Multimodal Attention-Aware Convolutional Neural Networks for Classification of Hyperspectral and LiDAR Data

Haotian Zhang, Jing Yao, *Member, IEEE*, Li Ni, Lianru Gao, *Senior Member, IEEE*, and Min Huang

***Abstract*—The attention mechanism is one of the most influential ideas in the deep learning community, which has shown excellent efficiency in various computer vision tasks. Thus, this article proposes the convolution neural network method with the attention mechanism to enhance the feature extraction of light detection and ranging (LiDAR) data. Meanwhile, our elaborately designed cascaded block contains a short path architecture beneficial for multistage information exchange. With the full exploitation of elevation information from LiDAR data and efficient utilization of the spatial-spectral information underlying hyperspectral data, our method provides a novel solution for multimodal feature fusion. Experiments are conducted on the LiDAR and hyperspectral dataset provided by the 2013 IEEE GRSS Data Fusion Contest and multisource Trento dataset to demonstrate the effectiveness of the proposed method. The experimental results have shown the superior results of the proposed method on both LiDAR and multimodality remote sensing data in comparison with several popular baselines.**

***Index Terms*—Attention mechanism, convolution neural network (CNN), hyperspectral, light detection and ranging (LiDAR), multimodality.**

I. Introduction

REMOTE sensing image classification task plays an essential role in Earth observation, which could be used for analyzing critical information related to urban planning, natural resources management, climate change, environmental monitoring, and so on. Remote sensing data acquired from various sensors could exploit multiple physical characteristics of ground objects [1], [2], [3], [4], [5]. With the blooming development of remote sensing sensors, more and more researchers in the remote sensing community are active in the algorithm innovations to better extract the most valuable features among the multimodal remote sensing data [6]. Mostly, it is complicated and challenging for some algorithms to extract the feature of ground objects efficiently. For example, it is hard to distinguish different ground objects in the downtown area with a high building density. In this case, various remote sensing data could facilitate the algorithm to improve the image classification results more precisely [7].

Hyperspectral image (HSI) can provide detailed spectral information of various ground cover types due to its broad coverage of wavelength and high sampling rate [8], [9]. Usually, HSI contains dozens or hundreds of spectral information ranging from the visible light (0.4–0.7 μm) bands to the short-wave infrared (almost 2.4 μm) bands. Thus, HSI with sufficient spectral information could discriminate ground objects with similar spatial features [10]. Nevertheless, hyperspectral data could not contain height information of ground objects as well as high-resolution spatial information. Meanwhile, there are complex mixed pixels and noising signal, which prevent the precise classification results [11], [12]. Classifying ground objects with similar spectral and spatial features could hardly distinguish ground objects only with the HSI, and lots of researchers have tried to improve the HSI classification accuracy [13], [14], [15], [16], [17], [18], [19], [20]. To this end, LiDAR data can provide elevation information to extract more precise features of various ground objects. Consequently, LiDAR data provide elevation information, which is a beneficial source for complementing the information provided solely by HSI [21].

Researchers have proposed a series of methods to better realize remote sensing image classification tasks using HSI and LiDAR data in recent years. To further strengthen the spatial feature, filtering-based methods have been proposed, which mostly could extract the regional geometrical feature, meanwhile, preserve the most critical spatial characteristic of HSI [22], [23], [24], [25], [26]. However, the filtering-based methods mostly increase the dimension of multimodal remote sensing data, which probably introduces the curse of dimensionality, decreasing the accuracy of classification results. Furthermore, the nonlinear characteristic of spectral information in HSI would be amplified when integrated with LiDAR data. The methods based on deep learning [27], [28], [29], [30], [31] could extract more complex and hierarchical features of multimodal remote sensing data, which have been experimented with in

Manuscript received 11 May 2022; revised 15 June 2022; accepted 26 June 2022. Date of publication 1 July 2022; date of current version 14 April 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFB3900502, and in part by the National Natural Science Foundation of China under Grant 62201553 and Grant 42030111. *(Corresponding author: Jing Yao.)*

Haotian Zhang is with the Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, and also with the School of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: zhanghaotian19@mails.ucas.ac.cn).

Jing Yao, Li Ni, Lianru Gao, and Min Huang are with the Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China (e-mail: jasonyao92@gmail.com; nili@aircas.ac.cn; gaolr@aircas.ac.cn; huangmin@aircas.ac.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3187730

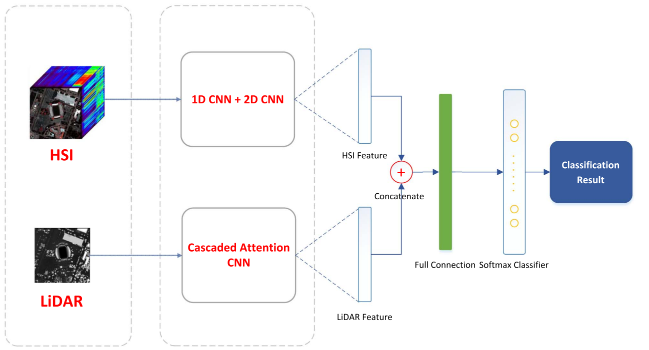

Fig. 1. Architecture of training multimodal data separately with 1D CNN + 2D CNN and cascaded attention CNN.

recent years with better classification results than other classical machine learning methods (e.g., support vector machine [32] and extreme learning machine [33], [34]).

A. Motivation

The methods based on deep learning behave perform better on the extraction of the complicated multimodal feature than other traditional machine learning methods [25], [31]. Among the methods based on deep learning, most would use the different kinds of convolution neural networks (CNNs) to extract the features acquired from different modalities. Meanwhile, there is lots of work trying to combine the filtering-based method with CNN to introduce more expert experience [25]. Then the features could be fused by being concatenated or point-wisely added. Besides, Hong et al. designed the common subspace representations to extract the integrated multimodal remote sensing data feature with EndNet followed a deep encoder–decoder network architecture [27]. Furthermore, researchers also attempt to combine the graph-based method with CNN to preserve the spatial edge information of ground objects [35].

Attention mechanism methods are popular in the natural language processing and computer vision area these years [36]. In the remote sensing community, researchers have also conducted experiments to explore the positive impact of the attention mechanism on deep learning-based methods [37], [38], [39], [40], [41], [42], [43]. When combing with the CNN, the attention mechanism could focus on the most vital features and weaken the impact of unnecessary features.

Hence, there is a potential space for us to explore the impact of attention mechanisms on multimodal remote sensing image classification. In the following section, we will illustrate the main contribution we made to this research.

B. Contribution

The framework of our proposed method is shown in Fig. 1. More concretely, the significant contributions in this article could be concluded as the following two aspects.

