This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Deep Encoder–Decoder Networks for Classification of Hyperspectral and LiDAR Data

Danfeng Hong, *Member, IEEE*, Lianru Gao, *Senior Member, IEEE*, Renlong Hang, *Member, IEEE*, Bing Zhang, *Fellow, IEEE*, and Jocelyn Chanussot, *Fellow, IEEE*

***Abstract*— Deep learning (DL) has been garnering increasing attention in remote sensing (RS) due to its powerful data representation ability. In particular, deep models have been proven to be effective for RS data classification based on a single given modality. However, with one single modality, the ability in identifying the materials remains limited due to the lack of feature diversity. To overcome this limitation, we present a simple but effective multimodal DL baseline by following a deep encoder–decoder network architecture, EndNet for short, for the classification of hyperspectral and light detection and ranging (LiDAR) data. EndNet fuses the multimodal information by enforcing the fused features to reconstruct the multimodal input in turn. Such a reconstruction strategy is capable of better activating the neurons across modalities compared with some conventional and widely used fusion strategies, e.g., early fusion, middle fusion, and late fusion. Extensive experiments conducted on two popular hyperspectral and LiDAR data sets demonstrate the superiority and effectiveness of the proposed EndNet in comparison with several state-of-the-art baselines in the hyperspectral-LiDAR classification task. The codes will be available at https://github.com/danfenghong/IEEE_GRSL_EndNet, contributing to the RS community.**

***Index Terms*— Classification, deep learning (DL), encoder–decoder, hyperspectral, light detection and ranging (LiDAR), multimodality, remote sensing (RS).**

I. Introduction

With the rapid development of different imaging techniques, spaceborne or aerial remote sensing (RS) data acquired by either active, e.g., synthetic aperture radar (SAR) [1], light detection and ranging (LiDAR) [2], or passive devices, e.g., hyperspectral imagery (HSI) [3] and multispectral imagery (MSI), become freely and openly available on an increasingly larger scale. However, using single modalities may limit the classification performances in various applications. For example, the MSI fails to finely discriminate those classes that hold very similar spectral signatures, or the HSI is not able to distinguish the differences between the material "grass" on the ground and one on the roof due to the missing elevation information or the effects of various spectral variabilities [4]. As a result, the joint use of multimodal data enables the recognition of materials of interest at a more accurate level.

Up until now, there have been many successful showcases by the means of multimodal RS data to enhance the spatial and spectral resolutions, e.g., HSI and MSI fusions [5], to effectively identify the materials with different heights, e.g., by using HSI and LiDAR data [6], or to provide the geometrically compensatory information by fusing the optical and SAR data [7]. In this letter, we pay more attention to the classification task of HSI and LiDAR data. In [8], the morphological profiles (MPs) are used to fuse the HSI and LiDAR data on the subspace obtained by a graph-based embedding technique. Rasti *et al.* [9] extracted more advanced extinction profiles (EPs) for the fusion of HSI and LiDAR data. Similarly, the same investigators further extended their fusion study by jointly modeling sparse and low-rank properties [10]. Inspired by the recent success of deep learning (DL) techniques, Chen *et al.* [11] first employed two convolutional neural networks (CNNs) to extract feature representations from HSI and LiDAR data, respectively, and then, one fully connected (FC) network is used over the extracted features for further fusion.

Different from the concatenation-based fusion strategy introduced earlier, the idea of the learning shared or common subspace representations [12] has been paid growing attention in the multi-modality learning (MML) task. This topic can be roughly categorized into image- and feature-level learnings. A typical image-level work was proposed in [13] to enhance the spectral resolution of the MSI using partially overlapped HSI. For the feature-level one, Hong *et al.* [14], for the first time, proposed to learn a common subspace from the limited HS-MS correspondences, and the learned subspace can be seen as a "bridge" to well transfer different information across modalities. Hong *et al.* [15] further improved their work by adaptively learning an aligned graph structure, yielding a state-of-the-art semisupervised model. Very recently,

Manuscript received March 25, 2020; revised June 10, 2020 and July 9, 2020; accepted August 14, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB0500502, in part by the National Natural Science Foundation of China under Grant 41722108, and in part by AXA Research Fund. *(Corresponding author: Lianru Gao.)*

Danfeng Hong is with the Université Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France (e-mail: hongdanfeng1989@gmail.com).

