Multimodal Remote Sensing Classification with Cascaded Attention Convolution Neural Network

Haotian Zhang, Li Ni and Min Huang

May 31, 2022

Multimodal Remote Sensing Classification with Cascaded Attention Convolution Neural Network

Haotian Zhang[1,2], Li Ni[1,*] and Min Huang[1]

[1] *The Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China*

[2] *School of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China*

Abstract

Multimodal remote sensing classification tasks always encounter the data problem of unbalanced feature distributions from various information sources. In this paper, we adopt the attention mechanism with a cascaded multi-scale training strategy to enhance the performance feature extraction of one data source. We have utilized the hyperspectral and LiDAR data to provide the proposed algorithm's efficiency with multimodal Trento dataset. Finally, we have achieved better classification performance on the ground object categories with close similarity on height features owing to strengthening the feature extraction by our methodology.

Keywords

Multimodal Remote Sensing Classification, Attention Module, Cascaded Convolution Neural Network

1. Introduction

The remote Sensing classification task is one of the most critical tasks in the earth observation study area. With the development of the satellite industry and high-efficiency computation resources, more researchers began to follow the remote sensing classification sector [1, 2, 3, 4, 5]. It is still challenging and complex for us to utilize the multimodal remote sensing data for more precise classification results. One of the most common problems, when researchers would like to use more modality remote sensing data to help improve the task efficiency, it does not always work better, introducing more information with the same algorithm model. One of the illustrations of this presence is that the feature extracted from diverse modalities maintains largely different distribution characteristics. This resulted in the original simple model could hardly distinguish which kind of feature information is more vital and helpful for the classification task [6]. So in this paper, we attempt to utilize the attention mechanism to help us gain more high-quality image features to improve the results. Air-bone Hyperspectral and LiDAR data have been chosen to prove the proposed algorithm in this paper [7, 8].

Hyperspectral image (HSI) is equipped with sufficient spectral information of ground objects due to its broad wavelength range and high spectral sampling rate. Hyperspectral data contains dozens to hundreds of spectral bands ranging from visible light to short-wave infrared bands, which could distinguish the ground objects with high similarity from the perspective of human vision. [9, 10] However, coins have two sides, and HSI could hardly be capable of a high-resolution spatial ratio generally to centimeter-level. In this case, the single hyperspectral data is tough to achieve good performance in the precise remote sensing scenarios such as ports and downtown regions. Researchers have investigated the possibility of introducing LiDAR data to compensate for the weakness of hyperspectral data in tackling complex remote sensing classification scenarios. The information of LiDAR could hardly be impacted by lots of environmental factors such as illumination, wind strength, and so on with high-accuracy ground object height features [11]. The features converging the precise spectral information of HSI and height information of LiDAR would improve the classification performance from the perspective of physical theory explanation [12].

The attention mechanism has been proved to with great significance in remote sensing community in the last research literature when applying attention mechanism into the methods based on deep learning framework [13, 14, 15, 16].

Before, The methods based on deep learning could extract more complex and hierarchical features of multimodal remote sensing data, which have been experimented with in recent years with better classification results than other classical machine learning methods (e.g., Support Vector Machine, Extreme Learning Machine) [17, 18, 17, 19]. In this paper, we will introduce the attention mechanism combining with deep learning methods to improve the multimodal classification performance.

2nd Workshop on Complex Data Challenges in Earth Observation (CDCEO) 2022, April 01, 2022

[*] Corresponding author.

zhanghaotian19@mails.ucas.ac.cn (H. Zhang); nili@aircas.ac.cn (L. Ni); huangmin@aircas.ac.cn (M. Huang)

CEUR Workshop Proceedings (CEUR-WS.org)

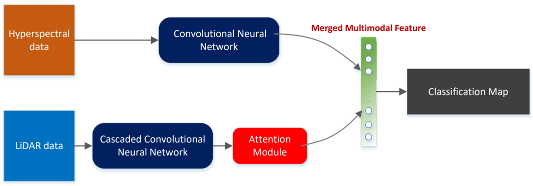

Hyperspectral data
Convolutional Neural Network
Merged Multimodal Feature
Classification Map
LiDAR data
Cascaded Convolutional Neural Network
Attention Module

Figure 1: The architecture of whole algorithm which separately extract Hyperspectral and LiDAR feature with CNN.

