Cross-Stage Fusion Network Based Multi-modal Hyperspectral Image Classification

Yuegong Sun[1](✉), Zhening Wang[1], Ao Li[1], and Hailong Jiang[2]

[1] School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, China
sunyuegong96@163.com

[2] Department of Computer Science, Kent State University, Kent, USA

Abstract. With the development of satellite technology and airborne platforms, there are more and more methods to acquire remote sensing data. The remote sensing data acquired by multiple methods contain different information and internal structures. Nowadays, single-mode hyperspectral image (HSI) data are no longer satisfactory for researchers' needs. How to apply and process the information of multimodal data poses a great challenge to researchers. In this paper, we propose a deep learning-based network framework for multimodal remote sensing data classification, where we construct an advanced cross-stage fusion strategy using a fully connected network as the backbone, called CSF. Like the name implies, CSF incorporated two separate stages of fusion strategies for moreeffective fusion of multimodal data: fusion at the pre-structure and fusion at the tail of the network. This strategy prevents the preservation of excessive redundant information in the pre-fusion and the details of information lost due to late fusion. Moreover, a plug-and-play cross-fusion module for CSF is implemented. On the Houston 2013 dataset, our model strategy outperformed the fusion strategy of each stage and the single-modal strategy, which also demonstrated that multimodal feature fusion has promising performance.

Keywords: Multi-modal · Feature Fusion · Hyperspectral Image Classification · Remote Sensing

1 Introduction

Compared with common RGB images, remote sensing data contains more information. However, the more information the more troublesome it is to process. Hyperspectral image is a kind of remote sensing data with a large number of spectral bands, which is characterized by rich information and high resolution, so it is widely used in objective detection [1], environmental exploration [2], mineral exploration [3], agricultural resource survey [4] and ocean research [5]. Since it contains a large amount of band information and the feature resemblance between adjacent bands is strong, it largely increases the computational complexity of hyperspectral image classification. Therefore, feature learning is needed for hyperspectral images to remove redundant

A. Li et al. (Eds.): 6GN 2022, LNICST 504, pp. 77–88, 2023.
https://doi.org/10.1007/978-3-031-36011-4_7

information, reduce data dimensionality, and improve classification accuracy. However, hyperspectral images are distorted by clouds and atmosphere, which causes significant problems for researchers, who have to allocate a considerable amount of effort to work on removing clouds and noise for the data, but the results are still unsatisfactory. However, Light Detection and Ranging (LiDAR) images are not affected by cloud cover and have image features that hyperspectral images lack, for example, the height and shape of land-covered objects. The image classification improves the classification accuracy after fusing the features of hyperspectral images and LiDAR images.

The technology in the field of hyperspectral image classification is gradually maturing. Traditionally, the main approach is feature dimensionality reduction and feature selection for hyperspectral images. Learning the potential subspaces of hyperspectral images or their intrinsic flow structures, and finding the information-rich bands of hyperspectral images among the redundant bands [6]. In the domain of deep learning, descending and feature extraction can be performed by modules such as fully connected networks, convolutional networks and attention mechanisms, and finally image classification by logistic regression [7]. The multimodal data classification of hyperspectral images and LiDAR images can be borrowed from the classification method of hyperspectral images, which is mainly studied in how to fuse the two features. This paper focuses on the performance impact of both features on the fusion stage in a fully connected network. Specifically, the contributions of this paper are summarized as follows.

1. Multi-modal data image classification framework dominated by fully connected networks with advanced cross-stage fusion strategy modules is designed.
2. A plug-and-play cross-stage fusion strategy module enables more effective reduction of redundant information while enhancing detailed information during processing of data fusion.

The rest of this paper is presented below. In Sect. 2, related work is presented. Section 3, presents the structure of our proposed model. Section 4, demonstrates the experimental results and provides an analysis. Section 5, discusses the limitations of the model and future work.