Lianru Gao is with the Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China (e-mail: gaolr@aircas.ac.cn).

Renlong Hang is with the Jiangsu Key Laboratory of Big Data Analysis Technology, School of Automation, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: renlong_hang@163.com).

Bing Zhang is with the Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, and also with the College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: zb@radi.ac.cn).

Jocelyn Chanussot is with the Université Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France, and also with Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China (e-mail: jocelyn@hi.is).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2020.3017414

1545-598X © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Jocelyn Chanussot. Downloaded on August 31,2020 at 15:21:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

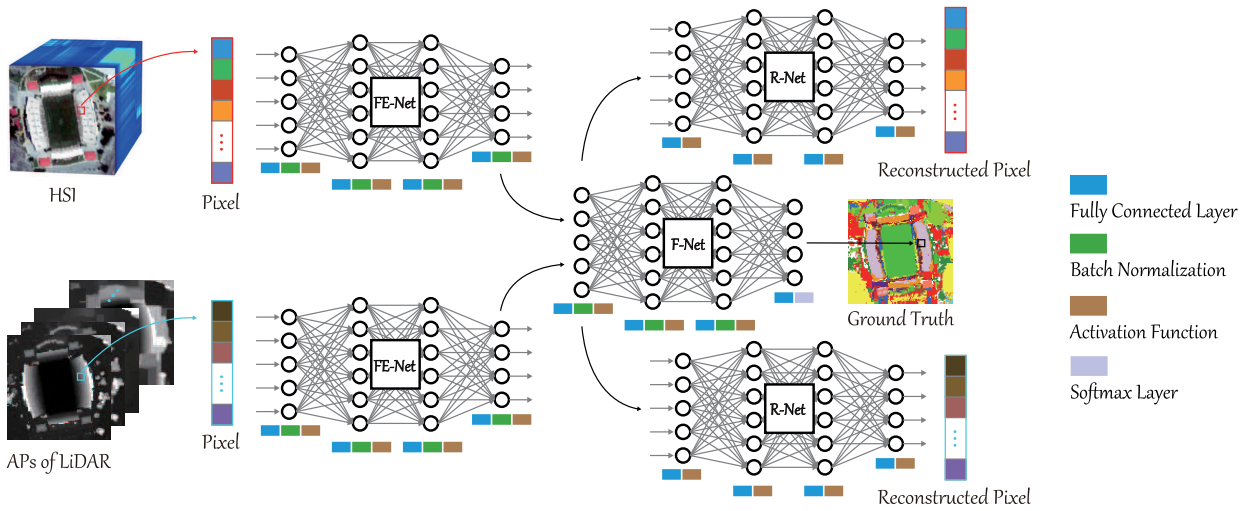

Fig. 1. Illustration of the proposed EndNet for classificaiton of HSI and LiDAR data.

Li *et al.* [16] proposed a robust capsule network by utilizing maximum correntropy criterion with the applications to HS image classification as well as HS and LiDAR data joint classification. However, there is still room for improvement in feature representations, e.g., more compact information blending and more effective information transfer.

To overcome this challenge, we develop a simple but effective deep network, which is shaped based on an encoder–decoder architecture, called EndNet, for classification of HSI and LiDAR data. EndNet makes full use of the advantages of the encoder–decoder module, aiming at fusing HSI and LiDAR data in a more compact fashion. It is worth noting that the proposed EndNet is an FC network. The main reasons lie in two aspects. One is to avoid the issue of information leak (e.g., using CNNs) when the overlap occurs between training and testing samples [17], [18]. Another reason that we expect to clarify is that our main motivation or goal is to provide a novel solution (fusion baseline) for the HSI and LiDAR classification beyond previous concatenation-dominated ones.

