EMFNet: Enhanced Multisource Fusion Network for Land Cover Classification

Chengxiang Li, Renlong Hang, *Member, IEEE*, and Behnood Rasti, *Senior Member, IEEE*

***Abstract*—Feature extraction and fusion are two critical issues for the task of multisource classification. In this article, we propose an enhanced multisource fusion network (EMFNet) to address them in an end-to-end framework. Specifically, two convolutional neural networks are employed to extract features from two different sources. Each network is mainly comprised of three convolutional layers. For each convolutional layer, feature tuning modules are designed to enhance the extracted feature of one source by taking advantage of the other source. After getting the features of two sources, a weighted summation method is used to fuse them. Considering that fusion weights should vary for different inputs, a feature fusion module is designed to achieve this goal. In order to test the performance of our proposed EMFNet, we compare it with state-of-the-art fusion models, including the traditional models and the deep-learning-based models, on two real datasets. Experimental results show that the EMFNet can achieve competitive classification results in comparison with them.**

***Index Terms*—Convolutional neural network (CNN), feature fusion module, feature tuning module, multisource fusion.**

I. Introduction

ACCURATE and up-to-date land cover classifications are fundamental and important for a variety of applications. In comparison with ground surveys, remote sensing techniques are capable of providing a larger view and faster acquisition of land cover information, thus becoming a dominant tool for land cover classification. With the development of imaging technologies, different kinds of remote sensors have been applied, providing different observation views of land covers. Typically, hyperspectral sensors record rich spectral information of land covers through the visible bands to the infrared bands. In contrast to the passive sensing of hyperspectral sensors, light detection and ranging (LiDAR) is an active sensing method. It adopts laser light as an illumination source and provides the height and shape information of the land surface.

Manuscript received March 6, 2021; revised March 19, 2021; accepted April 2, 2021. Date of publication April 16, 2021; date of current version May 6, 2021. This work was supported in part by the Natural Science Foundation of China under Grant 61906096 and Grant 61802199, and in part by the Natural Science Foundation of Jiangsu Province, China under Grant BK20180786. *(Corresponding author: Renlong Hang.)*

Chengxiang Li and Renlong Hang are with the Jiangsu Key Laboratory of Big Data Analysis Technology, Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: chengxiangli@nuist.edu.cn; renlong_hang@163.com).

Behnood Rasti is with the Helmholtz-Zentrum Dresden-Rossendorf, Helmholtz Institute Freiberg for Resource Technology, 09599 Freiberg, Germany (e-mail: b.rasti@hzdr.de).

Digital Object Identifier 10.1109/JSTARS.2021.3073719

Since the spectral properties of different materials vary, hyperspectral data are effective to discriminate objects with different materials, which have been widely used in land cover classification [1]–[3]. However, if two different objects are comprised of the same material, e.g., the building roof and the road, the individual use of hyperspectral data is difficult to differentiate them. Similarly, LiDAR is powerful enough to separate objects with different elevations, but cannot distinguish different objects with the same elevation, e.g., asphalt and concrete roads. In order to cope with their respective shortcomings, an intuitive method is fusing the complementary information from hyperspectral and LiDAR data [4]–[6]. To achieve this goal, one needs to answer two critical issues. The first one is feature extraction. How to effectively represent hyperspectral and LiDAR data is the precondition of the fusion problem. The second one is how to combine the multisource information based on the extracted features.

During the past few decades, a lot of methods have been proposed to deal with the first issue. For example, as an extension of a morphological profile, an attribute profile (AP) has been adopted to extract spatial features from both hyperspectral and LiDAR data in [7] and [8]. Due to the high dimensionality of hyperspectral data, dimensionality reduction is needed to acquire some principal components before extracting APs. To further improve the capability of the AP, the extinction profile was proposed [9] and employed to hyperspectral and LiDAR data [10]–[12]. In addition to these morphology-related features, a local binary pattern was also used to extract features from both sources [13]. In recent years, the success of deep learning [14]–[17] has motivated its applications to the field of multisource fusion. Typical models include convolutional neural network (CNN) [18]–[21], autoencoder [22], [23], and recurrent neural network [24], [25]. In [18], Chen *et al.* proposed a two-branch fusion framework based on 2-D CNNs. One CNN was used to extract convolutional features for hyperspectral data, and the other was for LiDAR data. Differently, Xu *et al.* [19] simultaneously employed the 1-D CNN and the 2-D CNN to extract spectral and spatial features from hyperspectral data, respectively, thus achieving better representations for hyperspectral data. In [23], Hong *et al.* designed two autoencoders to, respectively, process hyperspectral and LiDAR data. Compared to the CNN-related models, the training of autoencoders is based on the reconstruction strategy, alleviating the requirement of labeling samples. To sum up, *most of the existing models attempt to extract handcrafted or deep features from different sources*

(e.g., hyperspectral and LiDAR) separately, while ignoring their complementary information. We argue that such information will enforce the discriminative ability of extracted features if fully explored.

For the second issue, decision-level fusion and feature-level fusion are two popular strategies to combine multisource information. The former one aims to derive a classification result from each source and then integrate them to get the final result. The key is selecting proper classifiers and integration strategies. In [26], support vector machine (SVM) was adopted as classifiers for each source, and a weighted summation strategy was employed to combine classification results. The combination weights were determined by the classification accuracies of each source. In [27], Zhong *et al.* used three different classifiers, including SVM, maximum likelihood classifier, and multinomial logistic regression, to classify the extracted features. Weighted voting was applied to obtain the final result, and the weights of each classification map were optimized by a differential evolution algorithm. In [13], the derived features of each source were classified by a collaborative representation-based classifier with Tikhonov regularization. Different from decision-level fusion, the purpose of feature-level fusion is to integrate features of different sources. Concatenation is a simple but efficient fusion method, which has been widely applied [7], [18], [19], [28]. To address the high-dimensionality issue induced by concatenation, Liao *et al.* [29] proposed a generalized graph model. Besides concatenation, a multiple-kernel learning model was constructed to integrate heterogeneous features in [30]. The weights of the kernels with different features were determined by finding a projection based on the maximum variance. In [31], canonical correlation analysis was adopted as a basic unit for fusing features. *Although a lot of works have been developed to combine the multisource information, most of them use fixed combination weights for different sources. We argue that the weights should vary for discriminating different objects.* For instance, larger weights should be given to LiDAR data when classifying the building roof and the road, both of which are comprised of concrete; on the contrary, hyperspectral data should have larger weights when discriminating roads made up of asphalt and concrete.

