)}$ represent the HSI features at three scales, while $\mathbf{V}_{\mathrm{L1}}^{(l+1)}, \mathbf{V}_{\mathrm{L2}}^{(l+1)}, \mathbf{V}_{\mathrm{L3}}^{(l+1)}$ represent the LiDAR features. Each matrix contains the node features that have been segmented into superpixels, which need to be mapped back to the original image size, denoted as $\mathbf{P}_{\mathrm{H1}}, \mathbf{P}_{\mathrm{L1}}, \mathbf{P}_{\mathrm{H2}}, \mathbf{P}_{\mathrm{L2}}, \mathbf{P}_{\mathrm{H3}}$ and $\mathbf{P}_{\mathrm{L3}}$. The correspondence between pixel and superpixel is recorded in relation matrices $\mathbf{Q}^{\mathrm{H}}$ and $\mathbf{Q}^{\mathrm{L}}$. Therefore, mapping superpixel features back to the original image size can be operated by

$$\begin{aligned} \mathbf{P}_{\mathrm{H1}} &= \mathbf{Q}^{\mathrm{H1}}\mathbf{V}_{\mathrm{H1}}^{(l+1)} \\ \mathbf{P}_{\mathrm{L1}} &= \mathbf{Q}^{\mathrm{L1}}\mathbf{V}_{\mathrm{L1}}^{(l+1)} \\ \mathbf{P}_{\mathrm{H2}} &= \mathbf{Q}^{\mathrm{H2}}\mathbf{V}_{\mathrm{H2}}^{(l+1)} \\ \mathbf{P}_{\mathrm{L2}} &= \mathbf{Q}^{\mathrm{L2}}\mathbf{V}_{\mathrm{L2}}^{(l+1)} \\ \mathbf{P}_{\mathrm{H3}} &= \mathbf{Q}^{\mathrm{H3}}\mathbf{V}_{\mathrm{H3}}^{(l+1)} \\ \mathbf{P}_{\mathrm{L3}} &= \mathbf{Q}^{\mathrm{L3}}\mathbf{V}_{\mathrm{L3}}^{(l+1)} \end{aligned} \tag{15}$$

where $\mathbf{Q}^{\mathrm{H1}}$ and $\mathbf{Q}^{\mathrm{L1}}$ denote the relation matrices corresponding to scale 1. The features of other scales 2 and scale 3 are generated in a similar manner. When the features from all three scales are generated, the fused feature $\mathbf{F}$ is defined by

$$\mathbf{F} = \mathrm{FC}(\mathrm{FC}([\mathbf{P}_{\mathrm{H1}} \otimes \mathbf{P}_{\mathrm{L1}}, \mathbf{P}_{\mathrm{H2}} \otimes \mathbf{P}_{\mathrm{L2}}, \mathbf{P}_{\mathrm{H3}} \otimes \mathbf{P}_{\mathrm{L3}}])) \tag{16}$$

where $[\cdot,\cdot,\cdot]$ is the concatenation operation of multiple features, and $\otimes$ is the multiplication by elements. The function $\mathrm{FC}(\cdot)$ denotes a fully connected layer. After feature fusion, the final features containing multiple scales are processed by the softmax function to predict the class for each pixel, which is defined by

$$\mathbf{Y} = \frac{e^{\mathbf{F}_i}}{\sum_{j=1}^{c} e^{\mathbf{F}_j}} \tag{17}$$

where $\mathbf{Y}$ is the class vectors output from the network. $\mathbf{F}_i$ and $\mathbf{F}_j$ denotes the ith and jth element of $\mathbf{F}$, respectively. c is the number of classes in the dataset. The softmax function defined by Equation (17) and the two fully connected layers in Equation (16) make up a typical softmax classifier [50]. The loss function of the network adopts the cross-entropy loss, which is commonly used in classification tasks.

3. Results

3.1. Datasets

To verify the effectiveness of the proposed MGEN in multi-modal data classification, we conducted experiments on existing remote-sensing HSI and LiDAR datasets. Three datasets are encompassed in the experiments to ensure data diversity, namely the Trento dataset [51], Missouri University and University of Florida dataset (MUUFL) [52,53], and the Houston dataset [54]. Detailed information about the datasets is presented below.

3.1.1. Trento Dataset

The Trento dataset was captured in rural areas surrounding the city of Trento, Italy, containing HSI and LiDAR images. The HSI consists of 600 × 166 pixels and includes 63 spectral bands ranging from 420.89 nm to 989.09 nm, with a spectral resolution of 9.2 nm and spatial resolution of 1 m. The LiDAR is a single-channel image containing altitudes corresponding to the ground positions, with the same dimensions as the HSI. The annotated information in the dataset consists of six classes. Approximately 10% of the labeled pixels

are used for training, while the remaining 90% are used for testing. Detailed information about the dataset and the number of samples for each class is presented in Table 1.

Table 1. The number of training and testing samples for each class on Trento dataset.

No.	Class	Train	Test	All
1	Apple trees	404	3630	4034
2	Buildings	291	2612	2903
3	Ground	48	431	479
4	Woods	913	8210	9123
5	Vineyard	1051	9450	10,501
6	Roads	318	2856	3174
Total		3025	27,189	30,214

3.1.2. MUUFL Dataset

The MUUFL dataset is a co-registered aerial HSI-LiDAR dataset. The two modal images in this dataset were acquired simultaneously during an aerial flight in November 2010, located in Mississippi, USA. The image dimensions are 325 × 220 pixels. The HSI contains 64 spectral bands, while the LiDAR image is a single-channel image with altitude information. The labeled pixels in the dataset include 11 classes. We randomly selected 10% of annotated pixels for training, while the remaining annotated pixels were used for testing. Further details about this dataset are provided in Table 2.

Table 2. The number of training and testing samples for each class on MUUFL dataset.

No.	Class	Train	Test	All
1	Trees	2325	20,921	23,246
2	Grass ground	427	3843	4270
3	Mixed ground	689	6193	6882
4	Dirt and sand	183	1643	1826
5	Road	669	6018	6687
6	Water	47	419	466
7	Buildings	224	2009	2233
8	Shadow	624	5616	6240
9	Sidewalk	139	1246	1385
10	Yellow curb	19	164	183
11	Cloth panels	27	242	269
Total		5373	48,314	53,687

3.1.3. Houston Dataset

The Houston dataset originates from the Data Fusion Contest initiated by the IEEE Geoscience and Remote Sensing Society (GRSS) in 2013. The HSI and LiDAR in this dataset were captured around the University of Houston and its surrounding neighborhoods with a spatial resolution of 2.5 m and a dimension of 349 × 1905 pixels. The HSI consists of 144 spectral bands covering a wavelength range from 380 to 1050 nm, and the single-channel LiDAR image indicates the altitude on the corresponding position. The annotation information in the dataset includes 15 classes. The number of training and testing samples for each class, along with other detailed information, are provided in Table 3.

Table 3. The number of training and testing samples for each class on Houston dataset.

