SPECTRAL-SPATIAL SYMMETRICAL AGGREGATION CROSS-LINKING MULTI-MODAL DATA FUSION NETWORK

Jinping Wang[1]*, Jun Li*[1]*, and Xiaojun Tan*[2,*]

[1]School of Intelligent Systems Engineering, Sun Yat-sen University, Guangzhou, 510006, China,
[2]School of Intelligent Systems Engineering, Sun Yat-sen University & Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai, 519082, China

ABSTRACT

In this paper, a spectral-spatial symmetrical aggregation cross-linking network (SACLNet) is developed for multi-modal data classification, which contains three modules as follows. First, the *Spectro-Spatial Feature Learning Module* is proposed, using the involution operation sliding over the spectral channels of hyperspectral image (HSI) and fused-sharing weight obtained from HSI and light detection and ranging (HSI-LiDAR) data for spectral and spatial information representation. Second, the *pyramid feature fusion module* interacts among spectral and spatial information to further share and guide each other. In this step, multistage features, including low-level, middle-level, and high-level, are fused and adjusted in a pyramidal and mutually guided learning process. Third, the fused Spectro-Spatial features are embedded into the *Multimodal Data Fusion Module*, obtaining the final classification results. Experimental results show that the proposed SACLNet has a satisfactory classification performance than the state-of-the-art methods.

Index Terms— Involution operation, cross-linking network, data fusion

1. INTRODUCTION

Recently, the combination of multi-modal data, especially hyperspectral image (HSI) and light detection and ranging (LiDAR) image data fusion, has been developed and successfully used in many applications [1, 2].

Deep learning methods are becoming more and more popular on multi-modal data fusion tasks. Generally, there are two categories for data fusion framework; i,e., the disjoint stream fusion strategy and cross-learning stream fusion strategy [3–6]. Specifically, the disjoint stream fusion strategy uses the two-stream network to independently extract spectral and spatial features. For example, a new unsupervised network using patch-to-patch convolutional neural network (CNN) strategy is developed for data fusion classification [7], which can also provide higher classification accuracies than a two branch CNN [8]. In 2020, a spatial, spectral, and multi-scale attention mechanism is proposed for HSI and LiDAR complementarity information classification [9], which possesses better-enhanced feature representation among spectral and spatial domains. By contrast, the cross-learning stream fusion strategy aims to extract the HSI and LiDAR features from each other [10]. For instance, [11], a coupled CNN is raised to integrate heterogeneous features at feature-level and decision-level using parameter-sharing strategy. Moreover, the concept of cross-attention is introduced into multi-model data fusion task [12], making a better collaborative feature learning ability.

However, there are still some issues that are overlooked in existing networks. First, they usually consider the channel-wise feature relationships independently. Second, the existing methods ignore the complementarity characteristics in multi-sources spatial feature representations. Third, the features generated in hierarchical stages are combined through a simple concatenating strategy. Meanwhile, the semantic relatedness of the same stage's features is restricted. Therefore, to conquer the issues, this paper develops a spectral-spatial symmetrical aggregation cross-linking network (SACLNet). The main contributions can be concluded as follows:

1) It is the first study that introduces the involution operation for multi-modal spectral characteristic representation. It shows a more robust ability for individual samples' channel-wise feature excavation and measuring when compared with traditional convolution operation.

2) Instead of extracting spatial features of multi-resources independently, this paper develops a spatial fusion encoding to integrate the spatial information of two modalities, which could make a better feature spatial expressions.

3) Comparison with simple element-wisely adding or concatenating strategies, it's the first time that a pyramidal and cross-linking strategy is used to integrate different hierarchical features, effectively promoting the proposed network's representative ability.

This work was supported by the Key-Area Research and Development Program of Guangdong Province under Grant 2020B090921003 and the Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai) under Grant SML2020SP011. (* Corresponding author: Xiaojun Tan. e-mail: tanxj@mail.sysu.edu.cn)