2 Related Work

2.1 Single-Modal Feature Learning

Compared with multi-modal, single-modal feature learning just requires model design considering the characteristics of its own data. For HSI, different ground objects may appear with the same spectral profile features; there may also be the same ground objects with different spectral profile features. One way is to pre-process the pixel points by extracting spatial texture information as well as morphological features to complete the task of feature extraction, and then input the extracted information into the classifier for classification. For example, the proposed Local binary pattern (LBP) algorithm [8] and the LBP improved feature extraction method [9] enable such local texture methods to be widely used. In another way, the pixel points to be classified and their neighborhood pixel points are directly input to the classification machine, which is performed by designing a high-quality and efficient classifier. An example is the support vector machine

(SVM) [10] classifier. Jia et al. [11] proposed a super pixel-level unsupervised linear discriminant analysis framework based on gabber to extract the most informative and the most discriminative features. These traditional methods mentioned above use manual feature extraction to obtain image features, while spectral images are often difficult to obtain features with high discriminative power due to interference from factors such as cloud noise. In recent years, deep learning has become the focus of image classification and has been widely used in the field of image classification. Hu et al. [12] argued that the core building block of convolutional neural networks is the convolutional kernel, which is usually viewed as an information aggregator that aggregates information on space and information on channels over local sensory fields, so that in addition to spatial information, channel information cannot be neglected as well. However, the low-level features in the early stage of the neural network in [13] are rich in spatial information but lack semantic information, while the high-level features in the later stage are rich in semantic information but lack spatial information, but the two are isolated from each other and difficult to fully utilize.

2.2 Multi-modal Feature Learning

In the literature [14], a three-layer point-to-point mapping was designed, while a point-to-point convolutional network was designed as a hidden layer in order to merge multi-scale features between two different sources and extract deeply fused features to obtain an accurate representation of hyperspectral image data. You et al. [15] proposed a multi-view common component discriminant analysis to jointly deal with view differences, discriminability and nonlinearity, mainly addressing the problem of nonlinear manifold subspaces leading to degraded classification performance by adding supervised information and local geometric information to the common component extraction process to learn to obtain discriminative common subspaces and to be able to deal with the nonlinear structure in the obtained multi-view data. A pixel-level decision fusion method fusing HSI and Lidar is proposed in [16]. The data are first processed using kernel principal component analysis. Then Gabor filter is used to obtain the amplitude and phase information, which is composed of three sets of data with the original data. Finally, inter-pixel information is obtained by super-pixel segmentation, with which the three sets of data are then fused for classification. In [17], a deep learning-based multimodal classification framework is proposed, in which convolutional neural networks (CNNs) are used as the backbone of a cross-channel reconstruction module. The cross-modal reconstruction strategy learns more compact fused representations of different data sources that can exchange information with each other more efficiently. In [18], provides a baseline solution for the simultaneous processing and analysis of multimodal data by developing a generic multimodal deep learning framework. In [19], CNN net is used to learn the spectral spatial features and the elevation information of the Lidar data. Using a composition of three convolutional layers, where the feature fusion is performed in the last two layers of the convolutional network. According to the sequence of feature fusion, these methods can be categorized as early fusion, intermediate fusion, and late fusion. An intuitive early fusion technique is to superimpose data from multiple modalities in the channel direction and input them to the network as 4- or 6-channel data. Marcos et al. [20] simply combined NIR, red-green spectra, and digital surface models (DSM)

as inputs to the network. This image-level fusion approach, by not taking full advantage of the relationships between heterogeneous information, can introduce redundant features in the training. The medium-term fusion approach, also known as hierarchical fusion, combines feature mapping from different levels of multimodal specific encoders and uses a single decoder to up-sample the fused features. Marmanis et al. [21] design parallel branching networks to extract DSM data features and perform modal feature interactions in the middle layer, but this massive structure has a large number of parameters, is hardware demanding, and can be time consuming in the training and inference phases. Post-fusion approaches usually design identical networks that are first trained individually in a specific modality and then use cascade or element-level summation to fuse feature mappings to the end of the network, typically represented by V-FuseNet [22], which uses two convolutional neural networks for spectral and DSM data, respectively, and element summation fusion is performed.

3 The Proposed Method

3.1 Different Stages of Feature Fusion

In multi-modal data processing, data fusion strategy is a major problem, and different stages of fusing features have different effects. The unprocessed HSI and LiDAR are connected according to the spectral band dimension and input to the network for feature extraction for processing, called early fusion.