No.	Class	Train	Test	All
1	Healthy grass	126	1125	1251
2	Stressed grass	126	1128	1254
3	Synthetic grass	70	627	697
4	Trees	125	1119	1244
5	Soil	125	1117	1242
6	Water	33	292	325
7	Residential	127	1141	1268

Table 3. *Cont.*

No.	Class	Train	Test	All
8	Commericial	125	1119	1244
9	Road	126	1126	1252
10	Highway	123	1104	1227
11	Railway	124	1111	1235
12	Parking lot 1	124	1109	1233
13	Parking lot 2	47	422	469
14	Tennis court	43	385	428
15	Running Track	66	594	660
Total		1510	13,519	15,029

3.2. Experimental Settings

The experimental environment is based on the Linux Ubuntu operating system. The PyTorch 1.7.0 deep learning framework with Python 3.6 is adopted to construct the network. The learning rate is set to 10^{-4} and the maximum number of epochs is set to 500 in training. Three scales in MGEN are set as $\lambda_1 = 100$, $\lambda_2 = 120$, and $\lambda_3 = 70$, which will be detailed in the parameter analysis. In terms of evaluation metrics, besides the classification accuracy for each class, we also utilize Overall Accuracy (OA), Average Accuracy (AA) and Kappa coefficient. OA represents the proportion of correctly classified samples to the total number of samples, while AA indicates the average of the classification accuracies for each class, and the Kappa coefficient measures the agreement between predicted and actual classifications.

3.3. Comparative Methods

In order to assess the performance of the proposed MGEN, various existing methods are selected as compared methods in the experiments, including Support Vector Machine (SVM), Laplacian Embedding (LE), Spatial Spectral Schrodinger Eigenmap (SSSE) [55], Contextual CNN (CCNN) [56], 3D CNN [57], Two-Branch CNN (TBCNN) [58], and Hierarchical Random Walk Network (HRWN) [59]. The characteristics of these methods are outlined below:

- SVM: This is a widely used supervised learning algorithm that maps image data into a high-dimensional feature space and identifies the optimal boundary to maximize separation between different classes.
- LE: It classifies hyperspectral images by reducing dimensionality and preserving local structure to better reveal relationships between pixels.
- SSSE: It combines Laplacian embedding and Schrodinger eigenmaps to extract spectral features and maintain important contextual details.
- CCNN: This method integrates spectral data with contextual information, allowing it to capture local relationships and preserve spatial structures.
- 3D CNN: The method processes multi-band data by directly performing 3D convolution operations, collecting spatial and spectral information simultaneously.
- TBCNN: The method harnesses image information by using two parallel convolutional branches to separately extract spectral and spatial features.
- HRWN: It utilizes a hierarchical structure combined with a random walk algorithm to incorporate local and global relationships, enhancing both spatial coherence and classification accuracy.

3.4. Experimental Results and Analysis

The comparative experimental results are exhibited in Tables 4–6 and Figures 3–5. The tables present the statistical accuracy of each class and three accuracy metrics. The figures display the class maps plotted in various colors, enabling a visual comparison of each method's classification results with the ground truth, which highlights the accuracy of the methods.

Table 4. Classification results of the proposed MGEN and other compared methods on Trento dataset.

Class No.	SVM	LE	SSSE	CCNN	3D CNN	TBCNN	HRWN	MGEN
1	81.90%	60.22%	**100.00%**	**100.00%**	99.97%	53.83%	99.19%	**100.00%** *
2	96.94%	98.12%	99.08%	94.80%	96.53%	98.53%	93.79%	**98.74%**
3	96.29%	93.04%	97.68%	96.06%	90.49%	98.73%	97.00%	**99.07%**
4	99.67%	99.00%	99.84%	**100.00%**	**100.00%**	**100.00%**	99.91%	**100.00%**
5	94.70%	89.96%	99.92%	99.95%	**100.00%**	99.95%	98.68%	**100.00%**
6	94.78%	90.76%	97.06%	97.13%	98.73%	**99.86%**	98.70%	98.53%
OA	94.74%	89.64%	99.49%	99.12%	99.38%	88.36%	98.60%	**99.71%**
AA	94.05%	88.52%	98.93%	97.99%	97.62%	91.82%	97.88%	**99.39%**
Kappa	92.97%	86.09%	99.32%	98.82%	99.17%	84.95%	98.13%	**99.61%**

* The maximum accuracy of each category is displayed in bold. The same mark is applied to the subsequent tables as well.

Table 5. Classification results of the proposed MGEN and other compared methods on MUUFL dataset.

Class No.	SVM	LE	SSSE	CCNN	3D CNN	TBCNN	HRWN	MGEN
1	95.21%	92.58%	**98.03%**	97.81%	96.54%	97.16%	96.79%	97.21%
2	64.87%	71.95%	93.08%	83.35%	**99.15%**	85.71%	69.81%	95.23%
3	83.87%	78.14%	93.25%	90.02%	46.90%	81.15%	**95.14%**	93.96%
4	71.09%	74.13%	91.66%	81.86%	85.80%	92.25%	93.21%	**97.71%**
5	91.86%	90.13%	96.91%	**94.84%**	94.42%	90.00%	93.79%	91.24%
6	91.41%	73.51%	95.23%	87.83%	89.31%	98.33%	95.98%	**96.76%**
7	43.40%	67.99%	91.94%	91.39%	92.19%	94.57%	84.80%	**96.03%**
8	92.82%	79.31%	96.97%	96.96%	94.48%	97.25%	**98.57%**	96.49%
9	18.38%	45.02%	58.67%	72.02%	76.71%	**85.54%**	84.86%	84.32%
10	78.05%	3.66%	1.22%	32.73%	25.63%	56.67%	**85.37%**	71.30%
11	66.53%	87.19%	**94.21%**	83.88%	62.40%	95.93%	88.65%	98.75%
OA	85.46%	83.87%	94.90%	93.35%	88.48%	92.56%	92.60%	**95.36%**
AA	72.50%	69.42%	82.83%	82.97%	78.50%	88.60%	89.73%	**92.64%**
Kappa	80.61%	78.69%	93.24%	91.18%	84.85%	90.14%	91.89%	**93.89%**

Table 6. Classification results of the proposed MGEN and other compared methods on the Houston dataset.

Class No.	SVM	LE	SSSE	CCNN	3D CNN	TBCNN	HRWN	MGEN
1	93.33%	90.40%	95.47%	73.45%	86.23%	88.85%	91.09%	**99.73%**
2	98.05%	95.57%	93.97%	94.77%	**99.91%**	96.50%	95.65%	98.58%
3	98.72%	98.41%	99.52%	99.52%	97.45%	**100.00%**	99.86%	**100.00%**
4	99.20%	94.92%	96.69%	99.91%	98.13%	99.55%	**100.00%**	**100.00%**
5	99.55%	95.40%	**100.00%**	94.90%	**100.00%**	99.64%	98.03%	**100.00%**
6	90.75%	85.27%	97.95%	88.01%	88.36%	99.59%	**100.00%**	97.94%
7	91.59%	83.51%	97.46%	97.63%	79.49%	97.46%	96.74%	**97.72%**
8	91.96%	89.31%	94.10%	86.52%	96.25%	**100.00%**	98.29%	98.03%
9	84.55%	69.16%	**93.43%**	64.51%	56.08%	88.13%	97.11%	90.48%
10	94.02%	53.24%	98.19%	99.37%	75.54%	89.72%	84.00%	**100.00%**
11	90.82%	77.77%	**100.00%**	98.65%	83.72%	91.21%	93.33%	**100.00%**
12	85.39%	54.40%	89.81%	73.51%	93.15%	76.66%	79.89%	**98.92%**
13	23.46%	15.60%	94.79%	44.08%	50.95%	**100.00%**	92.76%	99.29%
14	97.40%	90.95%	**100.00%**	92.73%	96.62%	**100.00%**	97.42%	**100.00%**
15	97.81%	92.46%	**100.00%**	99.83%	99.49%	**100.00%**	95.51%	99.16%
OA	91.26%	79.53%	96.38%	88.08%	87.08%	93.29%	93.50%	**98.52%**
AA	89.11%	79.09%	96.76%	87.16%	86.76%	95.15%	94.65%	**98.66%**
Kappa	90.53%	77.81%	96.09%	87.09%	86.02%	92.73%	92.97%	**98.39%**