- *LiDAR multiscale cascaded CNN:* We have designed a multiscale cascaded-based deep CNN to extract the spatial feature of LiDAR DSM images. Compared with conventional CNN, the cascade block we designed could better extract multiscale hierarchical spatial features of LiDAR data.
- *LiDAR attention module blocks:* The attention mechanism module was applied to emphasize the most meaningful information contained in the LiDAR data. In this way, the feature extracted from LiDAR could better contribute to the whole multimodal data feature and final classification.

II. Related Work

In this section, we will briefly introduce the background of CNN and attention mechanism.

A. Convolution Neural Network

The CNN is an efficient deep learning model to extract the hierarchical feature of image information. The CNN contains a series of convolution layers, pooling layers, and activation function [44]. Some researchers have explored the efficiency of algorithms based on CNN with multimodal remote sensing image classification tasks.

Hang et al. [28] designed a simple two-stream CNN to extract the feature of hyperspectral and LiDAR data separately. As remote sensing data have the property of a large covering area, CNN's input data are usually a patch derived from the remote sensing data (such as a LiDAR image patch with the size of 5×5). Besides, owing to the abundant spectral information in the HSI, Xu et al. designed a one-dimensional CNN and two-dimensional CNN to extract the spectral and spatial features separately.

However, the feature extracted by CNN is strongly influenced by the network architecture [45]. To further explore the potential feature, we have proposed a modified network to exploit multimodal hierarchical features better.

B. Attention Mechanism

As mentioned previously, the attention mechanism can recalibrate the significant impact on various feature derived from the CNN output. A few researchers have proved that attention mechanism positively impact on the HSI classification task [46], [47]. Mei et al. [48] proposed the spatial attention CNN and spectral attention recurrent neural network and proven the effectiveness of attention mechanism in HSI classification.

Nevertheless, we still need to evaluate the efficiency of attention mechanism on the multimodal remote sensing image classification task.

III. Methodology

In this section, we will first introduce the algorithm framework we proposed and how to realize the training process. Then the cascaded CNN will be illustrated. Finally, we will focus on introducing the cascaded attention CNN we proposed.

A. Method Overview

We introduce the hierarchical CNN to extract the image feature contained in the HSI and LiDAR DSM data separately. The extracted image feature will be a one-dimensional vector

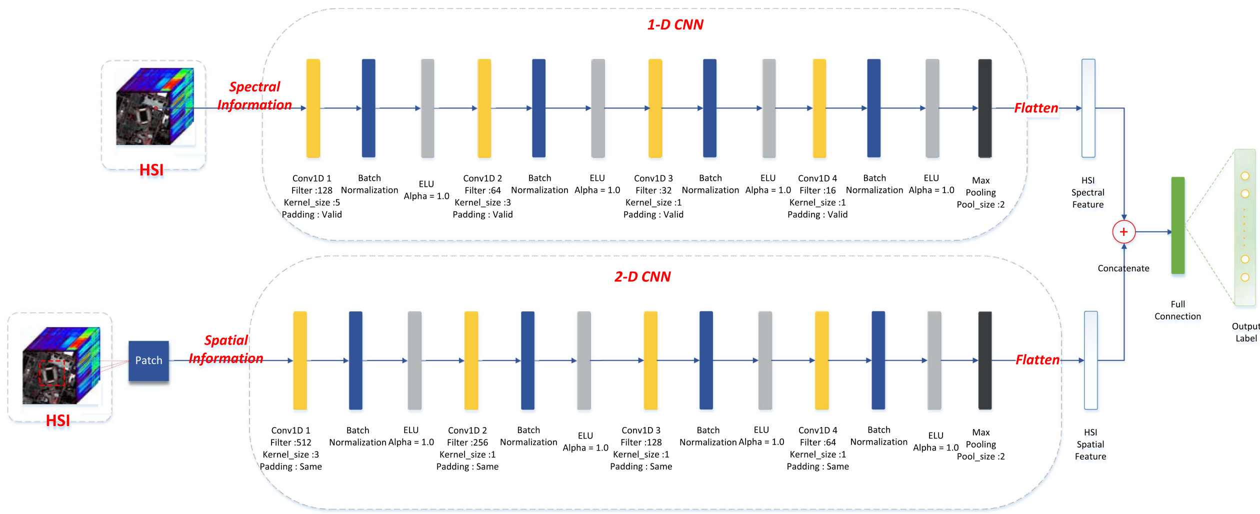

Fig. 2. Detailed network architecture of 1-D CNN and 2-D CNN in the proposed hyperspectral CNN.

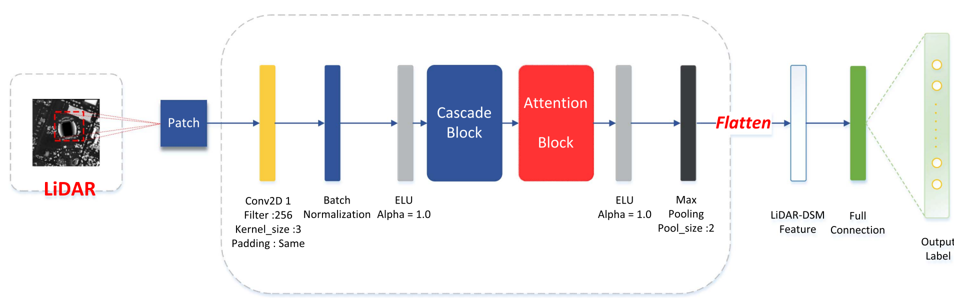

Fig. 3. Network of LiDAR information extraction, the input of LiDAR data patch will go through the cascade block and attention block.

as the result of CNN. We concatenate the derived feature as the multimodal image feature during the fusion stage. Then, a Softmax classifier is applied for the classification task.

B. Hyperspectral CNN

We have designed a Co-CNN hybrid network for the HSI image $\mathbf{H}^{\mathrm{M}\times\mathrm{N}\times\mathrm{K}}$ feature extraction CNN to exploit both two-dimensional spatial and one-dimensional spectral HSI features. To better gain the HSI spatial feature, we adopt the 9×9 patch $\mathbf{H}_{ij}^{\text{spatial}} \in \mathbb{R}^{9\times 9}$ as the training sample where the centered pixel $\mathbf{p}_{ij}$ has been labeled with ground truth as 2-D CNN input. We take the one-dimension spectral signal sample $\mathbf{H}_{ij}^{\text{spectral}} \in \mathbb{R}^{1\times\mathrm{K}}$ with ground truth as 1-D CNN input for the spectral signal data.