To address the above two issues, we propose an enhanced multisource fusion network (EMFNet), which adopts a two-branch framework. The basic structure of each branch is CNN, consisting of three convolutional layers. Different from the traditional CNNs, feature tuning modules are designed to enhance the convolutional features by taking advantage of the complementary information between two sources. Specifically, they are capable of using the channel features of the hyperspectral branch to enhance the convolutional features of LiDAR data, while using the spatial features of the LiDAR branch to enhance the features of hyperspectral data. When the third-layer convolution features are acquired from two sources, they will be fused in a weighted summation manner. The fusion weights are decided by a feature fusion module, whose inputs are the first-layer convolution features of two sources. The convolutional layers, feature tuning modules, and the feature fusion module are finally combined to construct an end-to-end network for multisource fusion. The major contributions of this article are summarized as follows.

1) To make full use of the complementary information between two sources during feature extraction, feature tuning modules are embedded into CNNs for enhancing the discriminative ability of each convolutional feature.
2) In order to sufficiently consider the contribution of each source during multisource information fusion, a feature fusion module is designed to dynamically allocate fusion weights according to the input objects.
3) Extensive experiments are conducted on two real hyperspectral and LiDAR data. The comparisons with state-of-the-art traditional and deep-learning-based models certify the effectiveness of our proposed model.

The rest of this article is organized as follows. Section II first presents the framework of our proposed method and then describes the detail of the feature tuning module and the feature fusion module. Section III introduces the experimental datasets and results in comparison with state-of-the-art models. Section IV concludes this article.

II. Methodology

The framework of our proposed EMFNet is shown in Fig.1, where "Conv i" represents the i th convolutional layer. The inputs of the EMFNet are comprised of hyperspectral data and LiDAR data. Considering the redundancy of high-dimensional spectral information in hyperspectral data, principal component analysis is adopted to extract h principal components. For a given pixel, a small cube $\mathbf{X}_H \in \Re^{w\times w\times h}$ centered at it is cropped from the dimensionality reduced hyperspectral data, where w is the selected cube size. In comparison with hyperspectral data, the LiDAR-derived digital surface model is relatively easier, which often contains a single channel. Therefore, we can directly crop a small patch $\mathbf{X}_L \in \Re^{w\times w}$ centered at the same position.

After preprocessing, $\mathbf{X}_H$ and $\mathbf{X}_L$ are separately fed into three convolutional layers to learn hyperspectral and LiDAR features. The kernel size of each convolutional layer is 3×3, while the channel sizes from the first to the third layer are set to 32, 64, and 128, respectively. Assume the output of the $(i+1)$th layer is $\mathbf{F}_j^{i+1}, i \in \{0,1,2\}, j \in \{H, L\}$; the process of feature learning can be expressed as

$$\mathbf{F}_j^{i+1} = \text{Conv}(\mathbf{F}_j^i) \tag{1}$$

where "Conv" denotes the convolutional operator, $\mathbf{F}_H^0$ equals $\mathbf{X}_H$, and $\mathbf{F}_L^0$ equals $\mathbf{X}_L$.

It is well known that hyperspectral data have abundant spectral information, leading to better channel features in $\mathbf{F}_H^{i+1}$ than that in $\mathbf{F}_L^{i+1}$. On the contrary, LiDAR data contain rich height and shape information, making $\mathbf{F}_L^{i+1}$ have discriminative spatial features. Inspired by such information, we attempt designing a feature tuning module to enhance the channel features of $\mathbf{F}_L^{i+1}$ by exploring $\mathbf{F}_H^{i+1}$, while enhancing the spatial features of $\mathbf{F}_H^{i+1}$ by taking advantage of $\mathbf{F}_L^{i+1}$. Therefore, (1) can be reformulated as

$$\mathbf{F}_j^{i+1} = \text{FT}(\text{Conv}(\mathbf{F}_j^i)) \tag{2}$$

where "FT" refers to the feature tuning operator.

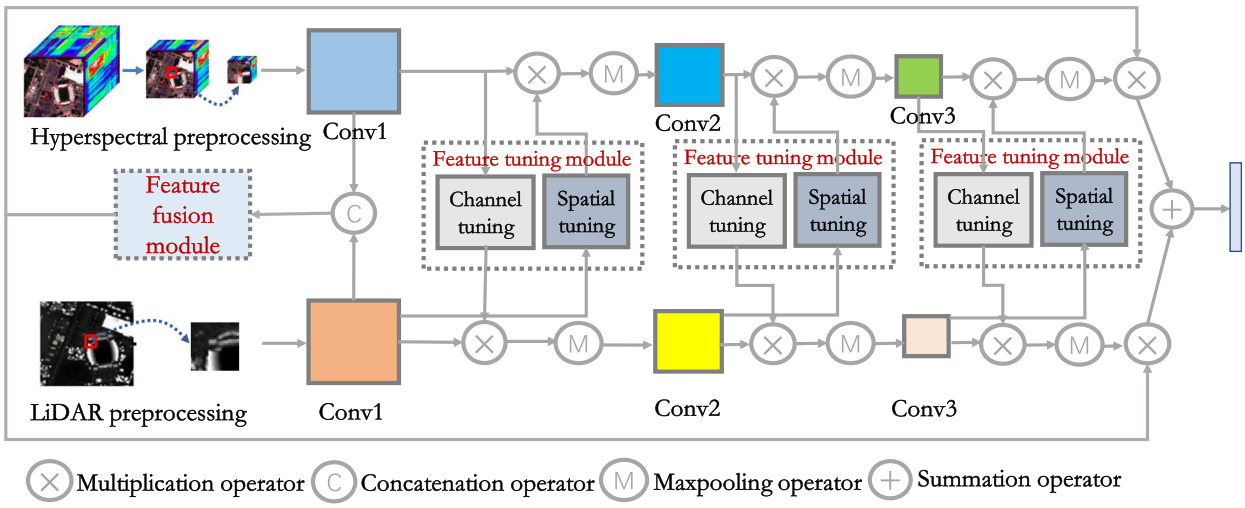

Fig. 1. Flowchart of the proposed network.