Let $\mathrm{X}_1 \in \mathbb{R}^{\mathrm{d}_1 \times N}$ and $X_2 \in \mathbb{R}^{d_2 \times N}$ denote multimodal data with d_1 and d_2 dimensions, N denote the number of samples, where $x_{1,i}$ denotes the i-th sample in the X_1 modal data. Let $Y \in \mathbb{R}^{C \times N}$ denote the same label information shared by multimodal data, with C categories and N samples, which is the one-hot code label matrix. With the definition as above, the input for early fusion can be expressed as $x_i = \left[x_{1,i}, x_{2,i}\right]$, $\mathrm{i} = 1, ..., N$. The fused features are input to the fully connected network for processing, and the output of the l-th layer can be denoted as

$$z_i^{(l)} = \begin{cases} h_{W^{(l)},b^{(l)}}(x_i) & l = 1, \\ h_{W^{(l)},b^{(l)}}(z_i^{(l-1)}) & l = 2, ..., p \end{cases} \tag{1}$$

where l indicates the number of network layers. $h(\cdot)$ denotes the linear regression equation, where $W^{(l)}$ and $b^{(l)}$ denote the weights and biases that can be learned in layer l. .. We introduce batch normalization (BN) layers to speed up convergence and training, as well as control the gradient to prevent gradient explosion and training overfitting, which is added to the output $z_i^{(l)}$

$$z_{BNi}^{(l)} = \alpha \hat{z}_i^{(l)} + \beta \tag{2}$$

where $\hat{z}_i^{(l)}$ is the $z - score$ result of $z_i^{(l)}$, α and β denote the network parameters to be learned by the BN layer. In order to make sense of the deep fully connected network, a nonlinear activation ReLU operation is performed on the output of each layer with the following equation

$$a_i^{(l)} = \mathrm{Re}LU(z_{BNi}^{(l)}) \tag{3}$$

Different from the earlier fusion, the mid-stage fusion first delivers the multimodal data to different fully connected layer networks for feature extraction separately. Then, the features are merged and sent to the fully connected layer network again for further feature fusion. Finally, the output results are used for classification. The connection of the two data after feature extraction is called later fusion, where the connected data are directly put into logistic regression for classification.

3.2 Cross-Stage Feature Fusion

Our proposed cross-stage feature fusion strategy is able to preserve the details of image data and remove the redundancy of image data. The reason is that the cross-stage feature fusion strategy combines the advantages of early fusion and later fusion by constructing pre-processing and post-processing modules. Pre-processing is used to retain the details lost in removing redundant information, whereas post-processing is used to boldly remove redundant information. The structure diagram is shown in Fig. 1.

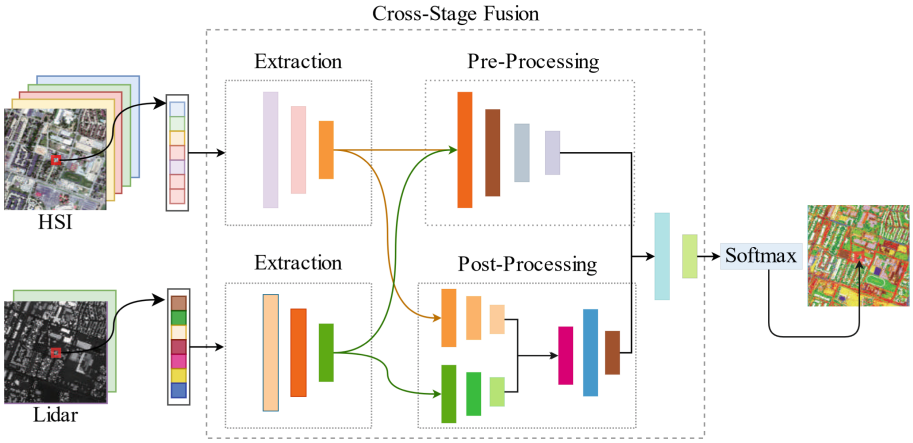

Fig. 1. Architecture diagram of cross-stage fusion method

CSF method mainly contains two parts: feature extraction and feature fusion. First, the samples are selected from the dataset for feature extraction through fully-connected (FC) network. Then, feature fusion is performed. Feature fusion is divided into a pre-processing of merging the two parts of data for deep fusion and a post-processing of continuing feature extraction to the end of the network to be merged and fused further, where the structure of the FC network is shown in Fig. 2.