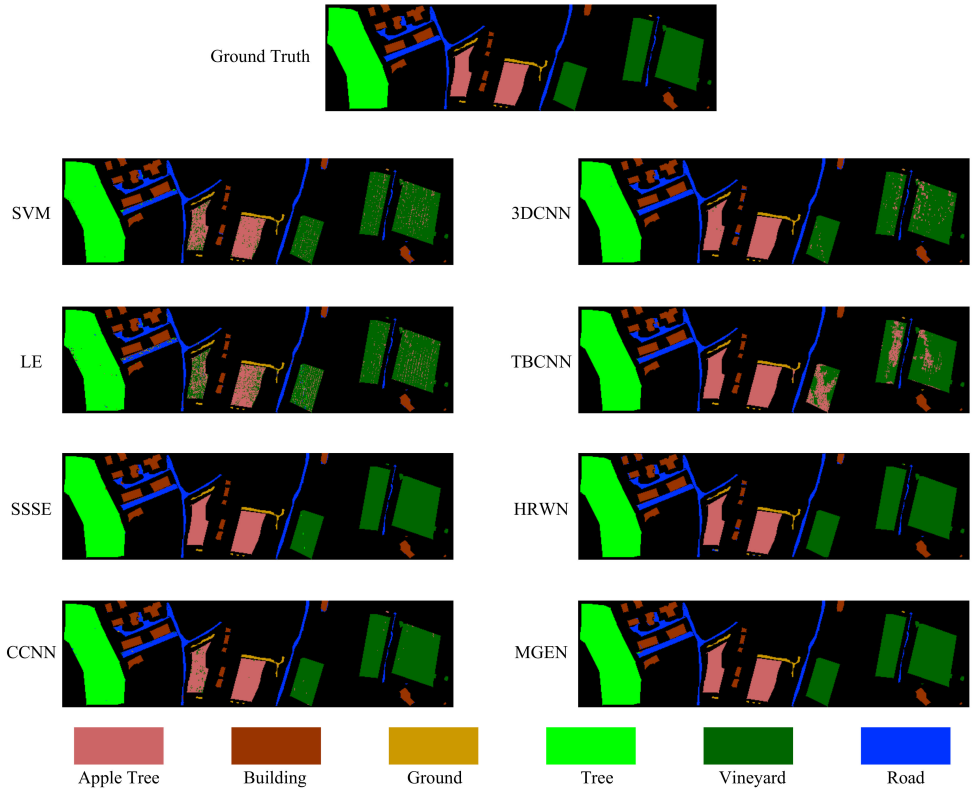

Figure 3. Visualized classification results of proposed MGEN and compared methods on Trento dataset.

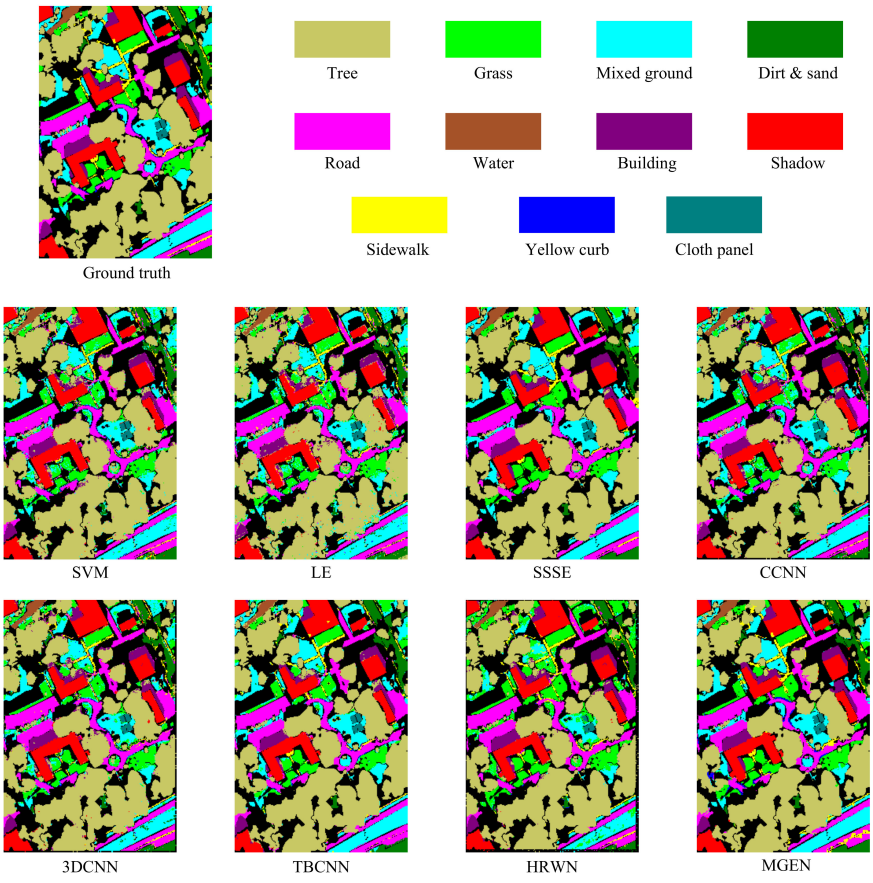

Figure 4. Visualized classification results of proposed MGEN and compared methods on MUUFL dataset.