2. Methodology

We will first illustrate the algorithm framework in this section and how could we utilise the multimodal remote sensing data (Hyperspectral and LiDAR data). And then highlight the main contribution that we proposed in this multimodal classification framework.

2.1. Algorithm Framework

Following the research work explored before by other researchers, we still adopt the multi-stream framework to deal with the multimodal remote sensing data. In the first step, Hypersepctral data and LiDAR data would go through different convolutional neural network to extract their image features. Following the multimodal feature extraction, the feature fusion strategy is applied here shown in Fig. (1).

2.1.1. Hyperspectral CNN

Towards high-dimensional hyperspectral data $\mathbf{H}^{M\times\mathrm{N}\times\mathrm{K}}$, we designed a Co-CNN hybrid network for the HSI image to separately exploit two-dimensional spatial and one-dimensional spectral features. For gaining more efficient hyperspectral spatial feature, the training input of two-dimensional CNN has been set as the $9{\times}9$ patch $\mathbf{H}_{ij}^{spatial} \in \mathbb{R}^{9\times 9}$ where the core pixel $\mathbf{p}_{ij}$ has been labeled with training ground truth. For further learning rich spectral feature, the one-dimension training sample $\mathbf{H}_{ij}^{\text{spectral}} \in \mathbb{R}^{1\times\mathrm{K}}$ will be adopted with ground truth.

The 1-D CNN and 2-D CNN are five convolution layers with batch normalization and ELU (Exponential Linear Unit) 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 $\mathbf{F}_{\text{HSI}} = [\mathbf{F}_{\text{spectral}}, \mathbf{F}_{\text{spatial}}] \in \mathbb{R}^{1\times(\mathrm{p}+\mathrm{q})}$ will go through full connection layer and the Softmax loss function to predict the classification results.

2.1.2. LiDAR Cascaded Attention CNN

For extracting ground height, we designed a two-dimension CNN with a cascaded feature extraction module with a multi-scaled kernel strategy and attention module to better enhance the LiDAR feature's weight. Given the LiDAR patch image, cascaded block and attention block will help us locate the key edge feature of ground object height following the procedure. Then ELU activation and Max Pooling and flatten functions help us gain the one dimensional LiDAR DSM feature shown in Fig. (2).

Following the training strategy that the kernel size of convolution operations descending sort gradually shown in Fig. (3), 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, dropout operation is highlighted to avoid trained features that lack multi-scale characteristics.

The extracted multi-scale feature will be fed into the attention module. It is mainly composed of spatial attention module and channel attention module. The detail network architecture is as Fig. (4). 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:

$$\mathbf{F}'' = \mathbf{f}_{\textbf{spatial}}\left(\mathbf{F}'\right) \otimes \mathbf{F}' \tag{1}$$

$$\mathbf{F}' = \mathbf{f}_{\textbf{channel}}\left(\mathbf{F}\right) \otimes \mathbf{F} \tag{2}$$

$$\mathbf{f}_{\textbf{spatial}}\left(\mathbf{F}'\right) = \varsigma\left(\mathbf{f}_{\textbf{conv}}[\mathbf{f}_{\textbf{Avg}}\left(\mathbf{F}'\right) \oplus \mathbf{f}_{\textbf{Max}}\left(\mathbf{F}'\right)]\right) \tag{3}$$

$$\mathbf{f}_{\textbf{channel}}\left(\mathbf{F}\right) = \varsigma\left(\mathbf{f}_{\mathrm{MLP}}[\mathbf{f}_{\mathbf{Avg}}\left(\mathbf{F}\right)] \oplus \mathbf{f}_{\mathrm{MLP}}[\mathbf{f}_{\mathbf{Max}}\left(\mathbf{F}\right)]\right) \tag{4}$$

where Eq. (1) and Eq. (3) represent the spatial attention module, Eq. (2) and Eq. (4) represent the channel attention module. The operation $\otimes$ represents element-wise multiplication between features, operation $\oplus$ represents element-wise sum between features, operation ς represents ELU activation function, $\mathbf{f}_{\text{Avg}}$ represents average pooling function, $\mathbf{f}_{\textbf{Max}}$ 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. Separate descriptors will go through a weight parameter shared multilayer perception $\mathbf{f}_{\textbf{MLP}}$ with one hidden layer which would help us gain the channel attention map with $H \times 1 \times 1$ data size. Then an element-wise 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 smoother model training process.