2. THE PROPOSED APPROACH

Suppose $\mathbf{X}_{hsi} \in \mathbb{R}^{M\times N\times D}$ represents the input HSI and $\mathbf{X}_{lid} \in \mathbb{R}^{M\times N\times 1}$ is its corresponding input LiDAR data, where M and N refer to image width and image height of input data respectively and D is HSI channel numbers. After principal component analysis operation, $\mathbf{X}_{hsi} \in \mathbb{R}^{M\times N\times d}$. Creating a new HSI-LiDAR pair data $\mathbf{X}$ as training set, any corresponding image patch is noted as $\mathbf{X} = \left\{\mathbf{X}^i | (\mathbf{X}^i_{hsi}, \mathbf{X}^i_{lid})\right\}^T_{i=1}$ with shape of $m \times n$, and T means training sample numbers. Set ith image patch chosen from $\mathbf{X}_{hsi}$ and its corresponding $\mathbf{X}_{lid}$ as example, $\mathbf{X}^i_{hsi} \in \mathbb{R}^{m\times n\times d}$ and $\mathbf{X}^i_{lid} \in \mathbb{R}^{m\times n\times 1}$ represent the ith HSI and LiDAR patches. $\mathbf{Y} = \left\{y^i\right\}^T_{i=1}$ is the groundtruth labels for $\mathbf{X}$, and $y^i \in \{1,2,\ldots,C\}$, and C is the largest class number.

2.1. Spectro-Spatial Feature Learning Module (SSFL)

SSFL module is composed of three sub-modules, namely hyperspectral feature extractor, spectral weight encoding, and spatial weight encoding, see Fig 1.

Hyperspectral Feature Extractor: After three convolutional layers with 16, 32, and 64 filters, respectively, the primary features $\mathbf{F}_{hsi}$ of input data can be obtained, and $\mathbf{F}_{hsi}$ is of size $m \times n \times 64$.

Spectral Weight Encoding: A new involution [13] operation is used here for HSI's spectral information extraction. Suppose $\mathbf{X}^{i_x,i_y}_{hsi} \in \mathbb{R}^{1\times 1\times d}$ is one element located at (x,y) of image patch $\mathbf{X}^i_{hsi}$. For gth involution kernel $\mathscr{H}^{\cdot,\cdot,\cdot,\cdot,g} \in \mathbb{R}^{p\times p}$, it is specifically tailored for $\mathbf{X}^{i_x,i_y,g}_{hsi}$, where $g = 1,2,...,G$ and G is the largest channel group numbers and the same group shares the same involution kernel. $p \times p$ denotes a kernel size. The spectral weight encoding process is described as the following six steps, i.e., channel grouping, involution kernel, channel-to-space rearrangement generation, tensor multiplication, neighbor summation and dimension match. The detailed equation is shown as follows:

$$\begin{aligned}\mathfrak{R}^{i_x,i_y,g}_{spe} &= \phi_{spe}(\mathbf{X}^{i_x,i_y,g}_{hsi}) \\ &= \sum_{(u,v)\in\Delta_P} \mathscr{H}^{i_x,i_y,u+\lfloor p/2\rfloor,v+\lfloor p/2\rfloor,g}\mathbf{X}^{i_x+u,i_y+v,g}_{hsi},\end{aligned} \tag{1}$$

where $\Delta_P = [-\lfloor p/2\rfloor,...,\lfloor p/2\rfloor] \times [-\lfloor p/2\rfloor,...,\lfloor p/2\rfloor]$. For more details about $\phi_{spe}(\cdot)$, its kernel generation function is generated from $\mathbb{R}^d \mapsto \mathbb{R}^{p\times p\times G}$ following the equation (2).

$$\mathscr{H}^{i_x,i_y} = \phi_{spe}\left(\mathbf{X}^{i_x,i_y}_{hsi}\right) = \mathbf{W}_1\sigma\left(\mathbf{W}_0\mathbf{X}^{i_x,i_y}_{hsi}\right) \tag{2}$$

In this formula, $\mathbf{W}_0 \in \mathbb{R}^{d/r\times d}$ and $\mathbf{W}_1 \in \mathbb{R}^{(p\times p\times G)\times d/r}$ represent two linear transformations, where r is the intermediate channel dimension, and σ implies Batch Normalization and non-linear activation functions. After a linear transformation operation, the spectral feature weights $\mathfrak{R}_{spe}$ for image patch $\mathbf{X}^i_{hsi}$ is obtained as $\mathfrak{R}_{spe} = \{\mathfrak{R}^{i,g}_{spe}\}^G_{g=1} \in \mathbb{R}^{m\times n\times 64}$, where