II. Methodology

Fig. 1 illustrates the proposed EndNet framework that consists of three parts: feature extraction (or encoder) network (FE-Net), fusion module connecting with labels (F-Net), and reconstruction (or decoder) network (R-Net).

A. FE-Net

As the name suggests, the FE-Net extracts the features for each modality in a hierarchical manner. Given two input modalities $\mathbf{X}_1 \in \mathbb{R}^{d_1 \times N}$ and $\mathbf{X}_2 \in \mathbb{R}^{d_2 \times N}$ with d_1 and d_2 dimensions by N pixels, the extracted features for the ith pixel in the lth layer, denoted as $\mathbf{z}_{s,i}^{(l)}$, can be written as

$$\mathbf{z}_{s,i}^{(l)} = \begin{cases} f\left(\mathbf{W}_s^{(l)}\mathbf{x}_{s,i} + \mathbf{b}_s^{(l)}\right), & l = 1 \\ f\left(\mathbf{W}_s^{(l)}\mathbf{z}_{s,i}^{(l-1)} + \mathbf{b}_s^{(l)}\right), & l = 2, \ldots, p \end{cases} \tag{1}$$

where $s = 1, 2$ stands for different modalities; $f(\cdot)$ is the nonlinear activation function, e.g., well-known sigmoid or ReLU, with respect to the to-be-estimated network parameters: $\{\mathbf{W}_s^{(l)}\}_{l=1}^{p}$ and $\{\mathbf{b}_s^{(l)}\}_{l=1}^{p}$. To speed up the network optimization and avoid the problems of the gradient vanishment and explosion in the process of updating parameters, we adopt the batch normalization (BN) layer [19] prior to the use of the activation function.

B. F-Net

We start to fuse the features extracted from FE-Net in the F-Net. This fused features $\mathbf{a}_i$ can be formulated as follows:

$$\mathbf{a}_i^{(l+1)} = h\left(\left[\mathbf{z}_{1,i}^{(l)}, \mathbf{z}_{2,i}^{(l)}\right]\right). \tag{2}$$

Similar to (1), $h(\cdot)$ can be seen as a set of "Block" that consists of the encoder layer, the BN layer, and the ReLU activation layer. $[\bullet, \bullet]$ denotes a common concatenation operator. The output in F-Net is connected with one-hot encoded labels through a softmax layer.

C. R-Net

Inspired by the recent success of the encoder–decoder architecture in feature fusion [20], we enforce to reconstruct the input modalities from the fused features to improve the fusion level of the features extracted from FE-Net, resulting in the following formulation:

$$\mathbf{x}_{s,i} = g\left(\mathbf{a}_i^{(l+1)}\right), \quad s = 1, 2 \tag{3}$$

where $g(\cdot)$ denotes a set of reconstruction "Block" with sigmoid activation function and without the BN layer in R-Net. More specific network configuration can be found in Table I.

D. Objective Function

As stated, the overall objective function in EndNet can be given as follows:

$$\min_{\phi,\varphi} \underbrace{\sum_{s=1}^{2} \|\mathbf{X}_s - g_\varphi(f_\phi(\mathbf{X}_s))\|_{\mathrm{F}}^2}_{\textbf{encoder}-\textbf{decoder}} + \underbrace{\frac{-1}{N}\sum_{i=1}^{N}\left[\mathbf{y}_i \log \mathbf{a}_i^{(l+1)} + (1-\mathbf{y}_i)\log\left(1-\mathbf{a}_i^{(l+1)}\right)\right]}_{\textbf{cross}-\textbf{entropy}} \tag{4}$$

Authorized licensed use limited to: Jocelyn Chanussot. Downloaded on August 31,2020 at 15:21:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE I