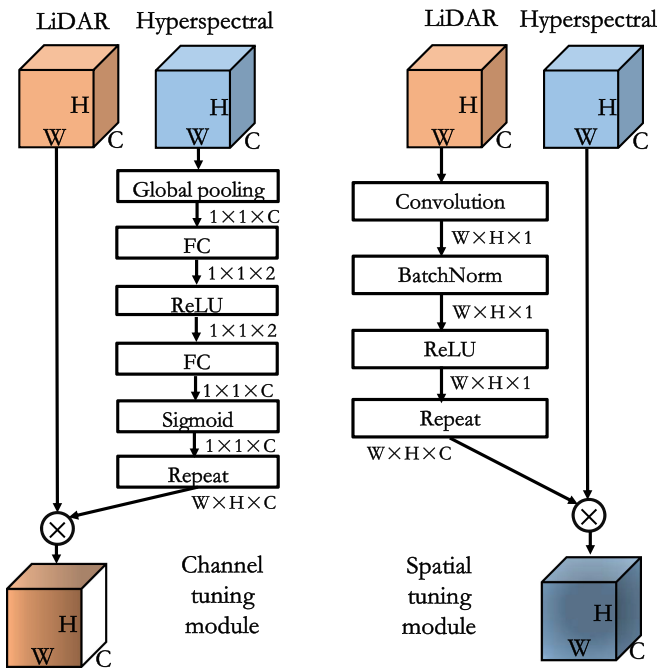

Fig. 2. Detail of our proposed feature tuning module. The left figure is the channel tuning module, while the right figure is the spatial tuning module. Note that "FC" represents the fully connected operators, "BatchNorm" denotes the batch normalization operator, and "⊗" is the multiplication operator.

A. Feature Tuning Module

Fig. 2 demonstrates the detailed structure of our proposed feature tuning module, which consists of a channel tuning module and a spatial tuning module. For the ith convolutional layer, a global pooling operator, shown in the left part of Fig. 2 is first adopted to squeeze the spatial dimension of $\mathbf{F}_H^i \in \Re^{W\times H\times C}$ to 1×1, because the channel tuning module mainly aims to make use of the channel features of $\mathbf{F}_H^i$. After that, two fully connected layers are sequentially used to $\mathbf{F}_H^i$, which can be formulated as

$$\mathbf{A}_H = f_2(\mathbf{W}_2 * f_1(\mathbf{W}_1 * \mathbf{F}_H^i)) \tag{3}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are the connection weights of the first and second fully connected layers, respectively. f_1 and f_2 correspond to the activation function ReLU and Sigmoid, respectively. The number of neurons in these two fully connected layers is empirically set to 2 and C, respectively. Finally, $\mathbf{A}_H \in \Re^{1\times 1\times C}$ is applied to refine the channel information of $\mathbf{F}_L^i$ as follows:

$$\mathbf{F}_L^i \leftarrow \mathbf{F}_L^i \odot \mathrm{Rep}(\mathbf{A}_H) \tag{4}$$

where $\mathrm{Rep}(\mathbf{A}_H)$ refers to repeating values in $\mathbf{A}_H$ along spatial dimensions, and "$\odot$" denotes elementwise multiplication.

In contrast to the channel tuning module, the spatial tuning module aims at taking advantage of the spatial features of $\mathbf{F}_L^i$. As shown in the right part of Fig. 2, The channel tuning module contains a 1×1 convolutional layer, a batch normalization layer, and an activation layer. The convolutional layer can reduce the channel dimension of $\mathbf{F}_L^i$ to 1, making it focus on the exploration of spatial features. After the channel tuning module, $\mathbf{F}_H^i$ is refined as

$$\mathbf{F}_H^i \leftarrow \mathbf{F}_H^i \odot \mathrm{Rep}(\mathbf{A}_L) \tag{5}$$

where $\mathbf{A}_L \in \Re^{W\times H\times 1}$ is the final result after the ReLU layer, $\mathrm{Rep}(\mathbf{A}_L)$ represents repeating values in $\mathbf{A}_L$ along the channel dimension, and "$\odot$" denotes elementwise multiplication.

B. Feature Fusion Module

The feature tuning module allows us to acquire an enhanced feature representation of hyperspectral data and LiDAR data, but how to combine them is also another key issue for multisource fusion. In this article, we adopt $\mathbf{F}_H^3$ and $\mathbf{F}_L^3$ as the final feature for hyperspectral and LiDAR data, respectively, and design a feature fusion module to adaptively combine them together. Fig. 3 shows the detailed structure of our proposed feature fusion module, whose inputs are $\mathbf{F}_H^1$ and $\mathbf{F}_L^1$. These two inputs are first concatenated along the channel dimension, followed by a global pooling layer. Then, two fully connected layers are employed to

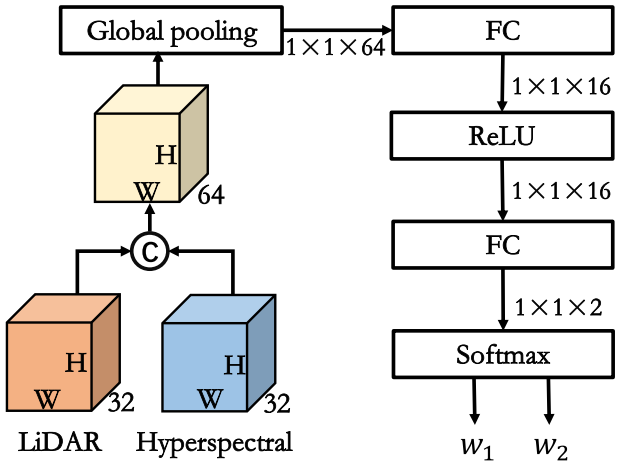

Fig. 3. Detail of our proposed feature fusion module. Note that FC represents the fully connected operators and "ⓒ" is the concatenation operator.

get the final result. The whole process can be described as

$$\mathbf{w} = f^2(\mathbf{W}^2 * (f^1(\mathbf{W}^1 * GP(\mathbf{F}_H^3 \text{ⓒ} \mathbf{F}_L^3)))) \tag{6}$$

where $\mathbf{w} \in \Re^2$ is the output of the feature fusion module, "ⓒ" represents the concatenation operator, and "GP" is the concatenation operator. $\mathbf{W}^1$ and $\mathbf{W}^2$ are the connection weights of the first and second fully connected layers, respectively. f^1 and f^2 correspond to the activation function ReLU and Softmax, respectively. Since the expected number of fusion weights is 2, the second fully connected layer contains two neurons. For the first fully connected layer, we empirically set its neurons to 16. Once the fusion weight $\mathbf{w}$ is obtained, we can derive the fusion result as

$$\mathbf{F} = \mathbf{F}_H^3 * w_1 + \mathbf{F}_L^3 * w_2 \tag{7}$$

where w_1 and w_2 are the two elements of $\mathbf{w}$. It is worth noting that $\mathbf{w}$ is dependent on the input samples. In other words, for different hyperspectral and LiDAR inputs, their fusion weights are different, which is more reasonable in real applications.