The 1-D CNN and 2-D CNN are five convolution layers with batch normalization and exponential linear unit (ELU) activation function. The batch normalization module could provide the training process with higher training efficiency. Besides, we adopt ELU activation functions to avoid exploding gradients problems and exceed the training process. The spatial feature $\mathbf{F}_{\text{spatial}} \in \mathbb{R}^{1\times\mathrm{p}}$ derived by HSI patches and the spectral feature $\mathbf{F}_{\text{spectral}} \in \mathbb{R}^{1\times\mathrm{q}}$ will be concatenated at the feature fusion stage. The fused feature $\mathbf{F}_{\text{HSI}} = [\mathbf{F}_{\text{spectral}}, \mathbf{F}_{\text{spatial}}] \in \mathbb{R}^{1\times(\mathrm{p+q})}$ will go through full connection layer and the Softmax loss function to predict the classification results. The prediction result as

$$\mathrm{Pred}\left(\mathbf{H}_{ij}^{\text{spatial}}, \mathbf{H}_{ij}^{\text{spectral}}\right) = \frac{\exp(\mathbf{F}_{\text{HSI}}\mathbf{W})}{\sum_{c=1}^{C}\exp(\mathbf{F}_{\text{HSI}}\mathbf{W})_c} \tag{1}$$

where, $\mathbf{W} \in \mathbb{R}^{(\mathrm{p+q})\times\mathrm{C}}$ represents the weights matrix in the prediction layer, C is the number of categories, $\exp(\mathbf{F}_{\text{HSI}}\mathbf{W})_c$ is the exponential function to each element corresponding to class c, and the predicted result on the left-hand side shows the probability of that pixel belongs to each category.

C. LiDAR Cascaded Attention CNN

The overall process of the LiDAR neural network contains the cascaded block and attention block. Following the descending kernel size strategy we mentioned in the cascaded block, the raw LiDAR patch data will be enrolled with a convolution with kernel size 3×3 with BN and ELU functions as in Fig. 3.

Given the LiDAR patch image, cascaded block and attention block will help us locate the key edge feature of ground object height. Then ELU activation and max pooling and flatten functions help us gain the one-dimensional LiDAR DSM feature.

1) Cascaded CNN: Toward the LiDAR image, we designed cascaded-based CNN to exploit the ground object height information in case of losing a key height feature in the propagation process. In detail, we follow the descending kernel size strategy

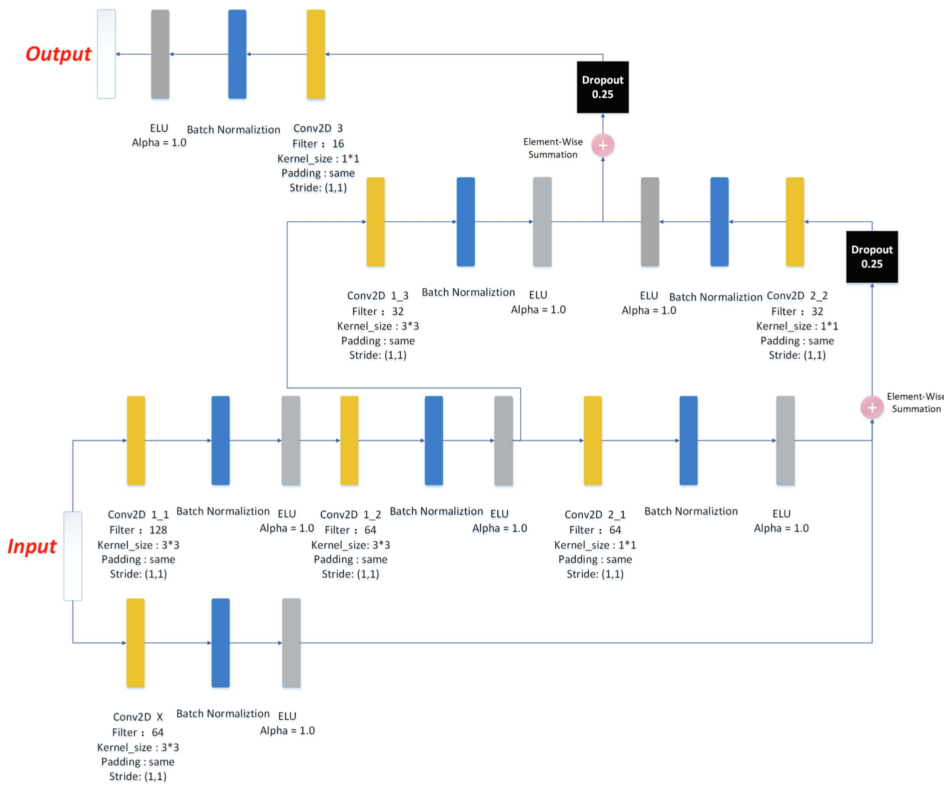

Fig. 4. Cascaded block designs skip connections between convolution layers with descending kernel size to capture multiscale LiDAR height feature.

with skip connection and drop-out operation to exploit the valuable height feature contained in LiDAR data.

Following the training strategy of HSI CNN shown in Fig. 4, in the cascaded block, we maintain the combination of batch normalization and ELU activation function to provide an effective and stable training process and parameters learning results. At the same time, drop-out operation is highlighted to avoid trained features that lack multiscale characteristics.

2) Attention module: The attention module is mainly composed of spatial attention module and channel attention module. The detail network architecture is as in Fig. 5.

The object height feature extracted by a cascaded block will be fed into an attention block, exploiting spatial and channelwise attention based on an efficient framework. The attention block is mainly composed of the channel attention module and spatial attention module, and we define the feature exploited by the cascaded block as $\mathbf{F} \in \mathbb{R}^{\mathrm{M}\times\mathrm{N}\times\mathrm{H}}$. Thus, the whole attention block could be demonstrated as follows:

$$\mathbf{F}'' = \mathbf{f}_{\mathbf{spatial}}\left(\mathbf{F}'\right) \otimes \mathbf{F}' \quad (2)$$

$$\mathbf{F}' = \mathbf{f}_{\mathbf{channel}}\left(\mathbf{F}\right) \otimes \mathbf{F} \quad (3)$$

$$\mathbf{f}_{\mathbf{spatial}}\left(\mathbf{F}'\right) = \varsigma\left(\mathbf{f}_{\mathbf{conv}}\left[\mathbf{f}_{\mathbf{Avg}}\left(\mathbf{F}'\right) \oplus \mathbf{f}_{\mathbf{Max}}\left(\mathbf{F}'\right)\right]\right) \quad (4)$$

$$\mathbf{f}_{\mathbf{channel}}\left(\mathbf{F}\right) = \varsigma\left(\mathbf{f}_{\mathrm{MLP}}\left[\mathbf{f}_{\mathbf{Avg}}\left(\mathbf{F}\right)\right] \oplus \mathbf{f}_{\mathrm{MLP}}\left[\mathbf{f}_{\mathbf{Max}}\left(\mathbf{F}\right)\right]\right) \quad (5)$$

where, (2) and (4) represent the spatial attention module, (3) and (5) represent the channel attention module. The operation $\otimes$ represents elementwise multiplication between features,