NETWORK CONFIGURATION IN EACH LAYER OF ENDNET. FC, d, AND C ARE THE ABBREVIATION OF FC LAYER, THE SIZE OF INPUT DATA, AND THE NUMBER OF CLASS. MOREOVER, THE VALUES IN THE LAST ROW ARE THE LAYERWISE SIZE

FE-Net				F-Net			R-Net			
Block1	Block2	Block3	Block4	Block5	Block6	Block7	Block8	Block9	Block10	Block11
FC	FC	FC	FC	FC	FC	FC	FC	FC	FC	FC
BN	BN	BN	BN	BN	BN	–	–	–	–	–
ReLU	ReLU	ReLU	ReLU	ReLU	ReLU	Softmax	Sigmoid	Sigmoid	Sigmoid	Sigmoid
16	32	64	128	128	64	C	64	32	16	d

TABLE II

QUANTITATIVE COMPARISON OF DIFFERENT NETWORKS ON THE HOUSTON2013 DATA SETS. THE BEST IS SHOWN IN BOLD

Method	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	OA	AA	κ
HSI	82.72	83.36	**100**	92.05	98.20	95.10	82.84	48.53	74.88	52.80	80.74	84.25	75.79	**100**	**98.73**	80.39	83.33	78.83
LiDAR	35.71	57.33	83.56	72.16	70.08	71.33	72.29	85.57	47.88	60.14	79.70	47.36	70.53	85.83	41.23	63.61	65.38	60.59
Early	80.44	80.73	**100**	**96.50**	98.67	83.22	82.74	82.24	75.92	71.04	84.72	79.54	82.46	**100**	98.52	84.88	86.45	83.66
Middle	80.34	84.02	**100**	92.90	99.53	95.10	82.46	82.05	86.21	75.87	85.48	81.65	**84.56**	**100**	98.52	86.62	88.58	85.56
Late	**82.81**	**84.21**	**100**	93.09	99.05	**98.60**	**88.90**	78.35	81.87	79.73	85.77	89.05	78.60	**100**	**98.73**	87.60	89.06	86.59
Li's	81.10	81.02	96.44	88.35	**100.00**	95.80	86.38	**90.88**	82.53	72.78	**88.99**	83.09	76.14	93.93	97.46	86.52	87.66	85.41
EndNet	81.58	83.65	**100**	93.09	99.91	95.10	82.65	81.29	**88.29**	**89.00**	83.78	**90.39**	82.46	**100**	98.10	**88.52**	**89.95**	**87.59**

where $\mathbf{y}_i$ is defined as the one-hot encoded label vector with C categories in the ith pixel, and the encoder–decoder term formulates how the reconstruction idea is embedded in the proposed end-to-end network. More details regarding the network architecture can be found in Fig. 1.

III. EXPERIMENTS

A. Data Description

Two popular and widely used HSI and LiDAR data sets are used to evaluate the performance of the proposed network both qualitatively and quantitatively.

1) Houston2013 Data: The first data set is available from the IEEE GRSS DFC2013.[1] One HSI with 144 bands in the spectral range from 364 to 1046 nm and one LiDAR image with 1 band by 349×1905 pixels are involved in this scene. Fig. 2 shows false-color HS and LiDAR images and the spatial distribution of training and testing samples in the studied scene, while the specific number of training and testing samples can be found in [16].

2) Trento Data: The second data set, which similarly consists of one HSI with 63 spectral bands and one LiDAR image, was acquired over the area from the city of Trento, Italy. More specifically, this scene consists of 166×600 pixels with a ground sampling distance (GSD) of 1m. Fig. 3 visualizes the HSI and LiDAR as well as the locations of training and testing samples, and the number of training and testing samples for each class can be detailed in [21].