III. Experiments

A. Data Description and Experimental Setup

In order to test the performance of our proposed EMFNet, we conduct comprehensive experiments on two hyperspectral and LiDAR fusion datasets. The first dataset contains *Houston* data, acquired over the University of Houston campus and the neighboring urban area in June 2012 [4]. The spatial size of both hyperspectral and LiDAR data is 349 × 1905 pixels, but not all of them are used in the experiments. Table I reports the specific number of training and test pixels in 15 classes. For the hyperspectral data, there exist 144 spectral bands. In Fig. 4, we show a pseudocolor image of the hyperspectral data using bands 64, 43, and 22, a grayscale image of the LiDAR data, and ground-truth maps of the training and test pixels. Note that the effect of cloud shadows in the Houston hyperspectral data was detected using the thresholding of illumination distributions calculated by the corresponding spectra. In this context, relatively small structures in the thresholded illumination map were eliminated w.r.t. the assumption that cloud shadows are larger than structures on the ground.[1]

TABLE I
Numbers of Training and Test Samples in Each Class of the Houston Data

Class No.	Class Name	Training	Test
1	Healthy grass	198	1053
2	Stressed grass	190	1064
3	Synthetic grass	192	505
4	Tree	188	1056
5	Soil	186	1056
6	Water	182	143
7	Residential	196	1072
8	Commercial	191	1053
9	Road	193	1059
10	Highway	191	1036
11	Railway	181	1054
12	Parking lot 1	192	1041
13	Parking lot 2	184	285
14	Tennis court	181	247
15	Running track	187	473
-	Total	2832	12197

TABLE II
Numbers of Training and Test Samples in Each Class of the Trento Data

Class No.	Class Name	Training	Test
1	Apple trees	129	3905
2	Buildings	125	2778
3	Ground	105	374
4	Wood	154	8969
5	Vineyard	184	10317
6	Roads	122	3252
-	Total	819	29595

The second dataset contains *Trento* data, which were captured over a rural area in the south of Trento, Italy. The hyperspectral data were acquired by the AISA Eagle sensor with 63 spectral bands, while the LiDAR data were acquired by the Optech ALTM 3100EA sensor. In comparison with the Houston dataset, this dataset has a smaller spatial size (i.e., 166 × 600 pixels). The total number of classes is 6. Table II lists the specific number of pixels in each class. Fig. 5 visualizes the hyperspectral data using bands 40, 20, and 10, the LiDAR data, and the distributions of the training and test pixels.

On these two datasets, we compare the EMFNet with state-of-the-art fusion models, including the traditional models and the deep-learning-based models. Some descriptions about them are as follows.

1) *SVM:* It adopts the fused hyperspectral data and LiDAR data as input and SVM as the classifier. Since the work in [32] also did the same experiment, it is reasonable to cite the results from [32].
2) *CHOTF:* It is a coupled higher order tensor factorization (CHOTF) model for hyperspectral and LiDAR fusion proposed in [33].

[1]The enhanced dataset was provided by Prof. Naoto Yokoya from RIKEN, Japan.