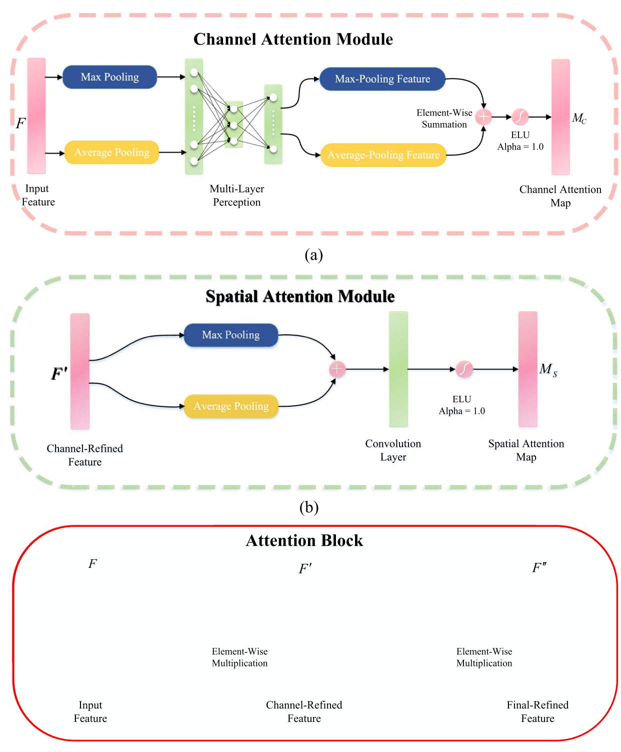

Fig. 5. Details of proposed attention block: (a) shows channel module supporting the interchannel connection of features; (b) shows spatial module strengthening the interspatial relationships of features; (c) shows the overall architecture by integrating abovementioned two blocks.

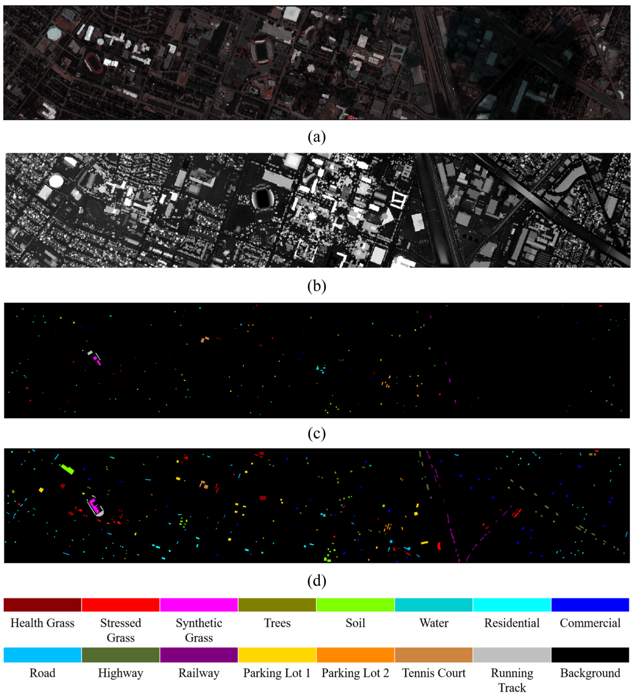

Fig. 6. Houston dataset was used in this experiment: (a) represents the pseudocolor HSI display; (b) represents the gray-scale LiDAR DSM data; (c) shows the training ground-truth samples; (d) shows the testing ground-truth samples.

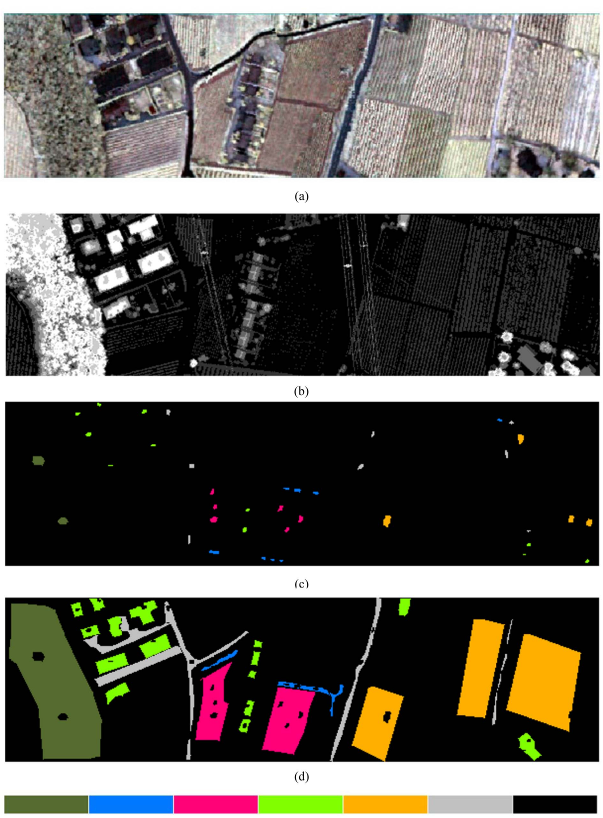

Fig. 7. Trento dataset was used in this experiment: (a) represents the pseudocolor HSI display; (b) represents the gray-scale LiDAR DSM data; (c) shows the training ground-truth samples; (d) shows the testing ground-truth samples.

TABLE I
COUNTS OF HOUSTON TRAINING AND TESTING GROUND TRUTH

Class	# Training samples	# Testing samples
Health grass	198	1053
Stressed grass	190	1064
Synthetic grass	192	505
Trees	188	1056
Soil	186	1056
Water	182	143
Residential	196	1072
Commercial	191	1053
Road	193	1059
Highway	191	1036
Railway	181	1054
Parking lot 1	192	1041
Parking lot 2	184	285
Tennis court	181	247
Running track	187	473
Total	**2832**	**12197**

TABLE II
COUNTS OF TRENTO TRAINING AND TESTING GROUND TRUTH

Class	# Training samples	# Testing samples
Apple trees	129	3905
Buildings	125	2778
Ground	105	374
Woods	154	8969
Vineyard	184	10317
Roads	122	3052
Total	**819**	**29395**

operation $\oplus$ represents elementwise sum between features, operation ς represents ELU activation function, $\mathbf{f}_{\text{Avg}}$ represents average pooling function, $\mathbf{f}_{\mathbf{Max}}$ represents max pooling function.

In the channel attention part, we operate max pooling and global average pooling separately for the input feature, gaining different descriptors, including edge and smooth features for the ground objects. Different descriptors will go through a weight parameter shared multilayer perception $\mathbf{f}_{\mathbf{MLP}}$ with one hidden layer, which would help us gain the channel attention map with $H \times 1 \times 1$ data size. Then, an elementwise summation will be applied toward max-pooling and average-pooling features. Finally, we also follow the network design strategy, allowing fused features to be activated by the ELU activation function for a smoother model training process.