We generate a spatial attention map to highlight the

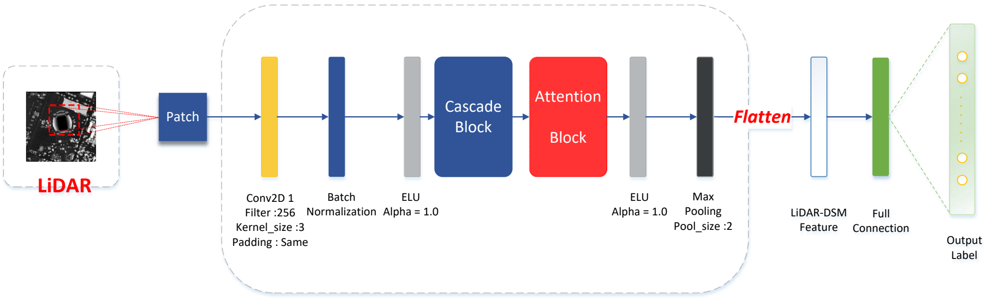


Figure 2: The network of LiDAR information extraction, the input of LiDAR data patch will go through the Cascade Block and Attention Block.

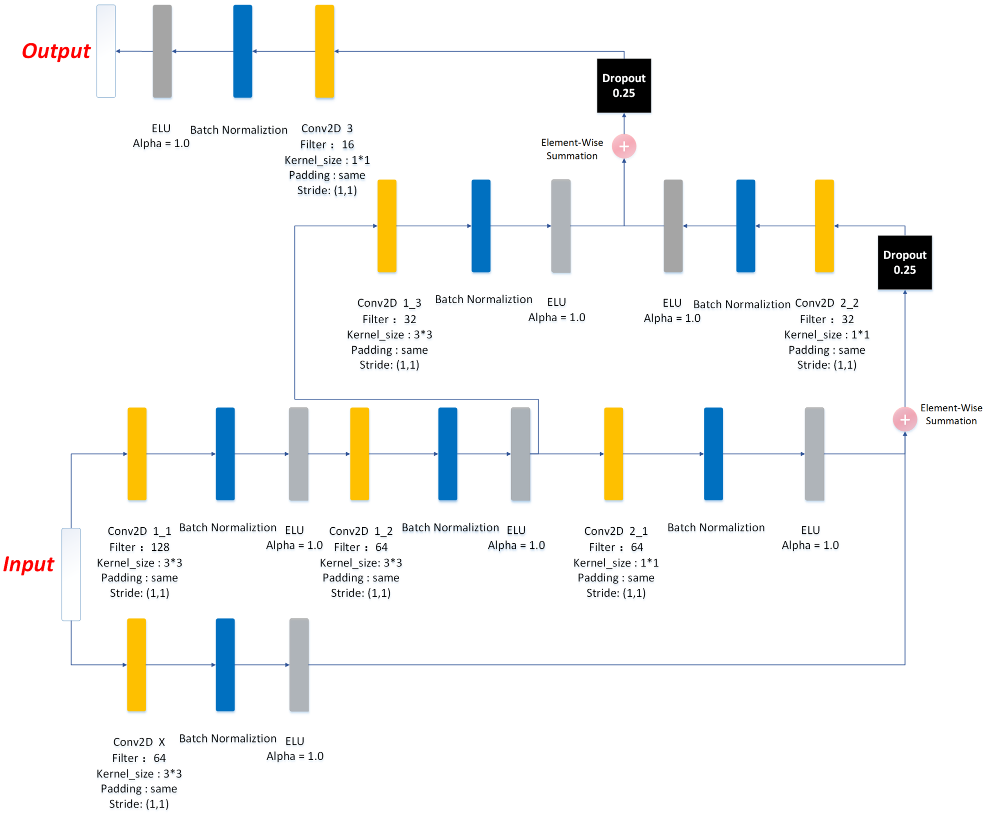


Figure 3: The Cascaded block designs skip connections between convolution layers with descending kernel size to capture multi-scale LiDAR height feature.

inter-spatial 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 element-wise summation. The extracted feature along the channel axis is then convolved and activated by the ELU function to get the final spatial attention map $\mathbf{f}_{\mathbf{spatial}}\left(\mathbf{F}'\right)$. As shown in Eqs. (1) and (2), the input feature will be multiplied by $\mathbf{f}_{\mathbf{spatial}}$ and $\mathbf{f}_{\mathbf{channel}}$ to get the enhanced feature $\mathbf{F}''$.