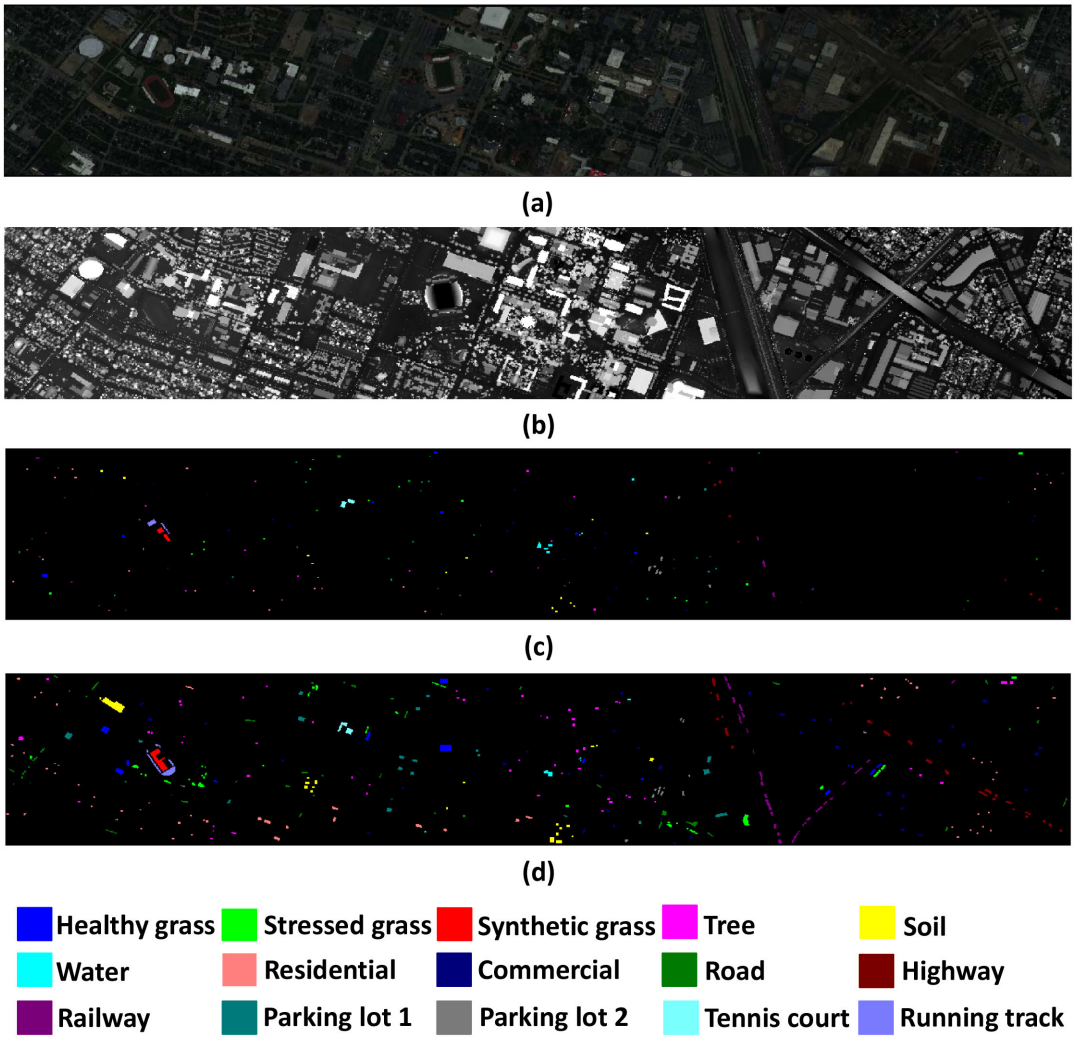

Fig. 4. Visualization of the Houston data. (a) Pseudocolor image for the hyperspectral data using 64, 43, and 22 as R, G, and B, respectively. (b) Grayscale image for the LiDAR data. (c) Training data map. (d) Test data map.

3) *CNN(H):* It is a three-layer CNN whose input is hyperspectral data only. The spatial size of all convolutional kernels is 3×3. The number of channels from the first to the third convolutional layers is set to 32, 64, and 128, respectively.
4) *CNN(L):* This network has the same structure as that of CNN(H), but with LiDAR data as input.
5) *EMFNet-FF-FT:* This network has the same structure as the EMFNet, but without using feature tuning modules in Fig. 2 and the feature fusion module in Fig. 3.
6) *EMFNet-FF-ST:* This network has the same structure as the EMFNet, but without using the spatial tuning modules in the right part of Fig. 2 and the feature fusion module in Fig. 3.
7) *EMFNet-FF:* This network has the same structure as the EMFNet, but without using feature tuning modules in Fig. 2.

All of the deep learning models are implemented in the PyTorch framework, and the Adam algorithm is chosen as the optimizer. The same as in [20], 11×11 patches are used as inputs for them. The classification performance of each model is evaluated by the overall accuracy (OA), the average accuracy (AA), the per-class accuracy, and the Kappa coefficient. The OA defines the ratio between the number of correctly classified pixels and the total number of pixels in the test set, the AA refers to the average of accuracies in all classes, and Kappa is the percentage of agreement corrected by the number of agreements that would be expected purely by chance.

B. Experimental Results

Table III compares the classification performance of different models on Houston data in terms of each class accuracies, OA, AA, and Kappa values. From this table, one can observe that CNN(H) achieves significantly higher OA, AA, and Kappa values than CNN(L), because hyperspectral data provide richer discriminant information than LiDAR data. In comparison with CNN(H), EMFNet-FF-FT, which combines hyperspectral data and LiDAR data together in a parallel framework, can improve OA, AA, and Kappa values. This conclusion certifies that hyperspectral data and LiDAR data contain complementary information, which is beneficial to the classification task. It is

TABLE III
CLASSIFICATION PERFORMANCE OF DIFFERENT FUSION MODELS APPLIED ON THE HOUSTON DATASET

Class No.	SVM	CHOTF	CNN(H)	CNN(L)	EMFNet-FF-FT	EMFNet-FF-ST	EMFNet-FF	EMFNet
1	82.43	83.00	82.91	60.30	83.29	**88.22**	83.10	83.57
2	82.05	95.68	99.91	24.34	99.44	99.25	**100.00**	99.62
3	99.80	**100.00**	91.29	66.53	99.60	98.02	**100.00**	98.02
4	92.80	95.83	95.93	88.73	98.48	99.72	99.91	**100.00**
5	98.48	99.91	**100.00**	24.81	99.81	**100.00**	**100.00**	**100.00**
6	95.10	95.10	93.71	25.87	97.20	**98.60**	95.80	95.80
7	75.47	89.93	91.60	61.19	94.40	95.15	96.74	**97.01**
8	46.91	82.43	87.18	84.33	91.45	94.59	95.54	**95.73**
9	77.53	94.43	86.87	40.32	85.55	86.50	**96.51**	91.88
10	60.04	68.24	**97.59**	53.86	96.81	95.17	95.95	96.24
11	81.02	99.15	89.56	80.46	95.54	94.21	98.10	**99.91**
12	85.49	96.06	91.16	29.30	94.72	95.29	89.72	**96.83**
13	75.09	80.70	**88.77**	81.05	88.07	84.91	87.02	83.51
14	**100.00**	99.60	89.07	52.63	91.09	89.88	98.79	**100.00**
15	98.31	98.94	90.91	29.81	97.67	99.58	99.37	**100.00**
OA	80.49	91.24	92.05	54.52	94.17	94.84	95.77	**96.10**
AA	83.37	91.93	91.76	53.57	94.21	94.61	95.77	**95.87**
Kappa	78.98	90.50	91.36	50.82	93.67	94.40	95.41	**95.76**