We generate a spatial attention map to highlight the interspatial object height information in the spatial attention sector to enhance the corresponding spatial feature. The feature separately goes through the max-pooling and average-pooling layers following the channel axis. Then, we fused these features with an elementwise summation. The extracted feature along the channel axis is then convolved and activated by the ELU function to get the final spatial attention map $\mathbf{f}_{\mathbf{spatial}}(\mathbf{F}')$. As shown in (2) and (3), the input feature will be multiplied by $\mathbf{f}_{\mathbf{spatial}}$ and $\mathbf{f}_{\mathbf{channel}}$ to get the enhanced feature $\mathbf{F}''$.

IV. EXPERIMENT

This section, we will introduce the experiment datasets, experiments settings, and final experiment results.

TABLE III
QUANTITATIVE COMPARISON RESULTS (%) OF DIFFERENT METHODS ON THE HOUSTON DATA

Class	SVM(H)	SVM(H+L)	ELM(H)	ELM(H+L)	Co-CNN(H)	Co-CNN(H+L)	Proposed(H)	Proposed(H+L)
Health grass	97.87	97.87	98.53	98.42	82.91	83.10	82.62	83.10
Stressed grass	99.22	99.22	98.13	98.02	84.49	85.06	85.15	85.15
Synthetic grass	99.80	99.80	94.22	99.80	99.21	100.00	99.60	100.00
Trees	99.08	99.08	98.18	98.58	89.77	93.28	91.95	**99.34**
Soil	97.74	97.74	95.80	97.94	99.81	99.81	100.00	100.00
Water	12.83	12.82	43.17	27.15	85.31	84.62	87.41	90.21
Residual	74.84	74.93	47.52	27.15	85.31	84.62	87.41	90.21
Commercial	87.19	87.19	85.86	85.20	73.79	76.92	79.77	76.26
Road	76.87	76.87	80.07	84.00	77.24	82.63	89.24	77.24
Highway	78.73	78.73	61.33	65.63	43.15	64.29	58.11	68.53
Railway	82.66	82.74	76.49	77.21	85.96	85.10	92.88	78.27
Parking lot 1	87.86	87.86	80.89	83.54	90.30	90.20	83.29	96.93
Parking lot 2	54.85	54.85	47.62	48.17	2.81	4.56	5.61	**73.68**
Tennis court	82.27	82.27	88.77	89.74	99.60	98.79	100.00	100.00
Running track	99.57	99.57	99.79	99.79	86.05	99.79	98.31	99.58
OA	80.99	81.00	78.36	79.48	79.43	84.33	82.80	**85.54**
AA	83.39	83.39	81.25	82.23	82.70	86.95	84.04	**87.19**
Kappa	79.50	79.51	76.55	77.79	77.69	83.01	81.36	**84.36**

The highest value in each line can be noted as bold.

TABLE IV
QUANTITATIVE COMPARISON RESULTS (%) OF DIFFERENT METHODS ON THE TRENTO DATA

Class	SVM(H)	SVM(H+L)	ELM(H)	ELM(H+L)	Co-CNN(H)	Co-CNN(H+L)	Proposed(H)	Proposed(H+L)
Apple trees	64.84	64.82	99.54	64.96	99.54	99.44	98.49	99.03
Buildings	73.87	74.13	95.46	78.59	95.46	99.42	94.74	97.80
Ground	63.15	63.15	91.71	64.94	91.71	91.18	99.47	89.04
Woods	94.63	94.70	89.36	95.15	89.36	98.33	99.52	99.61
Vineyard	93.90	93.87	91.32	95.44	91.32	89.24	98.62	96.62
Roads	83.66	84.19	71.63	89.54	71.63	85.45	71.66	**93.51**
OA	80.43	80.53	87.03	81.49	87.03	92.01	94.28	**96.72**
AA	85.48	85.59	86.14	86.28	86.14	88.60	89.57	**94.89**
kappa	85.16	85.24	90.17	85.94	90.17	93.96	95.72	**97.54**

The highest value in each line can be noted as bold.

TABLE V
TRAINING TIME AND NUMBERS OF PARAMETERS RESULTS COMPARISON

Dataset	Houston(H+L)		Trento(H+L)	
Methods	Co-CNN	Proposed	Co-CNN	Proposed
Time (s)	403	1046	211	286
Parameters (MB)	37.9	36.7	24.9	16.6
Training epochs	80	80	13	13

A. Dataset Description

In this experiment, we have conducted our algorithm on Houston and Trento datasets, which contains LiDAR and HSI information, to evaluate the efficiency of the cascaded CNN and attention modules.

Houston dataset [49] is captured in Houston, USA. The dataset contains one air-borne HSI and LiDAR DSM data with 349×1905 pixels. The spatial resolution has been registered on both HSI and LiDAR DSM data with 2.5 m. The HSI contains sufficient spectral information with 144 bands, and the hyperspectral sensor CASI-1500 captures 0.38–1.05 μm spectral data.

Trento dataset [50] is composed of HSI and LiDAR DSM data captured in Trento, Italy. The registration image size is 600×166 with a 1-m spatial resolution. The hyperspectral data contain 63 bands ranging from 0.42 to 0.99 μm. The HSI and LiDAR data are separately captured by AISA Eagle and Optech ALTM 3100EA sensors.

The original multimodal remote sensing datasets are two images with tiff format containing the ground-truth label. To modify the raw data as the standard model input data, we have normalized standardization and recorded the location index information for the ground-truth samples.

B. Experimental Setting

To evaluate the efficiency of our proposed method and compared methods, we conducted experiments on Intel(R) Core(TM) i7-7700HQ CPU, GTX 1060(Ti) GPU, 16 GB of RAM, and Ubuntu 18.04 version under the same experimental conditions. We have conducted overall accuracy (OA), average accuracy (AA), and Kappa coefficient metrics to prove the algorithm's performance. Meanwhile, to ensure the reliability of experiments, all the experiment results are the average results of ten experiments with the same parameter settings.

Towards the multimodal datasets, we have randomly selected half amount of training samples as the validation data to help optimize the performance. When training the Houston dataset, we set the batch size as 100 and the training epoch as 80. While training the Trento dataset, we set the batch size as 100 and the training epoch as 13.