Table 1
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**

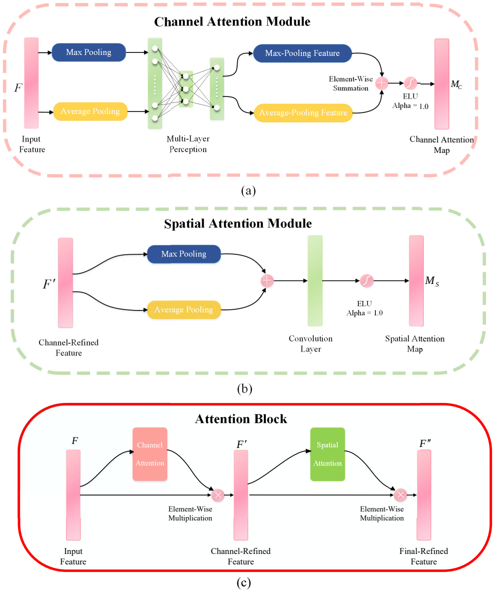


Figure 4: The details of proposed attention block: (a) shows channel module supporting the inter-channel connection of features; (b) shows spatial module strengthening the inter-spatial relationships of features; (c) shows the overall architecture by integrating above two blocks.

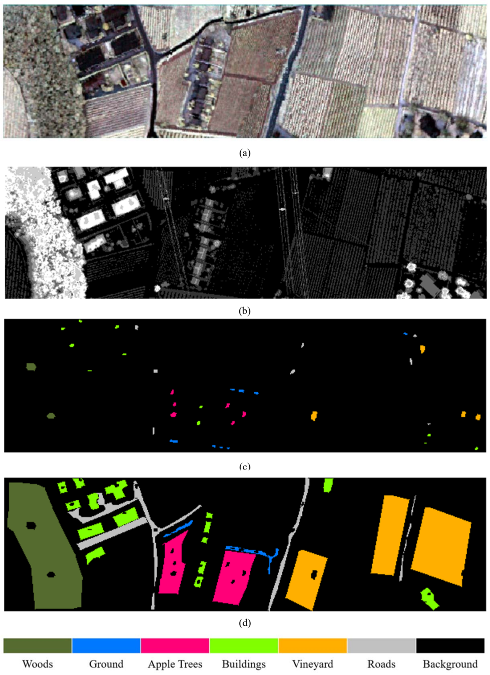


Figure 5: Trento dataset used in this experiments. (a) represents the pseudo-color hyperspectral image display; (b) represents the grey-scale LiDAR DSM data; (c) shows the training ground truth samples; (d) shows the testing ground truth samples.

3. Experiments

3.0.1. Dataset

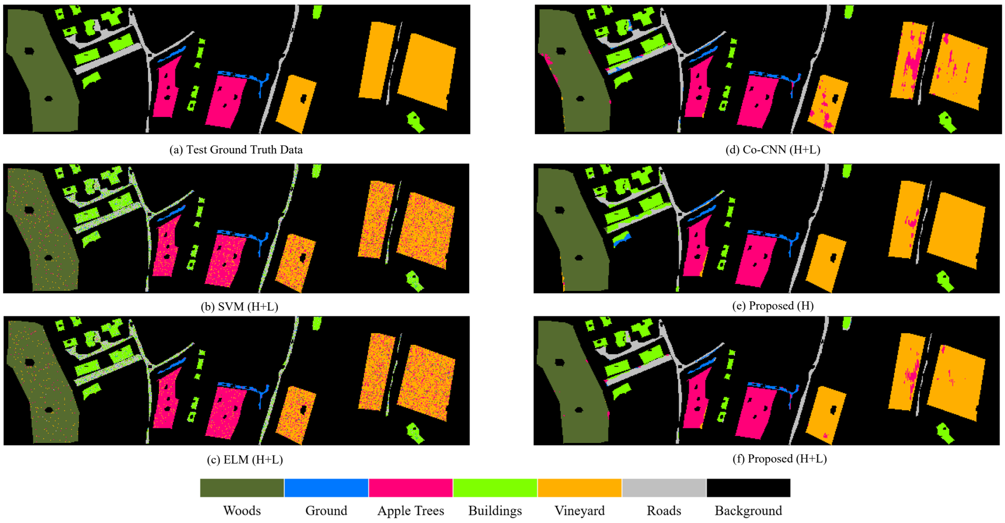
In this experiment, we have conducted our algorithm on the Trento dataset, which contains LiDAR and HSI information, to evaluate the efficiency of the cascaded convolutional neural network and attention modules.