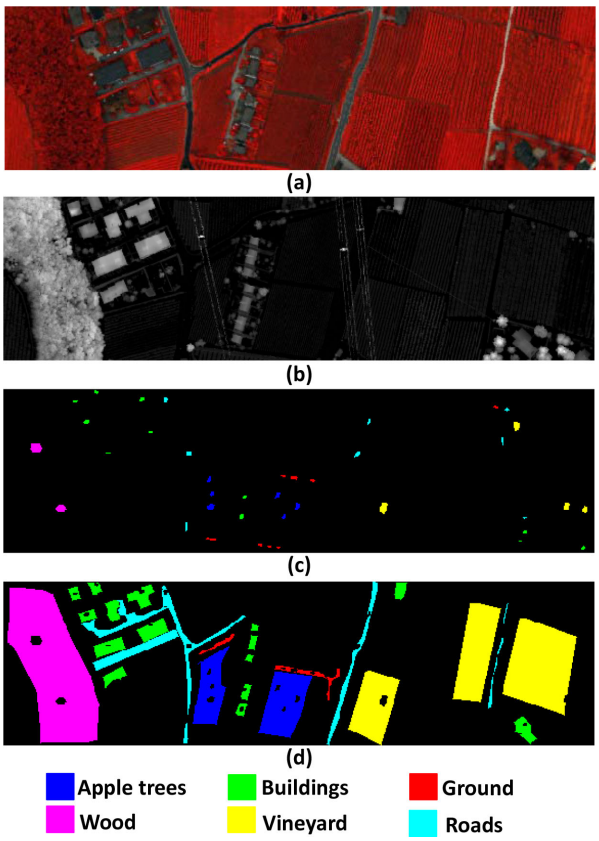

Fig. 5. Visualization of the Trento data. (a) Pseudocolor image for the hyperspectral data using 40, 20, and 10 as R, G, and B, respectively. (b) Grayscale image for the LiDAR data. (c) Training data map. (d) Test data map.

also worth noting that not all classes have been improved because CNN models need to balance the contributions of hyperspectral and LiDAR sources on the whole data. Although some classes (e.g., *Soil*) have been slightly degraded, more classes have been improved. For example, the accuracy of *Synthetic grass* has been improved from 91.29% to 99.60%. Upon EMFNet-FF-FT, when the channel attention module shown in Fig. 2 is added, EMFNet-FF-ST is capable of increasing the classification performance in terms of OA, AA, and Kappa. This can be explained by the enhanced feature representation ability of LiDAR with the help of hyperspectral data. If the spatial tuning module is added again, the feature representation ability of hyperspectral would be enhanced, resulting in better performance. Specifically, the OA, AA, and Kappa values have been increased about 1% using EMFNet-FF. Compared to all the aforementioned deep models, the EMFNet can obtain the best performance in terms of OA, AA, and Kappa values, which validate the effectiveness of our proposed model. In addition to the deep learning models, the EMFNet also shows significant improvement when compared with SVM and CHOTF. Besides the quantitative comparisons, Fig. 6 also qualitatively analyzes the classification maps of the deep-learning-related models. It is shown that our proposed EMFNet demonstrates more consistent results with the ground-truth map than the other models, but all of them tend to get oversmoothing results, especially in the object boundary areas.

Similar to Table III and Fig. 6, Table IV and Fig. 7 show quantitative and qualitative results on the Trento dataset, respectively. At the first glance, all models acquire better results than that on the Houston dataset because these data are much easier to classify. Nevertheless, some conclusions are the same as those in the Houston data. In particular, CNN(H) achieves better classification performance than CNN(L). By combining hyperspectral data and LiDAR data, EMFNet-FF-FT can improve the performance of CNN(H) by more than 2% in terms of OA, AA, and Kappa. Both EMFNet-FF-ST and EMFNet-FF outperform EMFNet-FF-FT, which can be attributed to the feature tuning modules employed in those techniques. Additionally, the conventional models (i.e., SVM and CHOTF) also obtain satisfactory performance. CHOTF shows slightly higher OA, AA, and Kappa values than EMFNet-FF. However, when the feature fusion module is incorporated, the proposed EMFNet will outperform CHOTF, which confirms the advantage of the proposed architecture.