We have utilized the fine-tune strategy to improve the multimodal attention algorithm performance during the training stage. We have separately trained the hyperspectral CNN and

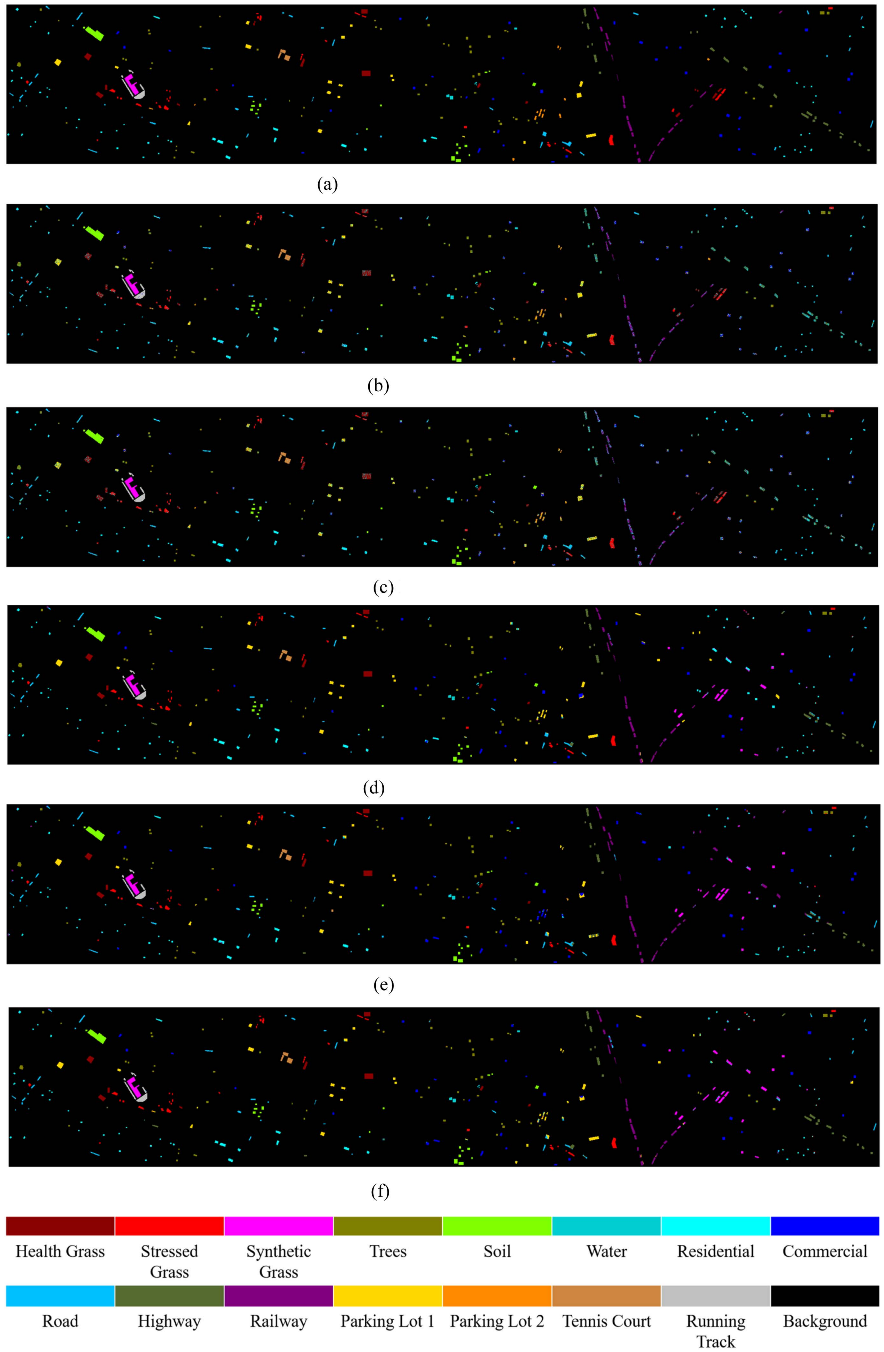

Fig. 8. Classification maps of various comparison algorithms for Houston datasets. (a) Visualization of used Houston test samples. (b) SVM (H+L). (c) ELM (H+L). (d) Co-CNN (H+L). (e) Proposed (H). (f) Proposed (H+L).

LiDAR cascaded CNN to save the trained model, then trained the multimodal neural network with initialization of the saved model. We have selected the Adam optimizer with 0.001 on training LiDAR data and 0.0001 on hyperspectral data. While conducting fine-tune training, we also choose Adam as an optimizer with a 0.001 learning rate. In case of overfitting the data, we design a 0.25 ratio dropout operation in the fusion stage. Other parameters has been listed in the framework.

C. Results and Analysis

This article compares the proposed method with classic machine learning methods, including SVM [32] and ELM [33]. Besides, we also introduced the fundamental Co-CNN [28] methods based on CNN to further prove the efficiency of proposed cascaded attention network. The final experiment results list in Table III and IV.

As shown in Tables III to V, under the same training epochs condition, although the proposed methods cost more training time, but achieve higher classification performance with nearly model parameters. The proposed methods have achieved better performance on OA, AA, and Kappa key metrics.

As the classification results shown in Figs. 8 and 9, the deep learning-based methods achieve better classification

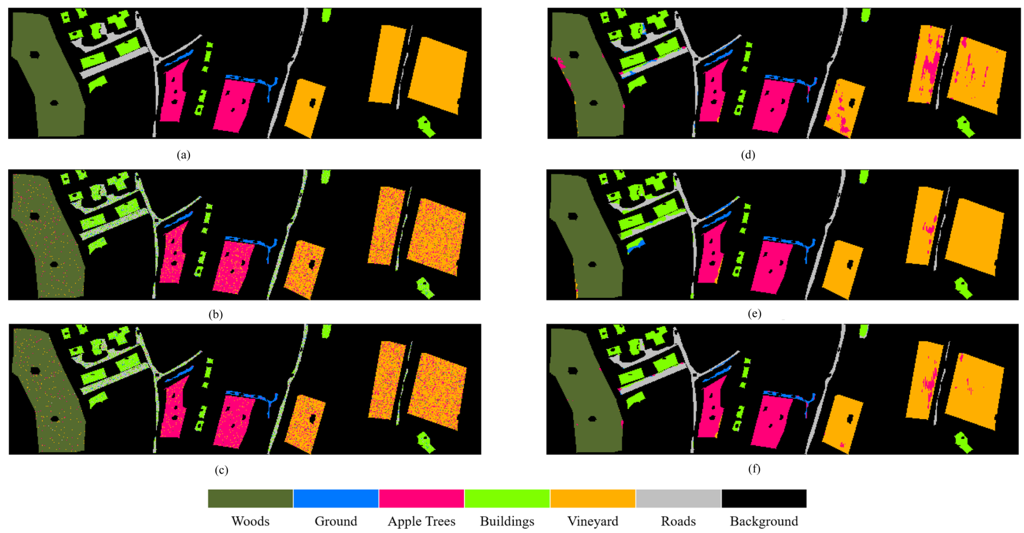

Fig. 9. Classification maps of various comparison algorithms for Trento dataset. (a) Visualization of used Trento test samples. (b) SVM (H+L). (c) ELM (H+L). (d) Co-CNN (H+L). (e) Proposed (H). (f) Proposed (H+L).

performance than classical machine learning methods on both datasets.