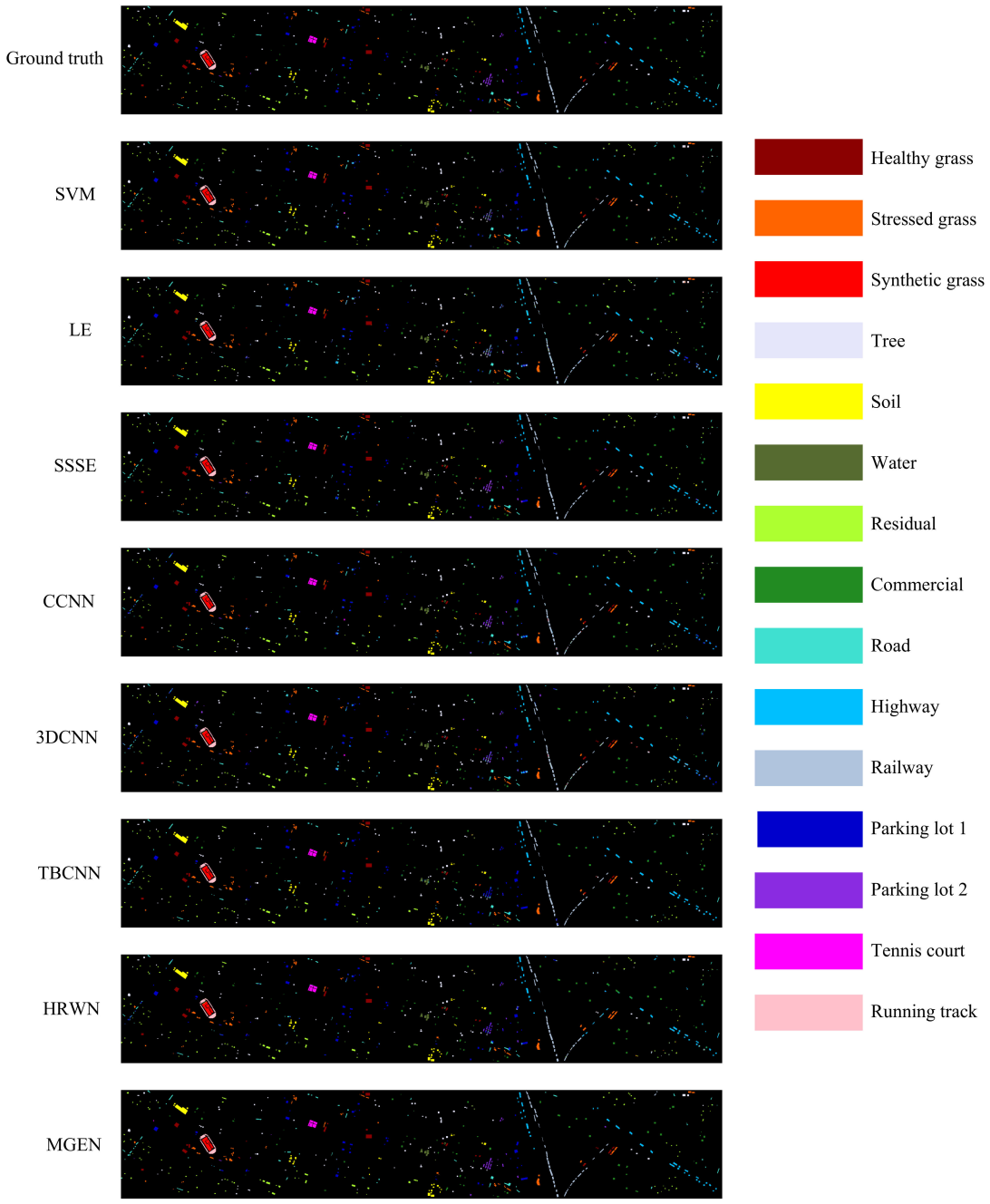

Figure 5. Visualized classification results of proposed MGEN and compared methods on Houston dataset.

Taking Trento as the tested dataset, the experimental results of the proposed MGEN and other compared methods are shown in Table 4. It can be seen that our MGEN achieves the highest accuracy among all methods for three metrics: OA at 99.71%, AA at 99.39%, and Kappa at 99.61%. Among the six classes, MGEN achieves the highest accuracy in each of first five classes, namely Apple trees, Buildings, Ground, Woods, and Vineyards. In the sixth class, Roads, the accuracy of MGEN (98.53%) is slightly lower than that of TBCNN (99.86%). SSSE achieves the best classification performance among the compared methods, second only to ours, with an OA of 99.49%, AA of 98.93%, and Kappa of 99.32%.

On the Trento dataset, the visualized classification results of each method are shown in Figure 3. It can be seen that the class maps predicted by the proposed MGEN are closest to the ground truth, with clearer contours of each area and fewer isolated misclassified pixels. Specifically, for the classes “Apple Tree”, “Tree”, and “Vineyard”, which have densely clustered sample distributions, the MGEN attains a fully unerring classification, showcasing

the network's prominent ability to capture fine local features. Meanwhile, for classes with more dispersed spatial distributions and larger positional spans, such as "Building" and "Ground", our proposed method still achieves almost perfect recognition, which can be attributed to its strong capability to integrate global information. Overall, MGEN excels in extracting both local and global features, leading to improved classification performance.

The classification results of all methods on the MUUFL dataset are shown in Table 5. We can see that our proposed MGEN achieves favorable results with an OA of 95.36%, an AA of 92.64%, and a Kappa of 93.89%. The dataset comprises a total of 11 classes, and our MGEN reaches the highest accuracy among all methods in three classes: Dirt and sand, Water, and Buildings. Compared to the baseline method SVM, MGEN improves OA by 9.9%, AA by 20.14%, and Kappa by 13.28%. It is noteworthy that many methods exhibit remarkably lower accuracy in the 10th class compared to others. This discrepancy might be due to the limited number of annotated pixels in this class, with only 19 pixels available for training after partitioning the dataset, significantly fewer than other classes. Despite such conditions, MGEN achieves a relatively high single-class prediction accuracy (71.30%), second only to the HRWN method.

Figure 4 illustrates the visualized class predictions of all methods on the MUUFL dataset. It can be observed that the predicted class map of the proposed MGEN closely resembles the ground truth, with relatively pure color blocks for each class and a small number of misclassified pixels. Particularly, the edges of categories like "Building" are accurately delineated in the subplot of MGEN, displaying much clearer contours. On the contrary, the boundaries generated by comparative methods appear to lack smoothness and exhibit some scattered misclassified pixels within the regions. This reveals that our MGEN outperforms other classification methods in fine feature representation and extraction. It effectively captures intricate details in complex scenes, leading to more precise classification.

On the Houston dataset, the classification results of the MGEN and other compared methods are presented in Table 6, demonstrating that our proposed network achieves high-level accuracy as well. The OA reaches 98.52%, the AA reaches 98.66%, and the Kappa is 98.39%, all of which are the highest among all the methods. In the total of 15 classes, the proposed MGEN results the best accuracy in nine classes individually (Healthy grass, Synthetic grass, Trees, Soil, Residential, Highway, Railway, Parking lot 1, Tennis court), indicating an excellent discriminative ability of MGEN across various types of land cover. Better than other compared methods, the SSSE method performs an excellent accuracy in five classes second only to our proposed MGEN, achieving an OA of 96.38%.

The visualized results of class predictions for all methods on the Houston dataset are shown in Figure 5, illustrating that the prediction results of MGEN are closer to the class map of ground truth, with fewer misclassified pixels, thereby achieving higher accuracy compared to other methods. Notably, samples of "Parking lot 2", colored in purple, present a challenge for many comparative methods due to their spectral similarity to "Parking lot 1", leading to frequent misclassifications and confusion. As a result, some of these methods struggle to achieve high classification accuracy for this class. Despite all this, our network still exhibits an impressive accuracy (99.29%). This can be ascribed to the excellent ability of MGEN to extract deep-seated spectral features of land cover with multiple scales in graph space, which enables effective differentiation between the samples with spectrally similar but distinct classes, thereby demonstrating the superior discriminative capability of MGEN in addressing challenging classification cases.

In summary, the experimental results on Trento, MUUFL, and Houston demonstrate that MGEN consistently outperforms other state-of-the-art methods across multiple datasets. Traditional machine learning methods, like SVM and LE, struggle to handle the high-dimensional and complex nature of multi-modal data, especially when compared to deep learning-based methods. The experimental results show that SVM performs significantly worse across all metrics. For example, on the MUUFL dataset, SVM achieved an OA of only 85.46%, which is nearly 10% lower than MGEN's OA of 95.36%. Methods using single-scale feature extraction such as TBCNN and 3D CNN are restricted by their ability

to adapt to the varying scales of objects within large remote sensing scenes, which can be observed in datasets like the MUUFL and Houston datasets, where objects can range from small, narrow roads to large, complex forests or urban areas. Compared to them, the multi-scale graph-based structure of MGEN captures both local and global spatial features effectively, resulting in clearer classification boundaries and fewer misclassifications. This is particularly evident in complex scenes where objects vary greatly in size and structure, such as the dense Parking lots 1 and 2 surrounded by Roads and Trees on the Houston dataset (Figure 5). Other graph-based methods like SSSE and HRWN, despite their global feature modeling capabilities, struggle to integrate multi-scale representations, limiting their ability to process features on smaller scales. This weakness leads to suboptimal performance when handling datasets that contain both large-scale (e.g., Trees) and small-scale objects (e.g., Cloth panels). The results on the MUUFL dataset highlight this issue, where MGEN outperformed HRWN with an accuracy on the class "Cloth panel" of 98.75% versus HRWN's 88.65%.