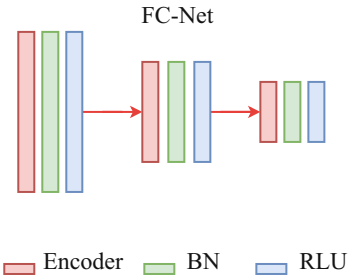

Fig. 2. Architecture diagram of FC-Net

The network structure is a fully connected network with internal blocks consisting of FC blocks as shown in Fig. 1. The purpose of the cross-stage fusion strategy is to minimize the mean squared error of the true value and the predicted value, and its loss function is as follows

$$\mathrm{L} = \frac{-1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1-y_i)\log\left(1-\hat{y}_i\right)\right] \tag{4}$$

where y_i and $\hat{y}_i$ are the true value and predicted value of N samples, respectively. In order to prevent overfitting in the training process and reduce the complexity of the model, additional constraints are imposed on the network parameters, and the loss function can be written as

$$\mathrm{L} = \frac{-1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1-y_i)\log\left(1-\hat{y}_i\right)\right] + \lambda\frac{1}{l}\sum_{i=1}^{l}\left\|W^{(l)}\right\|_2^2 \tag{5}$$

where λ is the parameter controlling the complexity, increasing the value of λ will reduce the complexity.

4 Experiments

4.1 Dataset

For the experiments, we used two datasets to test the effectiveness of our method, HSI-LiDAR Houston2013 Dataset and HSI-SAR Berlin Dataset [23], respectively.

HSI-LiDAR Houston2013 Dataset is grouped into two parts, HSI with 144 bands and Lidar with only 1 band, with a total of 349*1905 pixels, of which the number of training samples is 2832, and the number of testing samples is 12197, with the total number of categories being 15. Table 1 shows the number of training and testing samples for each category of the data set, respectively.

Table 1. Name, training set and test set of each category included in the Houston2013 data

Class	Class Name	Train Set	Test Set
Class 1	Healthy Grass	198	1053
Class 2	Stressed Grass	190	1064
Class 3	Synthetic Grass	192	505
Class 4	Tree	188	1056
Class 5	Soil	186	1056
Class 6	Water	182	143
Class 7	Residential	196	1072
Class 8	Commercial	191	1053
Class 9	Road	193	1059
Class 10	Highway	191	1036
Class 11	Railway	181	1054
Class 12	Parking Lot1	192	1041
Class 13	Parking Lot2	184	285
Class 14	Tennis Court	181	247
Class 15	Running Track	187	473
Total		2832	12197

HSI-SAR Berlin Dataset is composed of HSI with 244 bands and SAR with 4 bands, having 1723*476 pixels points, in which there are 2820 training samples and 461851 testing samples, totally 8 of classes. The number of training samples and the number of testing samples for each class of the dataset are shown in Table 2, accordingly.

Table 2. Name, training set and test set of each category included in the Berlin data

Class	Class Name	Train Set	Test Set
Class 1	Forest	443	54511
Class 2	Residential Area	423	268219
Class 3	Industrial Area	499	19067
Class 4	Low Plants	376	58906
Class 5	Soil	331	17095
Class 6	Allotment	280	13025
Class 7	Commercial Area	298	24526
Class 8	Water	170	6502
Total		2820	461851

Both datasets have similar numbers of total and training samples, whereas Berlin data has four times more test samples than Houston2013 data, but the number of categories is double that of the other.

4.2 Result Analysis

There are three metrics to judge the classification results of multimodal data, namely overall accuracy (OA), average accuracy (AA) and kappa coefficient (Kappa). Their equations are shown as follows

$$\mathrm{OA} = \frac{\mathrm{N}_p}{N_t} \tag{6}$$

$$\mathrm{AA} = \frac{1}{C}\sum_{i=1}^{C}\frac{\mathrm{N}_p^i}{N_t^i} \tag{7}$$

$$Kappa = \frac{OA - P_e}{1 - P_e} \tag{8}$$

where N_t and N_p denote the number of classified samples and the number of correct predictions, respectively. N_t^i and N_p^i denote the sample number for each class corresponding to N_t and N_p. P_e can be expressed as

$$P_e = \frac{\mathrm{N}_p^1 \times N_t^1 + \cdots + \mathrm{N}_p^i \times N_t^i + \cdots + \mathrm{N}_p^C \times N_t^C}{N_t \times N_t} \tag{9}$$

In order to verify the effectiveness of our proposed method CSF-Net, it is compared with the single-modal methods HIS-FC and Lidar-FC, which validates the effectiveness of the multimodal method, separately, in addition to the Early-Net, Mid-Net and Later-Net multimodal methods, which validates the efficiency of our proposed fusion strategy. It is also compared with other recent multimodal hyperspectral classification methods, such as LeMA, CapsNet and CoCNN. Our experiments use the Tensorflow framework. The optimizer used is Adam optimization. The initial learning rate is set to 0.001. The network parameters are regularized with $l_2 - norm$ to prevent overfitting.