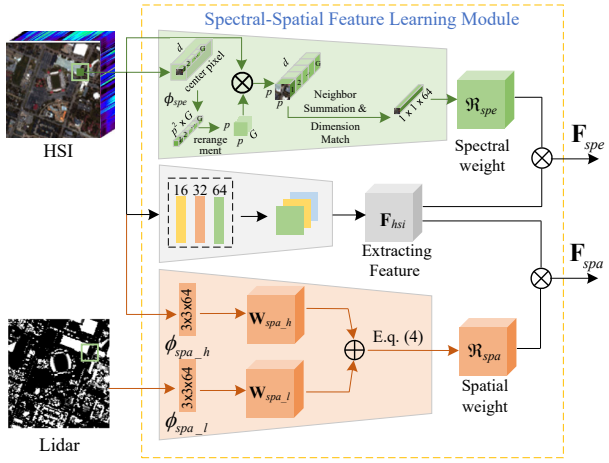

Fig. 1. The architecture of Spectro-Spatial Feature Learning Module.

$\mathfrak{R}^{i,g}_{spe}$ represents the gth group spectral-level feature weight for pixel $\mathbf{X}^{i,g}_{spe}$ and $g = 1,2,...,G$. Therefore, the initial spectral features can be obtained by multiplying the spectral feature weights $\mathfrak{R}_{spe}$ with the primary features $\mathbf{F}_{hsi}$ as follows:

$$\mathbf{F}_{spe} = \mathfrak{R}_{spe} \otimes \mathbf{F}_{hsi}, \tag{3}$$

where $\mathbf{F}_{spe} \in \mathbb{R}^{m\times n\times 64}$ denotes the initial spectral features, and $\otimes$ is the element-wise tensor multiplication operation.

Spatial Weight Encoding: Considering the complementarity of HSI and LiDAR in spatial feature domain, they can be captured jointly for better spatial information representation. Specifically, the classical convolution is used for sliding over the HSI-LiDAR pair $\mathbf{X}^i = (\mathbf{X}^i_{hsi}, \mathbf{X}^i_{lid})$. $\mathscr{F}^{\cdot,\cdot,\cdot,\cdot,k}_{spa\_h} \in \mathbb{R}^{p\times p\times d}$ and $\mathscr{F}^{\cdot,\cdot,\cdot,k}_{spa\_l} \in \mathbb{R}^{p\times p\times 1}$ represent the kth convolution kernel of the input HSI and LiDAR. $p \times p$ represents the filter kernel size. In this way, the initial spatial feature weights can be obtained by fusing HSI and LiDAR in spatial feature domain, as below.

$$\begin{aligned}\mathfrak{R}^{i_x,i_y,k}_{spa} = &\sum_{c=1}^{d}\sum_{(u,v)\in\Delta p} \mathscr{F}^{u+\lfloor p/2\rfloor,v+\lfloor p/2\rfloor,c,k}_{spa\_h}\mathbf{X}^{i_x+u,i_y+v,c}_{hsi} + \\ &\sum_{(u,v)\in\Delta p} \mathscr{F}^{u+\lfloor p/2\rfloor,v+\lfloor p/2\rfloor,k}_{spa\_l}\mathbf{X}^{i_x+u,i_y+v}_{lid},\end{aligned} \tag{4}$$

where $\mathfrak{R}_{spa} = \{\mathfrak{R}^k_{spa}\}^{64}_{k=1}$, it denotes the spatial-level feature weight of the HSI-LiDAR pair $\mathbf{X}^i$. To obtain the final spatial features $\mathbf{F}_{spa} \in \mathbb{R}^{m\times n\times 64}$, we multiply the spatial-level feature weight $\mathfrak{R}_{spa}$ by the primary feature $\mathbf{F}_{hsi}$ as follows:

$$\mathbf{F}_{spa} = \mathfrak{R}_{spa} \otimes \mathbf{F}_{hsi}. \tag{5}$$

2.2. Pyramid Feature Fusion Module (PFF)

In this paper, spectral-alone, spatial-alone, and Spectro-Spatial interactions alternately act on the stream of infor-