4. Discussion

4.1. Parameter Analysis

In our proposed network, three branches with different numbers of superpixels are utilized to segment the image over the spatial dimension, facilitating deep feature extraction from multi-modal data. As shown in Equation (1), the segmentation scale parameter λ in each branch affects the extraction of graph features, which in turn, influences the accuracy of land cover classification. Therefore, in this section, we will examine the impact of λ in three scales on classification accuracy.

Following the previously described experimental settings and taking the MUUFL dataset as an example, we conduct multiple experiments to evaluate the land cover classification accuracy of the proposed MGEN on multi-modal remote sensing images. The λ in Equation (1) controls the size of superpixels, which is crucial to determining the scale of feature extraction. When λ is small, ground objects may be over-segmented into multiple superpixels. When λ is large, superpixels may combine too many different land cover classes. In order to determine a coarse value range for λ, experiments are conducted on the MUUFL dataset. The results are shown in Figure 6. Specifically, the OA, AA and kappa curves across different values of λ, ranging from 10 to 200, are shown in the figure. The accuracies improve with increasing λ up to a value of 40. At $\lambda = 70$, OA reaches a local maximum. For values of λ greater than 150, no significant improvement in accuracy is observed. Local maxima of OA occur at λ = 70, 100, 120, and 150. The curve of AA and Kappa exhibit similar trends.

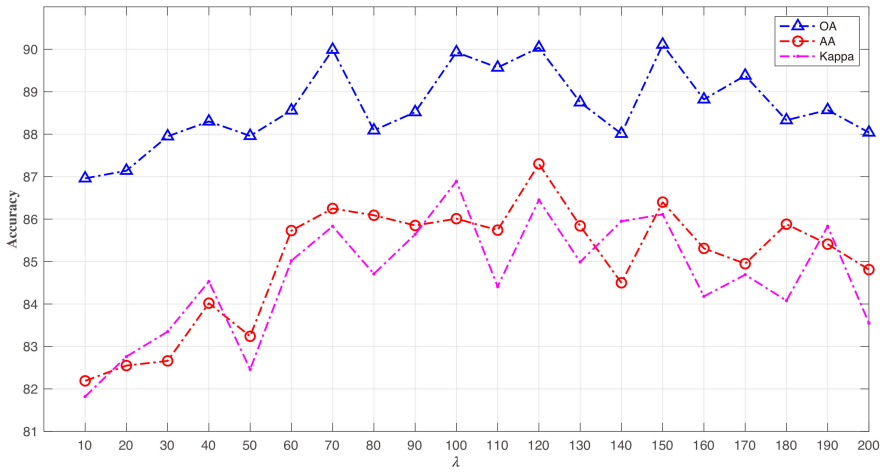

Figure 6. Experimental results of using a single scale with different values of λ.

Based on the observation of a single scale, the values of λ corresponding to these local maximum points, i.e., {70, 100, 120, 150}, are chosen for further investigation to find out the best combination of λ for different scales. We denote the λ in Equation (1) of Scale 1, Scale 2, and Scale 3, respectively, as λ_1, λ_2, and λ_3. Figure 7 presents the classification results for three metrics, with all other parameters held constant while varying only the three scale parameters. λ_1 is assigned one of the candidate values {70, 100, 120, 150}, and λ_2, as well as λ_3, is also set within the same range. In the figure, OA, AA, and Kappa are represented by blue, orange, and green, respectively, with the intensity of the colors indicating the magnitude of the values. The three rows of subfigures correspond to different values of λ_1, and each subfigure in the 4×3 grid (4 values of λ_1 and 3 metrics) shows the classification results obtained by MGEN with a fixed λ_1 while varying the values of λ_2 and λ_3.

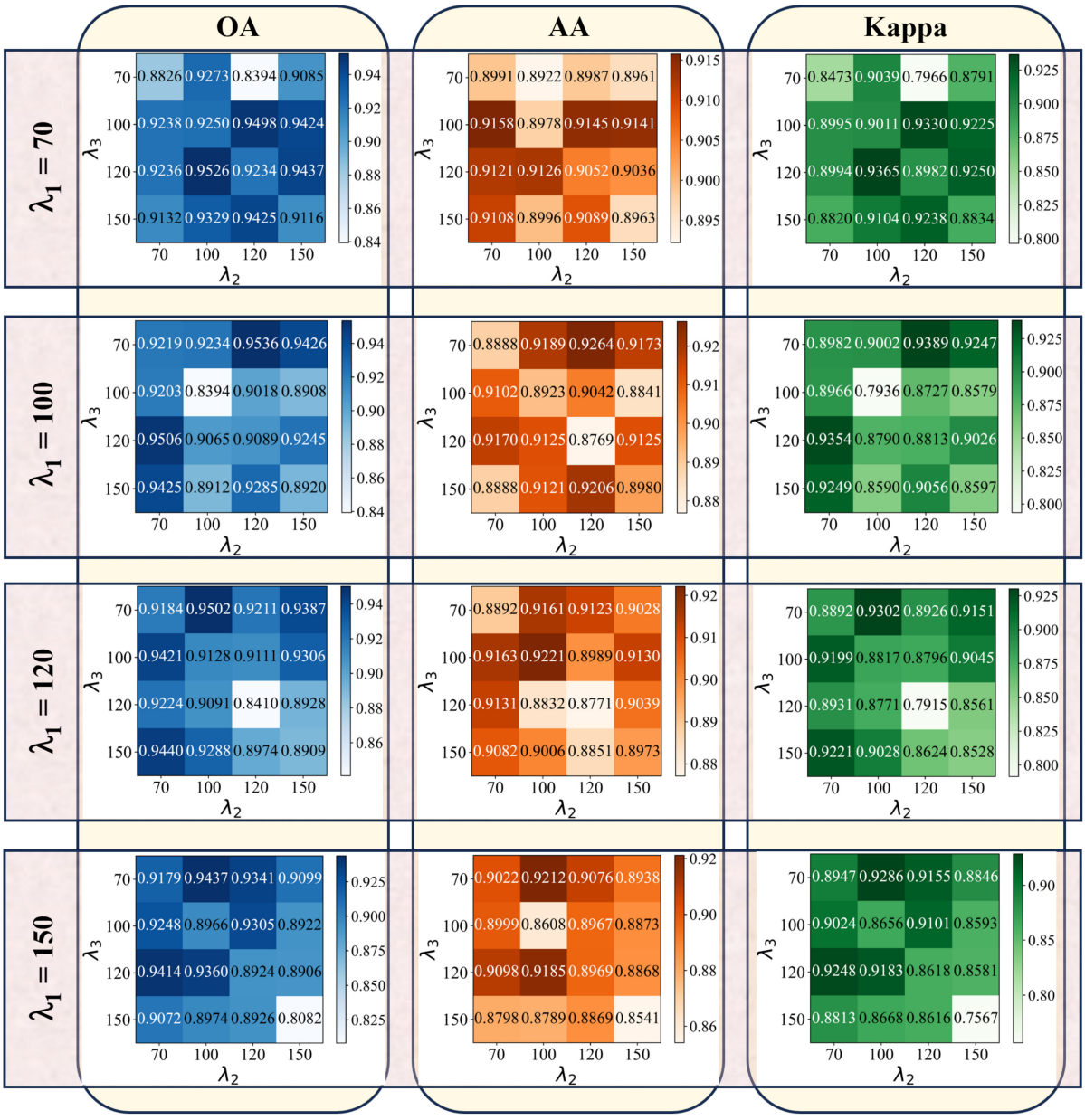

Figure 7. Classification accuracy results of the proposed MGEN with different scale parameters on MUUFL dataset. OA, AA, and Kappa are displayed by colors of blue, orange, and green, respectively, with the intensity of the colors indicating the magnitude of the values.