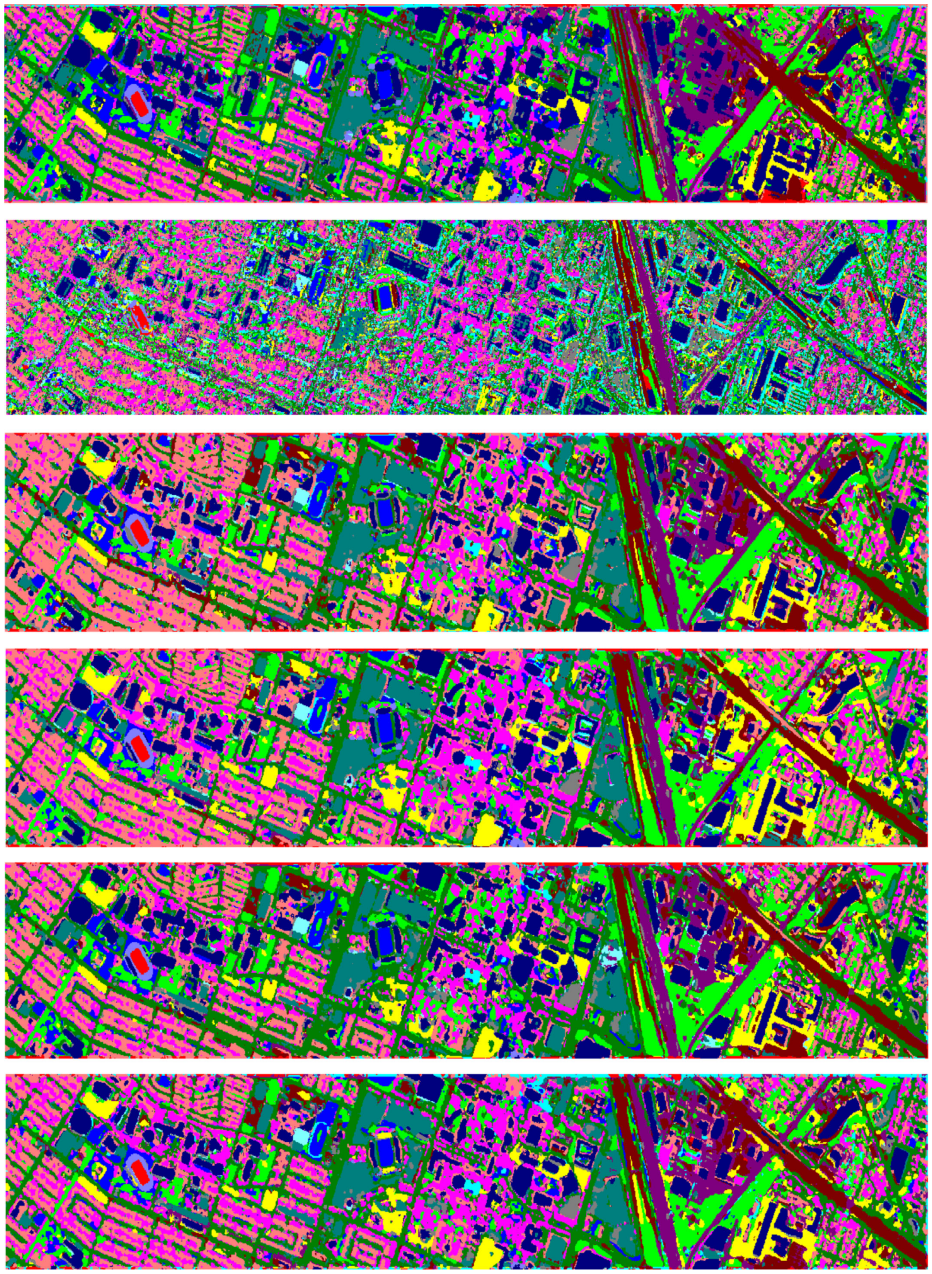

Fig. 6. Classification maps obtained for the Houston data using different fusion models. From top to bottom are the maps generated by CNN(H), CNN(L), EMFNet-FF-FT, EMFNet-FF-ST, EMFNet-FF, and EMFNet.

TABLE IV
CLASSIFICATION PERFORMANCE OF DIFFERENT FUSION MODELS APPLIED ON THE TRENTO DATASET

Class No.	SVM	CHOTF	CNN(H)	CNN(L)	EMFNet-FF-FT	EMFNet-FF-ST	EMFNet-FF	EMFNet
1	85.49	**100.00**	99.85	99.92	99.89	99.67	99.62	99.21
2	89.76	98.62	94.67	93.16	99.03	97.98	**99.53**	**99.53**
3	59.56	95.62	82.09	60.43	78.07	81.02	81.02	**97.86**
4	97.42	99.91	98.73	99.12	**100.00**	99.91	99.97	**100.00**
5	93.85	99.75	99.73	95.63	99.92	**100.00**	**100.00**	99.81
6	89.96	91.15	76.31	50.59	90.33	91.61	90.69	**94.10**
OA	92.30	98.76	96.31	91.91	98.58	98.63	98.69	**99.14**
AA	86.01	97.51	91.90	83.14	94.54	95.03	95.14	**98.42**
Kappa	89.71	98.30	95.05	89.17	98.10	98.16	98.24	**98.85**

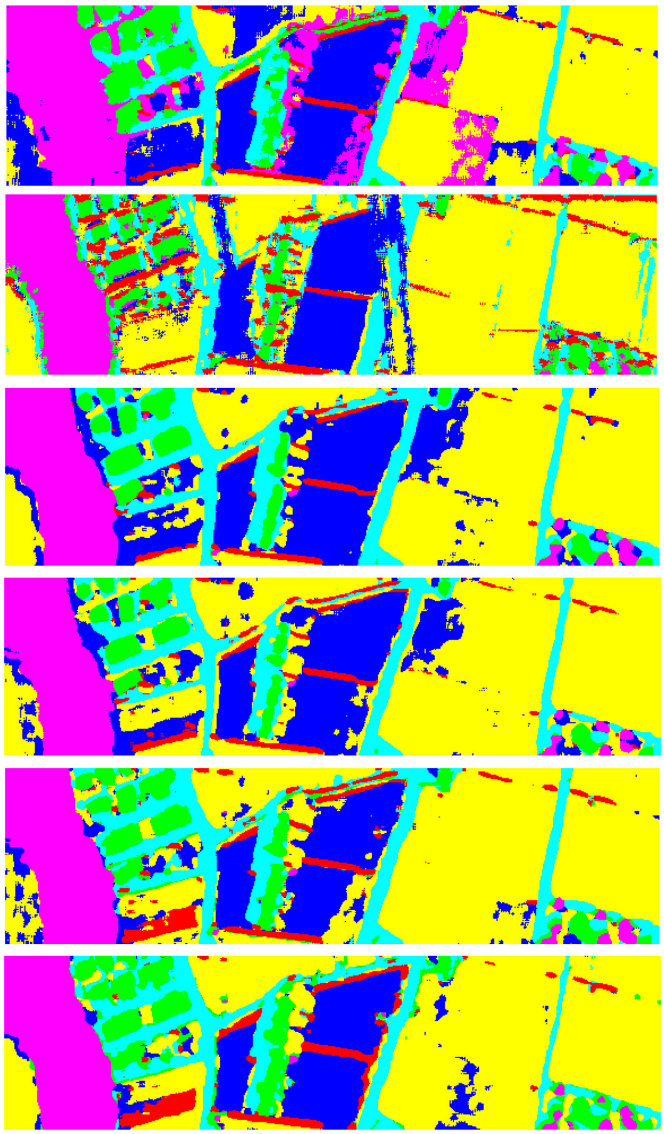

Fig. 7. Classification maps obtained for the Trento data using different fusion models. From top to bottom are the maps generated by CNN(H), CNN(L), EMFNet-FF-FT, EMFNet-FF-ST, EMFNet-FF, and EMFNet.

IV. Conclusion

This article proposed an EMFNet to classify hyperspectral and LiDAR data. It attempts to take advantage of the complementary information from both sources in the feature extraction phase and the feature fusion phase. In order to achieve this goal, two different modules, named feature tuning module and feature fusion module, are designed and incorporated into CNNs. The feature tuning module is comprised of two different modules, which are expected to enhance hyperspectral features and LiDAR features, respectively. The feature fusion module is desired to dynamically learn two fusion weights for hyperspectral and LiDAR data. To test the performance of our proposed network, several experiments are conducted on two different datasets, i.e., Houston data and Trento data. When compared with state-of-the-art fusion models, including two traditional models and five deep-learning-related models, our proposed model is capable of achieving better performance on both datasets in terms of classification accuracies.

References

[1] B. Rasti *et al.*, "Feature extraction for hyperspectral imagery: The evolution from shallow to deep: Overview and toolbox," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 4, pp. 60–88, Dec. 2020.

[2] 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.

[3] 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.*, to be published, doi: 10.1109/TGRS.2020.3015157.

[4] C. Debes *et al.*, "Hyperspectral and LiDAR data fusion: Outcome of the 2013 GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2405–2418, Jun. 2014.