Trento Dataset [20] is composed of HSI and LiDAR DSM data captured in Trento, Italy. The full image size is 600 × 166 with a 1-meter spatial resolution. And the HSI contains 63 bands ranging from $0.42 - 0.99\mu$m. The HSI and LiDAR data are separately captured by AISA Eagle and Optech ALTM 3100EA sensors.

3.1. Training Settings

Towards Trento multimodal remote sensing dataset, we have randomly selected half amount of training samples as the validation data to help optimize the performance during the training phase. And we have set the training batch size as 100 and the training epoch as 13.

Fine-tune strategy has been adopted in the multimodal algorithm framework to improve the performance during training. Firstly, separately training the HSI convolutional neural network and LiDAR cascaded convolutional neural network to save the trained model, and then


(a) Test Ground Truth Data

(d) Co-CNN (H+L)

(b) SVM (H+L)

(e) Proposed (H)

(c) ELM (H+L)

(f) Proposed (H+L)

Figure 6: 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).

training the multimodal framework with the saved initial parameters. Adam has been chosen as an optimizer with a 0.001 learning rate, and the optimizer parameter is 0.001 when training LiDAR data and 0.0001 for hyperspectral data. In case of over-fitting the data, we also adopt a dropout operation in the fusion stage.

3.2. Results

The proposed algorithm has been compared with classic machine learning methods including SVM [21] and ELM [22]. Besides we also introduced the fundamental Co-CNN [23] methods based on CNN to further prove the efficiency of proposed cascaded attention network. The final experiment results list in table 1.

As the classification results shown in Fig. (6), the deep learning based methods achieve better classification performance than classical machine leaning methods on both datasets. The proposed methods have achieved better performance on OA, AA and Kappa key metrics.

Our designed framework highlights the LiDAR ground objects' height information by utilizing an attention mechanism and cascaded multi-scale network. As shown in Fig. 6 (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 multi-scale cascaded network and attention mechanism to enhance precious LiDAR info to achieve 93.51% accuracy.

4. Conclusion

In this paper, our proposed cascaded attention convolution neural network has better solved the problem of using the multimodal feature with different distribution characteristics high-efficiently. The LiDAR data lack sufficient features when fused with hyperspectral on the feature level, which might harm the multimodal classification task. Our method has achieved outstanding performances on easily confusing categories and overall accuracy compared with classic machine learning methods (SVM and ELM) and other deep-learning-based methods. In the future work, we will continue to explore the perspectives of how to strengthen the weight of remote sensing sources with the weak feature or how to find a more general way to augment the weak source data for better utilizing contrasting multimodal characteristics.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62161160336 and Grant 42030111.

References

[1] J. Yao, D. Meng, Q. Zhao, W. Cao, Z. Xu, Nonconvex-sparsity and nonlocal-smoothness-based blind hy-

perspectral unmixing, IEEE Transactions on Image Processing 28 (2019) 2991–3006.

[2] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, J. Chanussot, Graph convolutional networks for hyperspectral image classification, IEEE Transactions on Geoscience and Remote Sensing 59 (2021) 5966–5978.

[3] J. Li, K. Zheng, J. Yao, L. Gao, D. Hong, Deep unsupervised blind hyperspectral and multispectral data fusion, IEEE Geoscience and Remote Sensing Letters 19 (2022) 1–5.

[4] J. Li, D. Hong, L. Gao, J. Yao, K. Zheng, B. Zhang, J. Chanussot, Deep learning in multimodal remote sensing data fusion: A comprehensive review, arXiv preprint arXiv:2205.01380 (2022).

[5] S. K. Roy, A. Deria, D. Hong, B. Rasti, A. Plaza, J. Chanussot, Multimodal fusion transformer for remote sensing image classification, arXiv preprint arXiv:2203.16952 (2022).