B. Experimental Configuration

Our EndNet is implemented in the Tensorflow framework. The Adam [22] is adopted to optimize our networks, and the learning rate is determined by multiplying the base one (e.g., 0.001) with "exponential" policy, e.g., $(1 - (\text{iter}/\text{maxIter}))^{0.5}$, at intervals of 30 epochs. To alleviate the effects of network overfitting, the ℓ_2-norm regularization parameterized by 0.01 is employed on weights. The network training would stop as long as the loss change on the validation set approximately reaches stable. Please note that the size for each FC layer and hyperparameters in networks, e.g., learning rate and regularization parameter, can be well determined by tenfold cross-validation on the training set. More specifically, we select the layer size and hyperparameters in the range of {16, 32, 64, 128, 256} and {0.0001, 0.001, 0.01, 0.1, 1}, respectively.

[1]http://www.grss-ieee.org/community/technical-committees/data-fusion/2013-ieee-grss-data-fusion-contest/

TABLE III

QUANTITATIVE COMPARISON OF DIFFERENT NETWORKS ON THE TRENTO DATA SETS. THE BEST IS SHOWN IN BOLD

Method	C1	C2	C3	C4	C5	C6	OA	AA	κ
HSI	89.99	84.95	96.79	94.78	75.68	62.22	83.16	84.07	77.83
LiDAR	79.46	92.98	37.97	93.50	36.90	81.75	69.79	70.43	62.51
Early	90.06	97.84	94.92	98.36	83.39	87.58	86.64	90.87	82.62
Middle	90.06	97.84	94.92	98.36	83.39	87.58	90.79	92.03	87.82
Late	**90.24**	97.70	93.85	96.64	81.30	89.22	89.70	91.49	86.44
Li's	89.14	95.10	**98.66**	98.74	86.23	84.21	91.22	92.01	88.34
EndNet	88.19	**98.49**	95.19	**99.30**	**91.96**	**90.14**	**94.17**	**93.88**	**92.22**

Moreover, to better utilize the spatial information of LiDAR data and make our FC network feasible, the attribute profiles (APs) [23] are used to extend the LiDAR image to 21-band profiles as the network input.

Furthermore, the classification performance of the EndNet is quantitatively evaluated by three commonly used indices, i.e., overall accuracy (OA), average accuracy (AA), and kappa coefficient (κ), in comparison with several popular and widely used fusion networks, i.e., early fusion, middle fusion, late fusion, and the latest work (i.e., the work of Li *et al.* [16]) as well as single modalities. One main reason to select these compared models is to verify the effectiveness of the proposed encoder–decoder fusion baseline compared with concatenation-based ones. For a fair comparison, we design the same network architecture for all methods, e.g., the number of the layer and the neuron, and activation function. A general network configuration for all models is given in Table I.

C. Results and Analysis

In Tables II and III, we make a quantitative comparison between different baseline networks by employing the classification as a potential application in terms of three indices:

Authorized licensed use limited to: Jocelyn Chanussot. Downloaded on August 31,2020 at 15:21:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Fig. 2. Classification maps of different compared methods on the Houston2013 data sets.