Figure 7 demonstrates that when the scale parameters in the multi-branch network are set to identical values, the classification accuracy tends to decrease, which can be attributed to the difficulty in fully utilizing the advantages of the multi-scale feature extraction at a single scale. In contrast, when the number of superpixels in the three branches is set to $\lambda_1 = 100$, $\lambda_2 = 120$, and $\lambda_3 = 70$, the classification metrics, depicted in the darkest colors of the figure, achieve their highest values: OA = 95.36%, AA = 92.64% and Kappa = 93.89%. Although other scale values also yield good classification accuracy, these specific settings optimize the performance of MGEN for land cover classification. Considering the actual size of land cover, the approximate size of 70 for a superpixel tends to correspond to a scale that emphasizes finer spatial details, allowing the model to capture small-scale features like Sidewalk and Road curbs. When the superpixel size comes to 120, the scale emphasizes larger context features, providing essential context for moderately sized objects such as Buildings or clusters of Trees. This multi-scale combination allows MGEN to capture local features while simultaneously considering larger, global patterns, improving the model's ability to classify both small-scale features (e.g., Roads, Yellow curb) and large-scale features (e.g., forests, water bodies). Consequently, these scale values were used in the comparative experiments of the last section to maximize the classification capabilities of the proposed MGEN.

4.2. Ablation Study

As shown in Figure 1, MGEN utilizes a multi-branch network with various scales for superpixel segmentation to extract graph features. To demonstrate the benefits of multi-scale branches in classification tasks, an ablation study is conducted here to assess the contributions of different superpixel segmentations. Using the MUUFL dataset as an example, we separately execute single-branch, two-branch, three-branch and four-branch networks in the experiments, with scale parameters λ_1, λ_2, λ_3 and λ_4 set to 100, 120, 70 and 150 according to the analysis in Section 4.1. All other parameters in experiments are kept the same as previously mentioned.

The experimental configurations of the ablation study are exhibited in the first five columns of Table 7, where ✓ and χ indicate the presence or absence of the corresponding branches, respectively. The table also reports the classification results for the three experiments, including the accuracy metrics of AA, OA and Kappa. In general, as the number of branches increases, the classification accuracy is obviously improved. The multi-branch classification network with three different scales, compared to the two-scale and single-scale networks, raised the OA by 1.78% and 5.43%, respectively. Assuming a single-branch network as the baseline, it can be seen that AA is progressively incremented by 4.94%, 1.66% and 0.49% along with the three increases in branching. Kappa likewise rose from 86.89% on a single-branch network to 91.71% on a double-branch one, before reaching 93.89% on a triple-branch one. It is evident that the proposed MGEN with multiple scales, like 3- or 4-scale, significantly outperforms the 1- or 2-scale models. This is attributed to the ability of multi-scale superpixel segmentation to fully leverage spatial-spectral information, and subsequently excavate deep-seated features in the multi-modal data from diverse levels. The land cover features obtained by the simultaneous introduction of three or more branches resulted in a significant improvement in the classification accuracy, yielding the best results in the ablation study of our proposed MGEN.

Table 7. Ablation study of the proposed MGEN with different branches on MUUFL dataset.

No.	Scale 1	Scale 2	Scale 3	Scale 4	OA	AA	Kappa
1	✓	χ	χ	χ	89.93%	86.01%	86.89%
2	✓	✓	χ	χ	93.58%	90.98%	91.71%
3	✓	✓	✓	χ	**95.36%**	92.64%	**93.89%**
4	✓	✓	✓	✓	95.22%	**93.13%**	92.90%

As shown in the last two rows of Table 7, the three-branch network achieves higher accuracies across all metrics compared to those with fewer branches, but the addition of an extra scale in the fourth experiment does not significantly enhance classification performance. This indicates that the classification accuracy of the proposed MGEN does not continuously improve with the addition of more branches. This outcome is likely due to the fact that the fourth scale parameter $\lambda_4 = 150$ is excessive, making it difficult to capture key information from each superpixel. Moreover, segmenting the image using three scales already provides sufficient information for the graph model, and there is no evidence that adding more scales boosts classification performance. Thus, the three-scale strategy used in the previous experiments is validated as reasonable.

In this section, we also discuss the effectiveness of the proposed long- and short-range attention graph convolution module in MGEN. As described in Section 2.3.2, the module integrates convolution strategies for both short-range and long-range nodes, enabling local and global feature extraction, respectively. Table 8 lists the classification accuracies obtained by various strategies on the MUUFL dataset, including a long-range branch solely, a short-range branch solely, and a combination of both branches. The three experiments used the completed framework of MGEN with all parameters consistent with previous settings, except for the long- and short-range strategies. The experimental results in Table 8 demonstrate that the combined strategy achieves much higher accuracies in OA, AA, and Kappa, outperforming either branch, especially the OA of the third experiment is improved by 1.25% and 3.44% compared to the short-range and long-range strategies, respectively. This can be attributed to the increased weight of features extracted by the short-range branch when nodes of the same class are highly concentrated, effectively leveraging local information. Vice versa, the attention weights of the long-range nodes are enhanced, inclining to exploit global information. Therefore, the MGEN, with the introduction of the long-range and short-range attention graph convolution module, is capable of extracting more comprehensive feature information from multi-modal data, thereby improving classification accuracy.

Table 8. Ablation study of the proposed MGEN with long- or short-range strategies.

No.	Long-Range	Short-Range	OA	AA	Kappa
1	✓	χ	91.92%	89.15%	89.81%
2	χ	✓	94.11%	90.73%	92.36%
3	✓	✓	**95.36%**	**92.64%**	**93.89%**

5. Conclusions

In this paper, we propose a multi-scale graph encoder–decoder network for multi-modal data classification, to address the challenge of integrating local-global information in remote sensing HSI and LiDAR images. Graph encoders are employed in the multiple branches to map multi-modal images with a variety of scales into the graph space, allowing for feature learning in the graph space. Subsequently, graph decoders are adopted to fuse multi-scale features and map them back to the original space for pixel-level classification. Specifically, the image is first segmented into a series of superpixels by the SLIC algorithm in a graph encoder, with each superpixel treated as a node in the graph space. By controlling the fineness of the segmentation algorithm, superpixels with different scales are generated. Based on the multi-scale superpixels, CNNs are employed for feature learning of the graph nodes to complete the mapping from images to topological graphs. Then, multi-branch graph CNNs are used for graph feature extraction. The graph decoder is responsible for fusing multi-scale features and mapping them to the original data scale to generate classification results. Experimental results on three HSI-LiDAR datasets demonstrate that the proposed MGEN achieves superior performance, surpassing many state-of-the-art multi-modal data classification methods.