[6] D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, J. Chanussot, Spectralformer: Rethinking hyperspectral image classification with transformers, IEEE Transactions on Geoscience and Remote Sensing 60 (2022) 1–15. Art no. 5518615, doi: 10.1109/TGRS.2021.3130716.

[7] J. Niemeyer, F. Rottensteiner, U. Soergel, Contextual classification of lidar data and building object detection in urban areas, ISPRS journal of photogrammetry and remote sensing 87 (2014) 152–165.

[8] W. Y. Yan, A. Shaker, N. El-Ashmawy, Urban land cover classification using airborne lidar data: A review, Remote Sensing of Environment 158 (2015) 295–310.

[9] D. Hong, W. He, N. Yokoya, J. Yao, L. Gao, L. Zhang, J. Chanussot, X. Zhu, Interpretable hyperspectral artificial intelligence: When nonconvex modeling meets hyperspectral remote sensing, IEEE Geoscience and Remote Sensing Magazine 9 (2021) 52–87.

[10] M. Ahmad, S. Shabbir, S. K. Roy, D. Hong, X. Wu, J. Yao, A. M. Khan, M. Mazzara, S. Distefano, J. Chanussot, Hyperspectral image classification-traditional to deep models: A survey for future prospects, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2021).

[11] R. Huang, D. Hong, Y. Xu, W. Yao, U. Stilla, Multi-scale local context embedding for lidar point cloud classification, IEEE Geoscience and Remote Sensing Letters 17 (2019) 721–725.

[12] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, B. Zhang, More diverse means better: Multimodal deep learning meets remote-sensing imagery classification, IEEE Transactions on Geoscience and Remote Sensing 59 (2021) 4340–4354.

[13] J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, J. Li, Visual attention-driven hyperspectral image classification, IEEE Transactions on Geoscience and Remote Sensing 57 (2019) 8065–8080.

[14] J. Yao, D. Hong, J. Chanussot, D. Meng, X. Zhu, Z. Xu, Cross-attention in coupled unmixing nets for unsupervised hyperspectral super-resolution, in: European Conference on Computer Vision, Springer, 2020, pp. 208–224.

[15] N. Liu, J. Han, M.-H. Yang, Picanet: Learning pixel-wise contextual attention for saliency detection, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3089–3098.

[16] Z. Han, D. Hong, L. Gao, J. Yao, B. Zhang, J. Chanussot, Multimodal hyperspectral unmixing: Insights from attention networks, IEEE Transactions on Geoscience and Remote Sensing 60 (2022) 1–13.

[17] P. Ghamisi, B. Höfle, X. X. Zhu, Hyperspectral and lidar data fusion using extinction profiles and deep convolutional neural network, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 10 (2017) 3011–3024. doi:10.1109/JSTARS.2016.2634863.

[18] Y. Chen, C. Li, P. Ghamisi, X. Jia, Y. Gu, Deep fusion of remote sensing data for accurate classification, IEEE Geoscience and Remote Sensing Letters 14 (2017) 1253–1257. doi:10.1109/LGRS.2017.2704625.

[19] D. Hong, L. Gao, R. Hang, B. Zhang, J. Chanussot, Deep encoder-decoder networks for classification of hyperspectral and lidar data, IEEE Geoscience and Remote Sensing Letters (2020).

[20] B. Rasti, P. Ghamisi, R. Gloaguen, Hyperspectral and lidar fusion using extinction profiles and total variation component analysis, IEEE Transactions on Geoscience and Remote Sensing 55 (2017) 3997–4007.

[21] C. Cortes, V. Vapnik, Support vector machine, Machine learning 20 (1995) 273–297.

[22] G.-B. Huang, Q.-Y. Zhu, C.-K. Siew, Extreme learning machine: A new learning scheme of feedforward neural networks, in: 2004 IEEE international joint conference on neural networks (IEEE Cat. No. 04CH37541), volume 2, IEEE, 2004, pp. 985–990.

[23] R. Hang, Z. Li, P. Ghamisi, D. Hong, G. Xia, Q. Liu, Classification of hyperspectral and lidar data using coupled cnns, IEEE Transactions on Geoscience and Remote Sensing 58 (2020) 4939–4950. doi:10.1109/TGRS.2020.2969024.