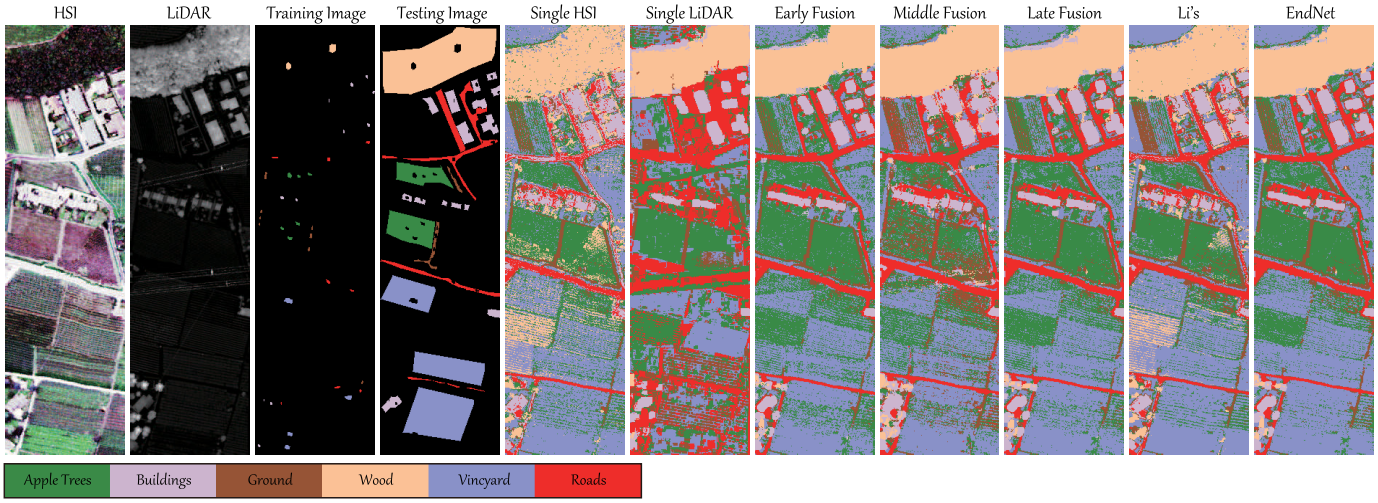

Fig. 3. Classification maps of different compared methods on the Trento data sets.

OA, AA, and κ as well as the classification accuracy for each category.

More specifically, although the APs are extracted from the LiDAR image, which is capable of further enhancing its spatial information, the network using HSI still performs better than LiDAR's due to the rich spectral information in HSI. Not unexpectedly, the classification performance of single modalities is inferior to that of the multimodal input (HSI and LiDAR in our case). Moreover, the multimodal data, particularly heterogeneous HSI and LiDAR data, usually fail to fuse in a very early stage, e.g., early fusion, due to a big gap between the original features from different modalities. This can well explain why late fusion and middle fusion that hierarchically extract feature representations before fusion can achieve higher classification accuracies. It should be noted that there is a similar performance trend between middle fusion and late fusion. One main reason could be high-level feature extraction, making it possible for the features from different modalities to be projected into a comparable feature space. Due to possibly insufficient information fusion, Li's method also achieves similar results compared with those aforementioned models, despite the use of more advanced network modules. More remarkably, the proposed EndNet outperforms other competitors and obtain the best classification results, which demonstrates its superiority and great potential in multimodal feature blending to a great extent.

D. Visual Comparison

Figs. 2 and 3 visualize the classification maps of different compared methods in the full scene. Referring to the false-color HSI and LiDAR image, we can see that the classification maps only using single modalities (HSI and LiDAR) are relatively worse than those jointly using the two modalities. In particular, the materials of Wood and Buildings in the second Trento data sets tend to be misclassified using single-modality networks, leading to the noisy classification maps.

Authorized licensed use limited to: Jocelyn Chanussot. Downloaded on August 31,2020 at 15:21:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Moreover, the latest Li's method does not well recognize the materials of Wood and Apple Trees, leading to some noisy pixels in the region of Wood and Apple Trees (see on the top and middle of classification map). This might be explained by the bottleneck in the fusion strategy. Intuitively, these concatenation-based networks behave similarly performance at least from the perspective of classification maps. In detail, the middle fusion, late fusion, and Li's work are able to deal with the details more robustly and effectively; for example, the material Commercial located in the region covered by the cloud is identified more accurately compared with that using early fusion in the Houston2013 data. As expected, our proposed EndNet achieves the best visual results, yielding smoother and more realistic classification maps on both data sets. There is a similar trend between visual results and quantitative ones listed in Tables II and III as well. Although these models, including compared and our networks, are all composed of FC layers rather than spatial-induced CNNs, yet the results obtained by the EndNet are still comparable. This shows the effectiveness of the EndNet in the classification of HSI and LiDAR data.