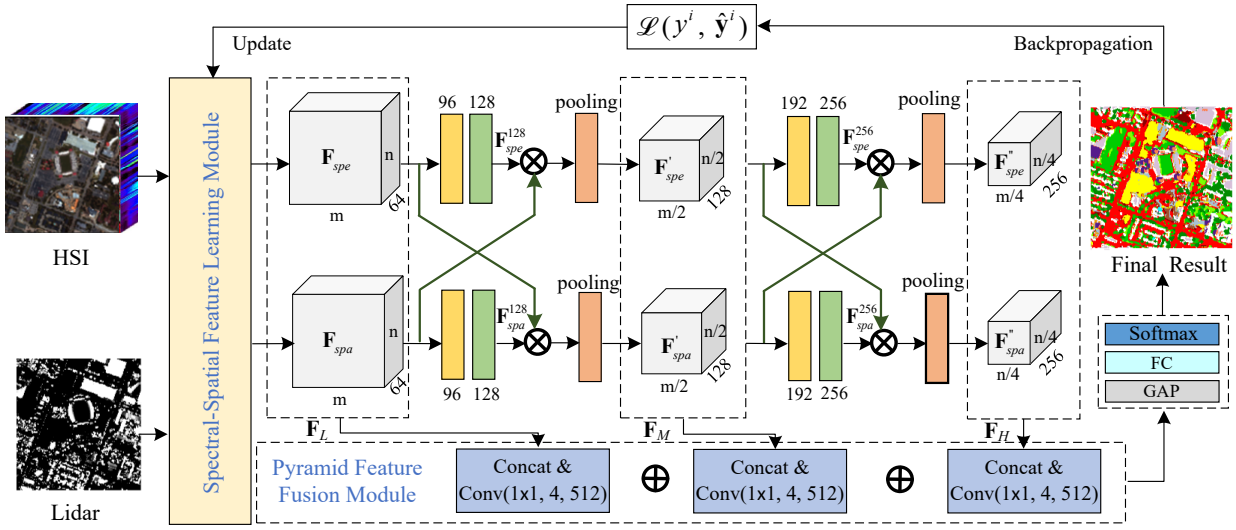

Fig. 2. The framework of the proposed method.

Table 1. Summary of hyper-parameters of the SACLNet.

Hyper-parameters	Value	Hyper-parameters	Value
Spatial Patch	25* 25	Dimension	15
Group Number	5	Optimizer	ADAM
Learning Rare	0.01	Padding	SAME
Training Epoch	100	Activation	ReLU

Table 2. The classification accuracies with different Spectral Weight Encoding methods.

Methods	OA	AA	Kappa
Conv-1D	89.93(2.36)	88.91(2.76)	88.74(5.61)
Involution	**96.59(0.31)**	**96.23(0.37)**	**97.01(0.34)**

mation propagation, collaboratively facilitating the representation capability of network architectures.

As shown in Fig. 2, the deep PFF-based feature learning process includes three components: the low-level local texture features $\mathbf{F}_L$, the middle-level interpretable features $\mathbf{F}_M$ and the high-level global semantic features $\mathbf{F}_H$. Specifically, $\mathbf{F}_L$ are generated by the combination of the low-level output layers as $\mathbf{F}_L$=$\mathbf{F}_{spe} \oplus \mathbf{F}_{spa}$, where $\oplus$ denotes tensor concatenation operation along the channel axis. Similarly, $\mathbf{F}_M$ and $\mathbf{F}_H$ are generated through the proposed PFF module, using $\mathbf{F}_L$, $\mathbf{F}_M$, respectively, as the inputs. $\mathbf{F}_M = \text{PFF}(\mathbf{F}_L)$, and $\mathbf{F}_H = \text{PFF}(\mathbf{F}_M)$, where PFF$(\cdot)$ denotes the PFF operations. To ensure that the three features have the same number of feature maps, the dimension-matching function $g(\cdot)$ is implemented on the feature maps before feature fusion, and then concatenate cross features. And the final fusion feature could be obtained by $\mathbf{O} = g_1(\mathbf{F}_L) + g_2(\mathbf{F}_M) + g_3(\mathbf{F}_H)$.

2.3. multi-modal Data Fusion Module (MDF)

For the ith sample, the maximum probability based on the prediction vector $\hat{\mathbf{y}}^i \in \mathbb{R}^{1\times C}$ is used for classification, where C is the total number of groundtruth classes. During the inference process, we use the *Crossentropy* loss function to evaluate the degree of inconsistency between the prediction results of model $\hat{\mathbf{y}}$ and the ground truth $\mathbf{y}$. The formula is as follows:

$$\mathscr{L}(\mathbf{y},\hat{\mathbf{y}}) = \frac{1}{N}\sum_i L_i = -\frac{1}{N}\sum_i \sum_{c=1}^{C} \mathbf{y}_{i,c} \log(\hat{\mathbf{y}}_{i,c}). \quad (6)$$

The default parameters of the SACLNet are shown in Table 1.