Author Contributions: Conceptualization, F.W. and X.D.; methodology, X.D. and W.Z.; software, X.D.; validation, F.W. and H.W.; formal analysis, H.W.; investigation, W.Z.; resources, J.M.; data curation, S.Z.; writing—original draft preparation, X.D.; writing—review and editing, F.W.; visualization, F.W.; supervision, L.N.; project administration, L.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by Key Scientific Research Plan of Education Department of Shaanxi Province, China, No. 20JY029, in part by Aeronautical Science Foundation of China, No. 202000190U1002.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Acknowledgments: The authors express their gratitude to the peer researchers for providing their source codes and the publicly available HSI and LiDAR datasets.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Xu, H.; Zheng, T.; Liu, Y.; Zhang, Z.; Xue, C.; Li, J. A joint convolutional cross ViT network for hyperspectral and light detection and ranging fusion classification. *Remote Sens.* **2024**, *16*, 489. [CrossRef]
2. Wang, G.; Chen, J.; Mo, L.; Wu, P.; Yi, X. Border-Enhanced Triple Attention Mechanism for High-Resolution Remote Sensing Images and Application to Land Cover Classification. *Remote Sens.* **2024**, *16*, 2814. [CrossRef]
3. Xu, H.; Chen, W.; Tan, C.; Ning, H.; Sun, H.; Xie, W. Orientational Clustering Learning for Open-Set Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 5508605. [CrossRef]
4. Liu, Y.; Jiang, S.; Liu, Y.; Mu, C. Spatial Feature Enhancement and Attention-Guided Bidirectional Sequential Spectral Feature Extraction for Hyperspectral Image Classification. *Remote Sens.* **2024**, *16*, 3124. [CrossRef]
5. Zhang, M.; Li, W.; Zhang, Y.; Tao, R.; Du, Q. Hyperspectral and LiDAR Data Classification Based on Structural Optimization Transmission. *IEEE Trans. Cybern.* **2023**, *53*, 3153–3164. [CrossRef]
6. Yao, H.; Chen, R.; Chen, W.; Sun, H.; Xie, W.; Lu, X. Pseudo-Label-Based Unreliable Sample Learning for Semi-Supervised Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5527116. [CrossRef]
7. Chen, Z.; Chen, Y.; Wang, Y.; Wang, X.; Wang, X.; Xiang, Z. DCFF-Net: Deep Context Feature Fusion Network for High-Precision Classification of Hyperspectral Image. *Remote Sens.* **2024**, *16*, 3002. [CrossRef]
8. Wang, A.; Dai, S.; Wu, H.; Iwahori, Y. Multimodal Semantic Collaborative Classification for Hyperspectral Images and LiDAR Data. *Remote Sens.* **2024**, *16*, 3082. [CrossRef]
9. Li, Z.; Liu, R.; Sun, L.; Zheng, Y. Multi-Feature, Cross Attention-Induced Transformer Network for Hyperspectral and LiDAR Data Classification. *Remote Sens.* **2024**, *16*, 2775. [CrossRef]
10. Hong, D.; Gao, L.; Yokoya, N.; Yao, J.; Chanussot, J.; Du, Q.; Zhang, B. More Diverse Means Better: Multimodal Deep Learning Meets Remote-Sensing Imagery Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 4340–4354. [CrossRef]
11. Ning, H.; Lei, T.; An, M.; Sun, H.; Hu, Z.; Nandi, A.K. Scale-wise interaction fusion and knowledge distillation network for aerial scene recognition. *CAAI Trans. Intell. Technol.* **2023**, *8*, 1178–1190. [CrossRef]
12. Li, J.; Liu, Y.; Song, R.; Liu, W.; Li, Y.; Du, Q. HyperMLP: Superpixel Prior and Feature Aggregated Perceptron Networks for Hyperspectral and Lidar Hybrid Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5505614. [CrossRef]
13. Bai, L.; Li, Y.; Cen, M.; Hu, F. 3D Instance Segmentation and Object Detection Framework Based on the Fusion of Lidar Remote Sensing and Optical Image Sensing. *Remote Sens.* **2021**, *13*, 3288. [CrossRef]
14. Wang, F.; Zhou, G.; Xie, J.; Fu, B.; You, H.; Chen, J.; Shi, X.; Zhou, B. An Automatic Hierarchical Clustering Method for the LiDAR Point Cloud Segmentation of Buildings via Shape Classification and Outliers Reassignment. *Remote Sens.* **2023**, *15*, 2432. [CrossRef]
15. Du, X.; Zheng, X.; Lu, X.; Doudkin, A.A. Multisource Remote Sensing Data Classification With Graph Fusion Network. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 10062–10072. [CrossRef]
16. Cao, M.; Zhao, G.; Lv, G.; Dong, A.; Guo, Y.; Dong, X. Spectral–Spatial–Language Fusion Network for Hyperspectral, LiDAR, and Text Data Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5503215. [CrossRef]
17. Li, J.; Liu, Y.; Song, R.; Li, Y.; Han, K.; Du, Q. Sal^{2}RN: A Spatial–Spectral Salient Reinforcement Network for Hyperspectral and LiDAR Data Fusion Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5500114. [CrossRef]
18. Chroni, A.; Vasilakos, C.; Christaki, M.; Soulakellis, N. Fusing Multispectral and LiDAR Data for CNN-Based Semantic Segmentation in Semi-Arid Mediterranean Environments: Land Cover Classification and Analysis. *Remote Sens.* **2024**, *16*, 2729. [CrossRef]
19. Lu, T.; Ding, K.; Fu, W.; Li, S.; Guo, A. Coupled adversarial learning for fusion classification of hyperspectral and LiDAR data. *Inf. Fusion* **2023**, *93*, 118–131. [CrossRef]
20. Zhang, W.; Wang, X.; Wang, H.; Cheng, Y. Causal Meta-Reinforcement Learning for Multimodal Remote Sensing Data Classification. *Remote Sens.* **2024**, *16*, 1055. [CrossRef]