Our designed framework highlights the LiDAR ground objects' height information by utilizing an attention mechanism and cascaded multiscale network. For Houston data, it is clear that the multimodal data with the proposed method has better performance on trees and Parking lot 2, which can easily be predicted as similar health grass and Parking lot 1 ground object categories. For Trento data, our proposed method has achieved tremendous results on roads, with similar strong object spectral and spatial features with buildings owing to the sensors' overlook perspective. As shown in Fig. 9(a), (d) and (f), the Co-CNN method does not perform well (85.45% accuracy) in the class of the road, in which several pixels have been classified as buildings because of lacking a specific height LiDAR feature. Besides, roads are easily predicted as ground owing to a similar height between road and ground class. Our proposed methods focus both on LiDAR contextual spatial info by multiscale cascaded network and attention mechanism to enhance precious LiDAR info to achieve 93.51% accuracy.

V. Conclusion

In this article, our proposed multimodal attention-aware convolutional network has fully utilized the height feature of ground objects provided by LiDAR data, which has achieved outstanding performances on easily confusing categories and overall accuracy. We have also conducted classic machine learning methods (SVM and ELM) and deep learning-based methods to compare the efficiency of the proposed methods. Because of the substantial feature divergence between various modalities, the proposed methods have proved how to strengthen the feature derived from the source without sufficient original image information. In the future work, we will continue to explore more possibilities to narrow the feature diversity among multiple modalities to achieve better classification performance, not limited the feature augmentation, fusion methods, robustness evaluation, or higher efficient training strategy.

Acknowledgment

The authors would like to thank the 2013 IEEE GRSS Data Fusion Contest Committee and Prof. Bruzzone for providing Houston and Trento multimodal remote sensing datasets. They also thank the Associate Editor and the anonymous Reviewers for their professional and meaningful comments that greatly help to improve the quality of this work.

References

[1] Y. Han, C. Deng, B. Zhao, and B. Zhao, "Spatial-temporal context-aware tracking," *IEEE Signal Process. Lett.*, vol. 26, no. 3, pp. 500–504, Mar. 2019.

[2] J. Yao, D. Meng, Q. Zhao, W. Cao, and Z. Xu, "Nonconvex-sparsity and nonlocal-smoothness-based blind hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2991–3006, Jun. 2019.

[3] Y. Han, C. Deng, B. Zhao, and D. Tao, "State-aware anti-drift object tracking," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 4075–4086, Aug. 2019.

[4] J. Li, K. Zheng, J. Yao, L. Gao, and D. Hong, "Deep unsupervised blind hyperspectral and multispectral data fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[5] J. Liet al., "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 112, 2022, Art. no. 102926.

[6] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: A review and future directions," *Proc. IEEE*, vol. 103, no. 9, pp. 1560–1584, Sep. 2015.

[7] B. Zhao, Y. Han, H. Wang, L. Tang, X. Liu, and T. Wang, "Robust shadow tracking for video SAR," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 821–825, May 2021.

[8] D. Honget al., "Interpretable hyperspectral artificial intelligence: When nonconvex modeling meets hyperspectral remote sensing," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 2, pp. 52–87, Jun. 2021.

[9] J. Yao, D. Hong, L. Xu, D. Meng, J. Chanussot, and Z. Xu, "Sparsity-enhanced convolutional decomposition: A novel tensor-based paradigm for blind hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022, Art no. 5505014, doi: 10.1109/TGRS.2021.3069845.

[10] B. Zhang, S. Li, X. Jia, L. Gao, and M. Peng, “Adaptive Markov random field approach for classification of hyperspectral imagery,” *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 5, pp. 973–977, Sep. 2011.
[11] B. Zhang, X. Sun, L. Gao, and L. Yang, “Endmember extraction of hyperspectral remote sensing images based on the ant colony optimization (ACO) algorithm,” *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 7, pp. 2635–2646, Jul. 2011.
[12] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, “An augmented linear mixing model to address spectral variability for hyperspectral unmixing,” *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.
[13] B. Zhang, W. Yang, L. Gao, and D. Chen, “Real-time target detection in hyperspectral images based on spatial-spectral information extraction,” *EURASIP J. Adv. Signal Process.*, vol. 2012, no. 1, 2012, Art. no. 142.
[14] L. Gaoet al., “Subspace-based support vector machines for hyperspectral image classification,” *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 2, pp. 349–353, Feb. 2015.
[15] H. Yu, L. Gao, W. Liao, B. Zhang, A. Pizurica, and W. Philips, “Multiscale superpixel-level subspace-based support vector machines for hyperspectral image classification,” *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 11, pp. 2142–2146, Nov. 2017.
[16] R. Hang, Q. Liu, D. Hong, and P. Ghamisi, “Cascaded recurrent neural networks for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5384–5394, Aug. 2019.
[17] D. Hong, X. Wu, P. Ghamisi, J. Chanussot, N. Yokoya, and X. X. Zhu, “Invariant attribute profiles: A spatial-frequency joint feature extractor for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 3791–3808, Jun. 2020.
[18] R. Hang, F. Zhou, Q. Liu, and P. Ghamisi, “Classification of hyperspectral images via multitask generative adversarial networks,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1424–1436, Feb. 2021.
[19] R. Hang, Z. Li, Q. Liu, P. Ghamisi, and S. S. Bhattacharyya, “Hyperspectral image classification with attention-aided CNNs,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2281–2293, Mar. 2021.
[20] R. Hang, Q. Liu, and Z. Li, “Spectral super-resolution network guided by intrinsic properties of hyperspectral imagery,” *IEEE Trans. Image Process.*, vol. 30, pp. 7256–7265, 2021.
[21] D. Hong, J. Hu, J. Yao, J. Chanussot, and X. X. Zhu, “Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 178, pp. 68–80, 2021.
[22] P. Ghamisi, J. A. Benediktsson, and S. Phinn, “Land-cover classification using both hyperspectral and lidar data,” *Int. J. Image Data Fusion*, vol. 6, no. 3, pp. 189–215, 2015.
[23] M. Pedergnana, P. R. Marpu, M. Dalla Mura, J. A. Benediktsson, and L. Bruzzone, “A novel technique for optimal feature selection in attribute profiles based on genetic algorithms,” *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 6, pp. 3514–3528, Jun. 2013.
[24] P. Ghamisi, R. Souza, J. A. Benediktsson, X. X. Zhu, L. Rittner, and R. A. Lotufo, “Extinction profiles for the classification of remote sensing data,” *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 5631–5645, Oct. 2016.
[25] P. Ghamisi, B. Höfle, and X. X. Zhu, “Hyperspectral and LiDAR data fusion using extinction profiles and deep convolutional neural network,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 6, pp. 3011–3024, Jun. 2017.
[26] D. Honget al., “SpectralFormer: Rethinking hyperspectral image classification with transformers,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022, Art no. 5518615, doi: 10.1109/TGRS.2021.3130716.
[27] D. Hong, L. Gao, R. Hang, B. Zhang, and J. Chanussot, “Deep encoder-decoder networks for classification of hyperspectral and LiDAR data,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[28] 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.
[29] D. Hong, N. Yokoya, G.-S. Xia, J. Chanussot, and X. X. Zhu, “X-modalNet: A semi-supervised deep cross-modal network for classification of remote sensing data,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 167, pp. 12–23, 2020.
[30] X. Wu, D. Hong, and J. Chanussot, “Convolutional neural networks for multimodal remote sensing data classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–10, 2022.
[31] Y. Chen, C. Li, P. Ghamisi, X. Jia, and Y. Gu, “Deep fusion of remote sensing data for accurate classification,” *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1253–1257, Aug. 2017.
[32] C. Cortes and V. Vapnik, “Support vector machine,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
[33] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, “Extreme learning machine: A new learning scheme of feedforward neural networks,” in *Proc. IEEE Int. Joint Conf. neural Netw.*, 2004, pp. 985–990.
[34] C. Deng, Y. Han, and B. Zhao, “High-performance visual tracking with extreme learning machine framework,” *IEEE Trans. Cybern.*, vol. 50, no. 6, pp. 2781–2792, Jun. 2020.
[35] D. Hong, L. Gao, J. Yao, B. Zhang, P. Antonio, and J. Chanussot, “Graph convolutional networks for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2021.
[36] A. Vaswaniet al., “Attention is all you need,” in *Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[37] K. Xuet al., “Show, attend and tell: Neural image caption generation with visual attention,” in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.
[38] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.
[39] Q. Wang, S. Liu, J. Chanussot, and X. Li, “Scene classification with recurrent attention of VHR remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2019.
[40] Q. Zhanget al., “Dense attention fluid network for salient object detection in optical remote sensing images,” *IEEE Trans. Image Process.*, vol. 30, pp. 1305–1317, 2021.
[41] B. Fang, Y. Li, H. Zhang, and J. C.-W. Chan, “Hyperspectral images classification based on dense convolutional networks with spectral-wise attention mechanism,” *Remote Sens.*, vol. 11, no. 2, 2019, Art. no. 159.
[42] X. Panet al., “Building extraction from high-resolution aerial imagery using a generative adversarial network with spatial and channel attention mechanisms,” *Remote Sens.*, vol. 11, no. 8, 2019, Art. no. 917.
[43] Z. Han, D. Hong, L. Gao, J. Yao, B. Zhang, and J. Chanussot, “Multimodal hyperspectral unmixing: Insights from attention networks,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.
[44] Y. Han, C. Deng, Z. Zhang, J. Li, and B. Zhao, “Adaptive feature representation for visual tracking,” in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 1867–1870.
[45] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, “A survey of deep neural network architectures and their applications,” *Neurocomputing*, vol. 234, no. 234, pp. 11–26, 2017.
[46] J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and J. Li, “Visual attention-driven hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 8065–8080, Oct. 2019.
[47] J. Yao, D. Hong, J. Chanussot, D. Meng, X. Zhu, and Z. Xu, “Cross-attention in coupled unmixing nets for unsupervised hyperspectral super-resolution,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 208–224.
[48] X. Meiet al., “Spectral-spatial attention networks for hyperspectral image classification,” *Remote Sens.*, vol. 11, no. 8, 2019, Art. no. 963.
[49] M. Khodadadzadeh, J. Li, S. Prasad, and A. Plaza, “Fusion of hyperspectral and LiDAR remote sensing data using multiple feature learning,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2971–2983, Jun. 2015.
[50] B. Rasti, P. Ghamisi, and R. Gloaguen, “Hyperspectral and LiDAR fusion using extinction profiles and total variation component analysis,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3997–4007, Jul. 2017.