[5] P. Ghamisi *et al.*, "Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state-of-the-art," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 1, pp. 6–39, Mar. 2019.

[6] Y. Xu *et al.*, "Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 IEEE GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1709–1724, Jun. 2019.

[7] M. Pedergnana, P. R. Marpu, M. D. Mura, J. A. Benediktsson, and L. Bruzzone, "Classification of remote sensing optical and LiDAR data using extended attribute profiles," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 7, pp. 856–865, Nov. 2012.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] B. Rasti, P. Ghamisi, J. Plaza, and A. Plaza, "Fusion of hyperspectral and LiDAR data using sparse and low-rank component analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 11, pp. 6354–6365, Nov. 2017.

[13] C. Ge, Q. Du, W. Li, Y. Li, and W. Sun, "Hyperspectral and LiDAR data classification using kernel collaborative representation based residual fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1963–1973, Jun. 2019.

[14] Q. Yuan *et al.*, "Deep learning in environmental remote sensing: Achievements and challenges," *Remote Sens. Environ.*, vol. 241, 2020, Art. no. 111716.

[15] J. Kang, R. Fernandez-Beltran, D. Hong, J. Chanussot, and A. Plaza, "Graph relation network: Modeling relations between scenes for multilabel remote-sensing image classification and retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4355–4369, May 2021.

[16] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogrammetry Remote Sens.*, vol. 152, pp. 166–177, 2019.

[17] J. Kang, R. Fernandez-Beltran, P. Duan, S. Liu, and A. J. Plaza, "Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2598–2610, Mar. 2021.

[18] 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.

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

[20] 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.

[21] D. Hong *et al.*, "More diverse means better: Multimodal deep learning meets remote sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.
[22] W. Liao, F. Van Coillie, L. Gao, L. Li, B. Zhang, and J. Chanussot, "Deep learning for fusion of apex hyperspectral and full-waveform LiDAR remote sensing data for tree species mapping," *IEEE Access*, vol. 6, pp. 68 716–68729, 2018.
[23] 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.*, to be published.
[24] Q. Liu, F. Zhou, R. Hang, and X. Yuan, "Bidirectional-convolutional LSTM based spectral-spatial feature learning for hyperspectral image classification," *Remote Sens.*, vol. 9, no. 12, 2017, Art. no. 1330.
[25] 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.
[26] W. Liao, R. Bellens, A. Pižurica, S. Gautama, and W. Philips, "Combining feature fusion and decision fusion for classification of hyperspectral and LiDAR data," in *Proc. IEEE Geosci. Remote Sens. Symp.*, 2014, pp. 1241–1244.
[27] Y. Zhong, Q. Cao, J. Zhao, A. Ma, B. Zhao, and L. Zhang, "Optimal decision fusion for urban land-use/land-cover classification based on adaptive differential evolution using hyperspectral and LiDAR data," *Remote Sens.*, vol. 9, no. 8, 2017, Art. no. 868.
[28] R. Luo *et al.*, "Fusion of hyperspectral and LiDAR data for classification of cloud-shadow mixed remote sensed scene," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3768–3781, Aug. 2017.
[29] W. Liao, A. Pižurica, R. Bellens, S. Gautama, and W. Philips, "Generalized graph-based fusion of hyperspectral and LiDAR data using morphological features," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 3, pp. 552–556, Mar. 2015.
[30] Y. Gu, Q. Wang, X. Jia, and J. A. Benediktsson, "A novel MKL model of integrating LiDAR data and MSI for urban area classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 10, pp. 5312–5326, Oct. 2015.
[31] F. Jahan, J. Zhou, M. Awrangjeb, and Y. Gao, "Inverse coefficient of variation feature and multilevel fusion technique for hyperspectral and LiDAR data classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 367–381, 2020.
[32] S. Mohla, S. Pande, B. Banerjee, and S. Chaudhuri, "FusAtNet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and LiDAR classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 92–93.
[33] Z. Xue, S. Yang, H. Zhang, and P. Du, "Coupled higher-order tensor factorization for hyperspectral and LiDAR data fusion and classification," *Remote Sens.*, vol. 11, no. 17, 2019, Art. no. 1959.

Chengxiang Li received the B.S. degree in electrical engineering and automation from Jiangsu Normal University Kewen College, Xuzhou, China, in 2019. He is currently working toward the master's degree in electronic information with the School of Automation, Nanjing University of Information Science and Technology, Nanjing, China.

His research interests include deep learning and remote sensing image processing.

Renlong Hang (Member, IEEE) received the Ph.D. degree from the Nanjing University of Information Science and Technology, Nanjing, China, in 2017.

He is currently a Lecturer with the School of Computer and Software, Nanjing University of Information Science and Technology. From 2018 to 2019, he was a Postdoctoral Researcher with the Department of Computer Science and Electrical Engineering, University of Missouri-Kansas City, Kansas City, MO, USA. He has authored or coauthored more than 30 peer-reviewed articles in international journals, such as IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, and IEEE GEOSCIENCE AND REMOTE SENSING LETTERS. His research interests include machine learning, pattern recognition, and their applications to remote sensing image processing.

Behnood Rasti (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in electronics and electrical engineering from the Department of Electrical Engineering, University of Guilan, Rasht, Iran, in 2006, and 2009, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Iceland, Reykjavik, Iceland, in 2014.

From 2015 to 2016, he was a Postdoctoral Researcher and a Seasonal Lecturer with the Department of Electrical and Computer Engineering, University of Iceland. From 2016 to 2019, he was a Lecturer with the Center of Engineering Technology and Applied Sciences, Department of Electrical and Computer Engineering, University of Iceland. Since 2020, he has been a Humboldt Research Fellow with the Machine Learning Group, Helmholtz-Zentrum Dresden-Rossendorf, Freiberg, Germany. His research interests include machine/deep learning, signal and image processing, data fusion, and remote sensing.

Dr. Rasti was the Valedictorian as an M.Sc. Student in 2009. He received the Doctoral Grant of the University of Iceland Research Fund and was awarded "The Eimskip University Fund" in 2013. In 2019, he received the prestigious "Alexander von Humboldt Research Fellowship Grant." He is an Associate Editor for IEEE GEOSCIENCE AND REMOTE SENSING LETTERS and *Remote Sensing*.