21. Guo, F.; Meng, Q.; Li, Z.; Ren, G.; Wang, L.; Zhang, J.; Xin, R.; Hu, Y. Multisource Feature Embedding and Interaction Fusion Network for Coastal Wetland Classification With Hyperspectral and LiDAR Data. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5509516. [CrossRef]
22. Song, W.; Dai, Y.; Gao, Z.; Fang, L.; Zhang, Y. Hashing-Based Deep Metric Learning for the Classification of Hyperspectral and LiDAR Data. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5704513. [CrossRef]
23. Wang, X.; Zhu, J.; Feng, Y.; Wang, L. MS2CANet: Multiscale Spatial–Spectral Cross-Modal Attention Network for Hyperspectral Image and LiDAR Classification. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 5501505. [CrossRef]
24. Zhao, X.; Tao, R.; Li, W.; Li, H.C.; Du, Q.; Liao, W.; Philips, W. Joint Classification of Hyperspectral and LiDAR Data Using Hierarchical Random Walk and Deep CNN Architecture. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7355–7370. [CrossRef]
25. Feng, Y.; Song, L.; Wang, L.; Wang, X. DSHFNet: Dynamic Scale Hierarchical Fusion Network Based on Multiattention for Hyperspectral Image and LiDAR Data Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5522514. [CrossRef]
26. Wang, M.; Gao, F.; Dong, J.; Li, H.C.; Du, Q. Nearest Neighbor-Based Contrastive Learning for Hyperspectral and LiDAR Data Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5501816. [CrossRef]
27. Xue, Z.; Yu, X.; Tan, X.; Liu, B.; Yu, A.; Wei, X. Multiscale Deep Learning Network With Self-Calibrated Convolution for Hyperspectral and LiDAR Data Collaborative Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [CrossRef]
28. Gao, H.; Feng, H.; Zhang, Y.; Xu, S.; Zhang, B. AMSSE-Net: Adaptive Multiscale Spatial–Spectral Enhancement Network for Classification of Hyperspectral and LiDAR Data. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5531317. [CrossRef]
29. Mohla, S.; Pande, S.; Banerjee, B.; Chaudhuri, S. Fusatnet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 416–425. [CrossRef]
30. Zhou, Y.; Wang, C.; Zhang, H.; Wang, H.; Xi, X.; Yang, Z.; Du, M. TCPSNet: Transformer and Cross-Pseudo-Siamese Learning Network for Classification of Multi-Source Remote Sensing Images. *Remote Sens.* **2024**, *16*, 3120. [CrossRef]
31. Wang, M.; Sun, Y.; Xiang, J.; Sun, R.; Zhong, Y. Joint Classification of Hyperspectral and LiDAR Data Based on Adaptive Gating Mechanism and Learnable Transformer. *Remote Sens.* **2024**, *16*, 1080. [CrossRef]
32. Wang, W.; Li, C.; Ren, P.; Lu, X.; Wang, J.; Ren, G.; Liu, B. Dual-Branch Feature Fusion Network Based Cross-Modal Enhanced CNN and Transformer for Hyperspectral and LiDAR Classification. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 5503605. [CrossRef]
33. Zhao, G.; Ye, Q.; Sun, L.; Wu, Z.; Pan, C.; Jeon, B. Joint Classification of Hyperspectral and LiDAR Data Using a Hierarchical CNN and Transformer. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5500716. [CrossRef]
34. Song, T.; Zeng, Z.; Gao, C.; Chen, H.; Li, J. Joint Classification of Hyperspectral and LiDAR Data Using Height Information Guided Hierarchical Fusion-and-Separation Network. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5505315. [CrossRef]
35. Yang, J.X.; Zhou, J.; Wang, J.; Tian, H.; Liew, A.W.C. LiDAR-Guided Cross-Attention Fusion for Hyperspectral Band Selection and Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–15. [CrossRef]
36. Zhang, Y.; Peng, Y.; Tu, B.; Liu, Y. Local Information Interaction Transformer for Hyperspectral and LiDAR Data Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 1130–1143. [CrossRef]
37. Ni, K.; Wang, D.; Zheng, Z.; Wang, P. MHST: Multiscale Head Selection Transformer for Hyperspectral and LiDAR Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 5470–5483. [CrossRef]
38. Sun, L.; Wang, X.; Zheng, Y.; Wu, Z.; Fu, L. Multiscale 3-D–2-D Mixed CNN and Lightweight Attention-Free Transformer for Hyperspectral and LiDAR Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 2100116. [CrossRef]
39. Feng, Y.; Zhu, J.; Song, R.; Wang, X. S2EFT: Spectral-Spatial-Elevation Fusion Transformer for hyperspectral image and LiDAR classification. *Knowl. Based Syst.* **2024**, *283*, 111190. [CrossRef]
40. Zhao, X.; Zhang, M.; Tao, R.; Li, W.; Liao, W.; Tian, L.; Philips, W. Fractional Fourier image transformer for multimodal remote sensing data classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *35*, 2314–2326. [CrossRef]
41. Feng, J.; Zhang, J.; Zhang, Y. Multiview Feature Learning and Multilevel Information Fusion for Joint Classification of Hyperspectral and LiDAR Data. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5528613. [CrossRef]
42. Cai, J.; Zhang, M.; Yang, H.; He, Y.; Yang, Y.; Shi, C.; Zhao, X.; Xun, Y. A novel graph-attention based multimodal fusion network for joint classification of hyperspectral image and LiDAR data. *Expert Syst. Appl.* **2024**, *249*, 123587. [CrossRef]
43. Wan, S.; Gong, C.; Zhong, P.; Pan, S.; Li, G.; Yang, J. Hyperspectral image classification with context-aware dynamic graph convolutional network. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 597–612. [CrossRef]
44. Cai, W.; Wei, Z. Remote sensing image classification based on a cross-attention mechanism and graph convolution. *IEEE Geosci. Remote Sens. Lett.* **2020**, *19*, 8002005. [CrossRef]
45. Sha, A.; Wang, B.; Wu, X.; Zhang, L. Semisupervised classification for hyperspectral images using graph attention networks. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 157–161. [CrossRef]
46. Zhu, W.; Zhao, C.; Feng, S.; Qin, B. Multiscale short and long range graph convolutional network for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5535815. [CrossRef]
47. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Süsstrunk, S. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2274–2282. [CrossRef]
48. Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P.S. A Comprehensive Survey on Graph Neural Networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 4–24. [CrossRef]

49. Hammond, D.K.; Vandergheynst, P.; Gribonval, R. Wavelets on graphs via spectral graph theory. *Appl. Comput. Harmon. Anal.* **2011**, *30*, 129–150. [CrossRef]
50. Qi, X.; Wang, T.; Liu, J. Comparison of Support Vector Machine and Softmax Classifiers in Computer Vision. In Proceedings of the 2017 Second International Conference on Mechanical, Control and Computer Engineering (ICMCCE), Harbin, China, 8–10 December 2017; pp. 151–155. [CrossRef]
51. Ghamisi, P.; Hofle, B.; Zhu, X.X. Hyperspectral and LiDAR Data Fusion Using Extinction Profiles and Deep Convolutional Neural Network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 3011–3024. [CrossRef]
52. Gader, P.; Zare, A.; Close, R.; Aitken, J.; Tuell, G. *MUUFL Gulfport Hyperspectral and LiDAR Airborne Data Set*; Technical Report REP-2013-570; University Florida: Gainesville, FL, USA, 2013.
53. Du, X.; Zare, A. *Scene Label Ground Truth Map for MUUFL Gulfport Data Set*; Technical Report; University of Florida: Gainesville, FL, USA, 2017.
54. Debes, C.; Merentitis, A.; Heremans, R.; Hahn, J.; Frangiadakis, N.; van Kasteren, T.; Liao, W.; Bellens, R.; Pižurica, A.; Gautama, S.; et al. Hyperspectral and LiDAR Data Fusion: Outcome of the 2013 GRSS Data Fusion Contest. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 2405–2418. [CrossRef]
55. Cahill, N.D.; Czaja, W.; Messinger, D.W. Schroedinger Eigenmaps with nondiagonal potentials for spatial-spectral clustering of hyperspectral imagery. In Proceedings of the Defense + Security Symposium, San Jose, CA, USA, 8–21 May 2014. [CrossRef]
56. Lee, H.; Kwon, H. Going Deeper with Contextual CNN for Hyperspectral Image Classification. *IEEE Trans. Image Process.* **2017**, *26*, 4843–4855. [CrossRef] [PubMed]
57. Ben Hamida, A.; Benoit, A.; Lambert, P.; Ben Amar, C. 3-D Deep Learning Approach for Remote Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 4420–4434. [CrossRef]
58. Xu, X.; Li, W.; Ran, Q.; Du, Q.; Gao, L.; Zhang, B. Multisource Remote Sensing Data Classification Based on Convolutional Neural Network. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 937–949. [CrossRef]
59. Zhao, X.; Tao, R.; Li, W. Multisource Remote Sensing Data Classification Using Deep Hierarchical Random Walk Networks. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 2187–2191. [CrossRef]