Haotian Zhang received the B.E. degree in remote sensing and technology from the China University of Geosciences, Beijing, China, in 2019. He is currently working toward the M.S. degree in cartography and geographical information system with the University of Chinese Academy of Sciences, Beijing.

His research interests include remote sensing and multimodal machine learning.

Jing Yao (Member, IEEE) received the B.Sc. degree in applied mathematics from Northwest University, Xi'an, China, in 2014, and the Ph.D. degree in mathematics from Xi'an Jiaotong University, Xi'an, in 2021.

He is currently an Assistant Professor with the Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. From 2019 to 2020, he was a visiting student with Signal Processing in Earth Observation, Technical University of Munich, Munich, Germany, and at the Remote Sensing Technology Institute, German Aerospace Center, Oberpfaffenhofen, Germany. His research interests include low-rank modeling, hyperspectral image analysis, and deep learning-based image processing methods.

Dr. Yao was a recipient of the Jose Bioucas Dias Award for recognizing the outstanding paper at the Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing in 2021. He also serves as a Guest Editor of Remote Sensing.

Li Ni received the Ph.D. degree in cartography and geographical information system from the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China, in 2015.

She is currently an Associate Researcher with the Aerospace Information Research Institute, Chinese Academy of Sciences. Her research interests include the land surface temperature retrieval and application of hyperspectral remote sensing.

Lianru Gao (Senior Member, IEEE) received the B.S. degree in civil engineering from Tsinghua University, Beijing, China, in 2002, and the Ph.D. degree in cartography and geographic information system from Institute of Remote Sensing Applications, Chinese Academy of Sciences (CAS), Beijing, in 2007.

He is currently a Professor with the Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, CAS. He also has been a Visiting Scholar with the University of Extremadura, Cáceres, Spain, in 2014, and at the Mississippi State University, Starkville, USA, in 2016. In last ten years, he was the PI of ten scientific research projects at national and ministerial levels, including projects by the National Natural Science Foundation of China (2016–2019, 2018–2020, 2022–2025), and by the National Key R&D Program of China (2021–2025) et al. He has authored or coauthored more than 200 peer-reviewed papers, and there are more than 120 journal papers included by Science Citation Index (SCI). He was a coauthor of three academic books including "Hyperspectral Image Information Extraction" et al. He received 29 National Invention Patents in China. His research focuses on hyperspectral image processing and information extraction.

Dr. Gao was a recipient of the Outstanding Science and Technology Achievement Prize of the CAS in 2016, and was supported by the China National Science Fund for Excellent Young Scholars in 2017, and won the Second Prize of The State Scientific and Technological Progress Award in 2018. He received the recognition of the Best Reviewers of the IEEE JSTARS in 2015, and the Best Reviewers of the IEEE TGRS in 2017.

Min Huang received the B.S. degree in optical engineering from University of Science and Technology of China, Hefei, China, in 1999, and the Ph.D. degree in optical engineering from the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences (CAS), Xi'an, China, in 2009.

He is currently a Professor with the Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, CAS. In last ten years, he was the PI of ten scientific research projects at national and ministerial levels, including projects by the Key Research Program of CAS's Support Technology, and by the National High-tech R&D Program. He has authored or coauthored more than 30 peer-reviewed papers and received more than 20 National Invention Patents in China. His research focuses on imaging spectrometry and computational optical imaging.

Dr. Huang was a recipient of the Second Prize of The State Scientific and Technological Progress Award in 2010.