IV. Conclusion

In this letter, we aim to propose a novel baseline (EndNet) in FC-dominated multimodal networks with the application for the classification of HSI and LiDAR data. Similar to these concatenation-based models, the proposed baseline starts to extract the feature representations from different modalities and better fuses them by additionally enforcing a regularization behavior before feeding into the softmax layer connecting with labels. Inspired by the effectiveness of the proposed EndNet in the fusion-related task baseline, we would like to further extend this baseline to improve the performance of some existing networks and develop more advanced modules in the future work.

Acknowledgment

The authors would like to thank Prof. L. Bruzzone from the University of Trento for providing the Trento data and the Hyperspectral Image Analysis Group, University of Houston, for providing the Compact Airborne Spectrographic Imager (CASI) data sets.

References

[1] J. Kang, D. Hong, J. Liu, G. Baier, N. Yokoya, and B. Demir, "Learning convolutional sparse coding on complex domain for interferometric phase restoration," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Apr. 9, 2020, doi: 10.1109/TNNLS.2020.2979546.

[2] R. Huang, D. Hong, Y. Xu, W. Yao, and U. Stilla, "Multi-scale local context embedding for LiDAR point cloud classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 4, pp. 721–725, Apr. 2020.

[3] B. Rasti *et al.*, "Feature extraction for hyperspectral imagery: The evolution from shallow to deep (overview and toolbox)," *IEEE Geosci. Remote Sens. Mag.*, early access, Dec. 15, 2020, doi: 10.1109/MGRS.2020.2979764.

[4] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

[5] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 2, pp. 528–537, Feb. 2012.

[6] D. Hong, J. Chanussot, N. Yokoya, J. Kang, and X. X. Zhu, "Learning-shared cross-modality representation using multispectral-LiDAR and hyperspectral data," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1470–1474, Aug. 2020.

[7] D. Hong, N. Yokoya, G.-S. Xia, J. Chanussot, and X. X. Zhu, "X-ModalNet: A semi-supervised deep cross-modal network for classification of remote sensing data," *ISPRS J. Photogramm. Remote Sens.*, vol. 167, pp. 12–23, Sep. 2020.

[8] W. Liao, A. Pizurica, 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.

[9] 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.

[10] 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.

[11] 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.

[12] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, early access, Aug. 18, 2020, doi: 10.1109/TGRS.2020.3015157.

[13] X. Sun, L. Zhang, H. Yang, T. Wu, Y. Cen, and Y. Guo, "Enhancement of spectral resolution for remotely sensed multispectral image," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 5, pp. 2198–2211, May 2015.

[14] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "CoSpace: Common subspace learning from hyperspectral-multispectral correspondences," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4349–4359, Jul. 2019.

[15] D. Hong, N. Yokoya, N. Ge, J. Chanussot, and X. X. Zhu, "Learnable manifold alignment (LeMA): A semi-supervised cross-modality learning framework for land cover and land use classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 147, pp. 193–205, Jan. 2019.

[16] H.-C. Li, W.-Y. Wang, L. Pan, W. Li, Q. Du, and R. Tao, "Robust capsule network based on maximum correntropy criterion for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 738–751, 2020.

[17] J. Nalepa, M. Myller, and M. Kawulok, "Validating hyperspectral image segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1264–1268, Aug. 2019.

[18] L. Zou, X. Zhu, C. Wu, Y. Liu, and L. Qu, "Spectral–spatial exploration for hyperspectral image classification via the fusion of fully convolutional networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, no. 1, pp. 659–674, Oct. 2020.

[19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167

[20] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning," in *Proc. ICML*, pp. 689–696, 2011.

[21] 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.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[23] D. Hong, X. Wu, P. Ghamisi, J. Chanussot, N. Yokoya, and X. X. Zhu, "Invariant attribute profiles: A spatial-frequency joint feature extractor for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 3791–3808, Jun. 2020.

Authorized licensed use limited to: Jocelyn Chanussot. Downloaded on August 31,2020 at 15:21:56 UTC from IEEE Xplore. Restrictions apply.