As shown in Fig. 3, there is a clear gap between the single-modal method and the multi-modal method, therefore, complementary information exists in both modalities which can be fused and utilized. Among the single-modal methods, there is a huge gap between the Lidar-Net method and the HSI-Net method, which shows that the HSI contains much more information than the Lidar. Our proposed method has the highest accuracy which proves that the cross-fusion strategy is a successful method.

As can be seen from Table 3, our method is the best in all three-evaluation metrics. Details of the classification of each category in the Houston2013 dataset are presented. Categories C3 and C14 are basically all predicted correctly, as synthetic grass and tennis courts, respectively, have distinct characteristics and regional invariance.

As shown in Table 4, our method is the best among the two-evaluation metrics. In the demonstration of the classification of each category, category C7 has a low correct prediction rate because of the chaotic features and regional irregularities of the commercial area. The overall evaluation metrics of Berlin dataset are low, indicating that this dataset has complex potential features and belongs to a relatively new and complex dataset. The next work will investigate this dataset more to increase the classification accuracy.

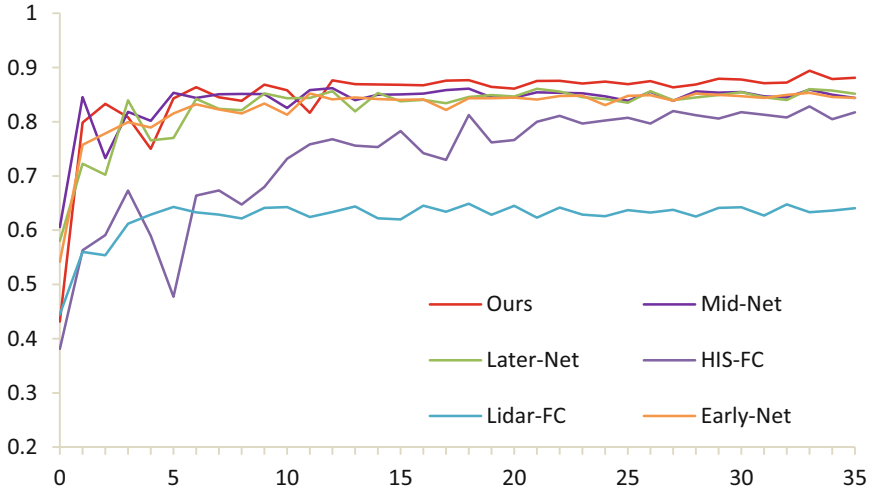

Fig. 3. Accuracy curves of the proposed method and other methods

Table 3. Experimental results for the Houston2013 dataset

Methods	HSIFC	EarlyNet	MidNet	LaterNet	LeMA	CapsNet	CoCNN	**Ours**
C1	82.91	82.15	82.81	8272	81.86	81.10	**83.10**	82.72
C2	83.74	83.74	82.33	83.83	83.80	81.02	**84.87**	82.24
C3	**100**	**100**	**100**	**100**	**100**	96.44	99.80	**100**
C4	91.57	92.52	93.18	93.18	**94.79**	88.35	92.42	92.52
C5	97.63	**99.62**	98.86	99.24	99.34	100	99.24	98.67
C6	95.10	83.92	95.80	95.10	**99.30**	95.80	95.80	95.10
C7	86.57	81.81	81.34	84.42	88.99	86.37	**95.20**	80.97
C8	45.30	81.58	79.58	78.73	74.26	90.10	81.86	81.67
C9	71.67	71.96	79.69	74.98	73.84	82.53	85.08	**86.87**
C10	**86.85**	79.63	70.85	70.27	72.20	72.78	61.10	79.63
C11	80.07	76.19	83.25	80.93	82.26	82.99	83.09	**83.30**
C12	86.84	81.56	90.20	**90.39**	90.30	83.09	91.26	88.95
C13	74.74	83.16	80.70	75.09	67.37	76.14	86.77	**87.37**
C14	**100**	**100**	**100**	99.60	**100**	93.93	91.09	**100**
C15	98.73	98.10	98.73	98.52	98.10	97.46	98.73	**99.52**
OA(%)	82.84	85.45	86.19	86.07	85.42	86.52	87.23	**89.42**
AA(%)	83.45	86.39	87.89	87.13	87.05	87.54	88.22	**89.24**
Kappa(%)	81.37	84.20	85.01	84.87	84.17	85.41	86.19	**88.51**