3. EXPERIMENTAL RESULTS AND ANALYSIS

We train and test our method on Houston dataset. It is composed of a HSI and a LiDAR-derived DSM, where the hyperspectral data were captured by an AISA Eagle sensor [14]. The dataset is 349 $\times$ 1905 pixels in size with a spatial resolution of 2.5 m, yielding 15 different classes in total.

A detailed comparison for different models, including the extreme learning machine (ELM) [15], deep convolution network (DeepCNN) [11], dual attention-based multi-modal fusion network (FusAtNet) [12], deep encoder-decoder network (EndNet) [16], and hierarchical random walk network (HRWN) [17], are established. All the compared methods use the default parameters provided by their original papers. For a fair comparison, experiment is repeated approximately 20 times to achieve reasonable results. Regarding the division of the dataset, we have adopted a general division method [12, 16, 17]. Here, the overall accuracy (OA), average accuracy (AA), and Kappa are used to evaluate the classifica-

Table 3. The classification accuracies as percentages for the different methods as averages after 20 repeated experiments. The number in parentheses indicates the standard variance for the repeated experiments.

Indexes	ELM [15]	DeepCNN [11]	FusAtNet [12]	EndNet [16]	HRWN [17]	SACLNet-H	SACLNet
OA	83.57(0.82)	86.97(0.77)	87.88(2.92)	90.72(1.71)	94.44(0.64)	92.08(0.84)	**96.59(0.31)**
AA	84.21(0.89)	86.78(0.41)	89.01(1.91)	91.93(2.15)	95.38(0.41)	92.93(0.81)	**96.23(0.37)**
Kappa	82.24(0.90)	86.69(0.76)	86.93(3.11)	89.94(1.96)	94.16(0.60)	93.39(0.88)	**97.01(0.34)**
Training Time (s)	7.15	131.94	826.87	91.36	45.06	54.82	58.29
Parameters (M)	-	3.57	38.11	0.27	1.44	2.64	2.65

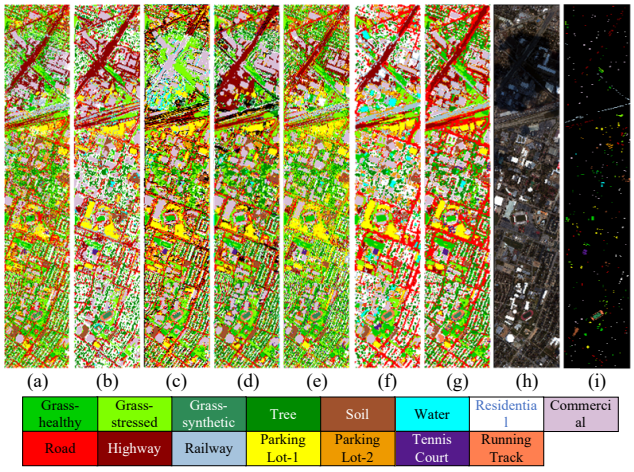

Fig. 3. Classification maps: (a) ELM, (b) DeepCNN, (c) FusAtNet, (d) EndNet, (e) HRWN, (f) SACLNet-H, (g) SACLNet, (h) pseudocolor images, and (i) GroundTruth Map.

Fig. 4. Local compared classification maps: (a) FusAtNet, (b) EndNet, (c) HRWN, and (d) SACLNet.

tion performance. All the experiments are implemented with an Intel i7-10700F processor and a NVIDIA 2070 GPU.

In addition, an ablation experiment is conducted on Houston dataset to evaluate the classification performance of the SACLNet under different spectral weight encoding methods, which can be shown in Table 2. As seen, instead of using Conv-1D to extract the spectral information of HSI under the Houston dataset, the involution can reach 96.59% classification accuracy, which is higher than that of Conv-1D (89.93%).

The classification accuracies and the corresponding classification maps obtained by all methods are recorded in Table 3 and Fig. 3. A local compared classification maps are given in Fig. 4. As seen, the proposed method can obtain the highest classification accuracies in terms of the OA, AA, and Kappa

Table 4. Performance comparison with the state-of-the-art neural network-based methods.

Model	OA	AA	Kappa
SLRCA [18]	91.30	91.95	90.56
OTVCA [19]	92.45	92.68	91.81
ODF-ADE [20]	93.50	94.07	92.99
E-UGF [21]	95.11	94.57	94.47
Deep Fusion [22]	91.32	91.96	90.57
PToPCNN [7]	92.48	93.55	91.87
EndNet [16]	90.72	91.93	89.94
HRWN [17]	94.44	95.38	94.16
SACLNet	96.59	96.23	97.01

and the best classification map when compared with other methods. Especially, SACLNet-H and SACLNet represent the classification performance of single HSI and HSI-LiDAR pair, respectively. compared with the SACLNet-H, the proposed SACLNet possesses a classification accuracy that is more than 4.51 points higher, indicating the advantages of fusion data in distinguishing different surface objects. In addition, compared with deep learning-based models, the training time of the proposed method also has certain advantages. At last, Table 4 reports more detailed comparison results for the state-of-the-art models developed in last five years, recorded according to [11]; Experimental results indicate that the proposed SACLNet indeed has an impressive feature learning strategy in feature transmission and extraction processes.

4. CONCLUSIONS

We develop a new framework, namely SACLNet, for data fusion and classification. Specifically, this paper uses the involution operation for spectral information extraction to explore the channel relevance among the spectrum channels more effectively than traditional CNNs. Second, this paper describes the spatial information in a multi-modal-oriented manner, so SACLNet can make full use of the complementarity of different modalities, thereby increasing its spatial features representation ability. In addition, a pyramidal and mutually guided learning process is conducted on hierarchical spectral-spatial information to generate more suitable and discriminate features for specific targets. Experimental results indicate the superiority of the SACLNet network, which also suggests that the SACLNet is easily developed for real data fusion applications.

5. REFERENCES

[1] H. Ma, G. Liu, and Y. Yuan, "Enhanced non-local cascading network with attention mechanism for hyperspectral image denoising," in *Proc. 2020 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2020, pp. 2448–2452.

[2] E. Pan, Y. Ma, X. Mei, F. Fan, and J. Ma, "Unsupervised stacked capsule autoencoder for hyperspectral image classification," in *Proc. 2021 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2021, pp. 1825–1829.

[3] H. Lee and H. Kwon, "Going deeper with contextual cnn for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4843–4855, 2017.

[4] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, 2019.

[5] 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, 2016.

[6] Y. Chen, C. Li, P. Ghamisi, C. Shi, and Y. Gu, "Deep fusion of hyperspectral and lidar data for thematic classification," in *proc. 2016 IEEE int. geosci. remote sens. symp. (IGARSS)*. IEEE, 2016, pp. 3591–3594.

[7] M. Zhang, W. Li, Q. Du, L. Gao, and B. Zhang, "Feature extraction for classification of hyperspectral and lidar data using patch-to-patch cnn," *IEEE Trans. Cybern.*, vol. 50, no. 1, pp. 100–111, 2018.

[8] 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, 2017.

[9] H. Li, W. Hu, W. Li, J. Li, Q. Du, and A. Plaza, "A^3CLNN: Spatial, spectral and multiscale attention convlstm neural network for multisource remote sensing data classification," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–15, 2020, Early Access, DOI: 10.1109/TNNLS.2020.3028945.

[10] J. Xia and Z. Ming, "Classification of hyperspectral and lidar with deep rotation forest," in *Proc. 2019 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2019, pp. 2197–2201.

[11] 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, 2020.

[12] S. Mohla, S. Pande, B. Banerjee, and S. Chaudhuri, "FusAtNet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification," in *Proc. 2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 92–93.

[13] D. Li, J. Hu, C. Wang, X. Li, Q. She, L. Zhu, T. Zhang, and Q. Chen, "Involution: Inverting the inherence of convolution for visual recognition," in *Proc. 2021 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 12 321–12 330.

[14] J. Xia, N. Yokoya, and A. Iwasaki, "A novel ensemble classifier of hyperspectral and lidar data using morphological features," in *Proc. 2017 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2017, pp. 6185–6189.

[15] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1-3, pp. 489–501, 2006.

[16] 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.*, pp. 1–5, 2020, Early Access, DOI: 10.1109/LGRS.2020.3017414.

[17] X. Zhao, R. Tao, W. Li, H.-C. Li, Q. Du, W. Liao, and W. Philips, "Joint classification of hyperspectral and lidar data using hierarchical random walk and deep cnn architecture," *Neurocomputing*, vol. 58, no. 10, pp. 7355–7370, 2020.

[18] 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, 2017.

[19] 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, 2017.

[20] 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, p. 868, 2017.

[21] J. Xia, N. Yokoya, and A. Iwasaki, "Fusion of hyperspectral and lidar data with a novel ensemble classifier," *IEEE Geosci. Remote. Sens. Lett.*, vol. 15, no. 6, pp. 957–961, 2018.

[22] 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, 2017.