Table 4. Experimental results for the Berlin dataset

Methods	HSIFC	EarlyNet	MidNet	LaterNet	LeMA	CapsNet	CoCNN	**Ours**
C1	75.62	65.95	75.66	76.29	64.18	**84.96**	84.09	64.18
C2	51.78	57.68	63.04	62.50	64.11	65.22	**68.48**	64.11
C3	51.19	53.55	53.01	49.79	56.62	48.42	49.09	**56.62**
C4	76.32	**84.65**	80.45	77.58	70.28	80.8	79.43	70.28
C5	78.29	**82.37**	71.44	80.00	76.00	69.18	81.25	76.00
C6	66.97	64.48	63.87	61.73	70.10	55.08	50.68	**70.20**
C7	30.95	30.00	27.86	35.56	30.11	26.12	26.16	**36.11**
C8	**68.99**	64.30	63.99	64.27	59.77	59.69	59.52	59.77
OA (%)	64.42	67.33	71.07	70.27	66.71	66.55	68.51	**71.68**
AA (%)	63.01	63.58	62.42	63.47	62.35	61.18	62.34	**63.69**
Kappa (%)	43.13	47.78	53.75	52.48	53.12	52.77	54.76	**54.79**

5 Conclusion

In this paper, we propose a multi-modal feature fusion strategy based on FC-Net, we mainly explore the effect of multi-modal fusion strategy and the effect of different fusion strategies. Based on different period fusion strategies, the most discriminative features of the two stage methods are fused, finally extracting the more robust and easily distinguishable features. Experiments prove that our proposed method is the best. The output of pre-processing and post-processing is fused to obtain features that eliminate redundant information and retain detailed information, which improves the classification results. Although the method has promising results, however, hyperspectral image region invariance is not exploited. Next, we will prepare to explore the deep network model for multi-modal data with a combination of spatial and spectral information.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China under Grant 62071157, National Key Research and Development Programme 2022YFD2000500 and Natural Science Foundation of Heilongjiang Province under Grant YQ2019F011.

References

1. Liang, J., Zhou, J., Tong, L., Bai, X., Wang, B.: Material based salient object detection from hyperspectral images. Pattern Recognit. **76**, 476–490 (2018). https://doi.org/10.1016/j.patcog.2017.11.024
2. Gao, B., et al.: Additional sampling layout optimization method for environmental quality grade classifications of farmland soil. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **10**, 5350–5358 (2017). https://doi.org/10.1109/JSTARS.2017.2753467

3. Zadeh, M.H., Tangestani, M.H., Roldan, F.V., Yusta, I.: Mineral exploration and alteration zone mapping using mixture tuned matched filtering approach on ASTER Data at the Central Part of Dehaj-Sarduiyeh Copper Belt, SE Iran. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **7**, 284–289 (2014). https://doi.org/10.1109/JSTARS.2013.2261800
4. Lu, B., Dao, P., Liu, J., He, Y., Shang, J.: Recent advances of hyperspectral imaging technology and applications in agriculture. Remote Sens. **12**, 2659 (2020). https://doi.org/10.3390/rs12162659
5. Kobryn, H.T., Wouters, K., Beckley, L.E., Heege, T.: Ningaloo reef: shallow marine habitats mapped using a hyperspectral sensor. PLoS ONE **8**, e70105 (2013). https://doi.org/10.1371/journal.pone.0070105
6. Tang, C., Liu, X., Zhu, E., Wang, L., Zomaya, A.: Hyperspectral band selection via spatial-spectral weighted region-wise multiple graph fusion-based spectral clustering. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, pp. 3038–3044. International Joint Conferences on Artificial Intelligence Organization, Montreal, Canada (2021). https://doi.org/10.24963/ijcai.2021/418
7. Chakraborty, T., Trehan, U.: SpectralNET: Exploring Spatial-Spectral WaveletCNN for Hyperspectral Image Classification. http://arxiv.org/abs/2104.00341 (2021)
8. Guo, Z., Zhang, L., Zhang, D.: A completed modeling of local binary pattern operator for texture classification. IEEE Trans. Image Process. **19**, 1657–1663 (2010). https://doi.org/10.1109/TIP.2010.2044957
9. Liu, L., Lao, S., Fieguth, P.W., Guo, Y., Wang, X., Pietikäinen, M.: Median robust extended local binary pattern for texture classification. IEEE Trans. Image Process. **25**, 1368–1381 (2016). https://doi.org/10.1109/TIP.2016.2522378
10. Camps-Valls, G., Bruzzone, L.: Kernel-based methods for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. **43**, 1351–1362 (2005). https://doi.org/10.1109/TGRS.2005.846154
11. Jia, S., et al.: Flexible gabor-based superpixel-level unsupervised LDA for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. **59**, 10394–10409 (2021). https://doi.org/10.1109/TGRS.2020.3048994
12. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7132–7141. IEEE, Salt Lake City, UT (2018). https://doi.org/10.1109/CVPR.2018.00745
13. Zhang, N., Li, J., Li, Y., Du, Y.: Global attention pyramid network for semantic segmentation. In: 2019 Chinese Control Conference (CCC), pp. 8728–8732 (2019). https://doi.org/10.23919/ChiCC.2019.8865946
14. Zhang, M., Li, W., Du, Q., Gao, L., Zhang, B.: Feature extraction for classification of hyperspectral and LiDAR data using patch-to-patch CNN. IEEE Trans. Cybern. **50**, 100–111 (2020). https://doi.org/10.1109/TCYB.2018.2864670
15. You, X., Xu, J., Yuan, W., Jing, X.-Y., Tao, D., Zhang, T.: Multi-view common component discriminant analysis for cross-view classification. Pattern Recognit. **92**, 37–51 (2019). https://doi.org/10.1016/j.patcog.2019.03.008
16. Jia, S., et al.: Multiple feature-based superpixel-level decision fusion for hyperspectral and LiDAR data classification. IEEE Trans. Geosci. Remote Sens. **59**, 1437–1452 (2021). https://doi.org/10.1109/TGRS.2020.2996599
17. Wu, X., Hong, D., Chanussot, J.: Convolutional neural networks for multimodal remote sensing data classification. IEEE Trans. Geosci. Remote Sens. **60**, 1 (2022). https://doi.org/10.1109/TGRS.2021.3124913
18. Hong, D., et al.: More diverse means better: multimodal deep learning meets remote-sensing imagery classification. IEEE Trans. Geosci. Remote Sens. **59**, 4340–4354 (2021). https://doi.org/10.1109/TGRS.2020.3016820

19. Hang, R., Li, Z., Ghamisi, P., Hong, D., Xia, G., Liu, Q.: Classification of hyperspectral and LiDAR data using coupled CNNs. IEEE Trans. Geosci. Remote Sens. **58**, 4939–4950 (2020). https://doi.org/10.1109/TGRS.2020.2969024
20. Marcos, D., Volpi, M., Kellenberger, B., Tuia, D.: Land cover mapping at very high resolution with rotation equivariant CNNs: towards small yet accurate models. ISPRS J. Photogramm. Remote Sens. **145**, 96–107 (2018). https://doi.org/10.1016/j.isprsjprs.2018.01.021
21. Marmanis, D., Wegner, J.D., Galliani, S., Schindler, K., Datcu, M., Stilla, U.: Semantic Segmentation of Aerial Images with an Ensemble of CNNS. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, pp. 473–480. Copernicus GmbH (2016). https://doi.org/10.5194/isprs-annals-III-3-473-2016
22. Audebert, N., Le Saux, B., Lefèvre, S.: Beyond RGB: very high resolution urban remote sensing with multimodal deep networks. ISPRS J. Photogramm. Remote Sens. **140**, 20–32 (2018). https://doi.org/10.1016/j.isprsjprs.2017.11.011
23. Hong, D., Hu, J., Yao, J., Chanussot, J., Zhu, X.X.: Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model. ISPRS J. Photogramm. Remote Sens. **178**, 68–80 (2021). https://doi.org/10.1016/j.isprsjprs.2021.05.011