An Enhanced 3-D Discrete Wavelet Transform for Hyperspectral Image Classification

Xiangyong Cao, *Member, IEEE*, Jing Yao, Xueyang Fu, *Member, IEEE*, Haixia Bi, and Danfeng Hong, *Member, IEEE*

***Abstract*— In the classification of hyperspectral image (HSI), there exists a common issue that the collected HSI data set is always contaminated by various noise (e.g., Gaussian, stripe, and deadline), degrading the classification results. To tackle this issue, we modify the 3-dimensional discrete wavelet transform (3DDWT) method by considering the noise effect on feature quality and propose an enhanced 3DDWT (E-3DDWT) approach to extract the feature and meanwhile alleviate the noise. Specifically, the proposed E-3DDWT method first applies classical 3DDWT method to the HSI data cube and thus can generate eight subcubes in each level. Then, the stripe noise is concentrated into several subcubes due to its spatial vertical property. Finally, we abandon these subcubes and obtain the feature cube by stacking the remaining ones. After acquiring the feature, we then adopt the convolutional neural network (CNN) model with an active learning strategy for classification since CNN has been verified to be a state-of-the-art feature extraction method for HSI classification, and active learning strategy can alleviate the insufficient labeled sample issue to some extent. In addition, we apply the Markov random field to enhance the final categorized results. Experiments on two synthetically striped data sets show that our proposed approach achieves better categorized results than other advanced methods.**

***Index Terms*— Classification, hyperspectral image (HSI), noise, wavelet transform.**

I. Introduction

HYPERSPECTRAL image (HSI) is a main data-type of remote sensing field, and can offer sufficient spatial–spectral information on some specific objective [1]–[3], which boosts the extensive research on HSI classification task in various applications [4]–[7]. For the past decade, researchers have proposed a ton of classification algorithms, which mainly revolve around extracting spatial–spectral features, such as patch-based methods [8]–[13], wavelet transform methods [14]–[16], and deep learning methods [17]–[20]. For a more comprehensive review, we can refer to [21].

Manuscript received February 14, 2020; revised March 27, 2020; accepted April 22, 2020. This work was supported in part by the China Postdoctoral Science Foundation under Project 2018M643655, in part by the Fundamental Research Funds for the Central Universities, and in part by the China NSFC Project under Contract 61906151 and Contract 61901433. *(Corresponding author: Xueyang Fu.)*

Xiangyong Cao and Jing Yao are with the School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China.

Xueyang Fu is with the School of Information Science and Technology, University of Science and Technology of China, Hefei 230052, China (e-mail: xyfu@ustc.edu.cn).

Haixia Bi is with the Faculty of Engineering, University of Bristol, Bristol BS8 1UB, U.K.

Danfeng Hong is with the German Aerospace Center (DLR), Remote Sensing Technology Institute (IMF), 82234 Wessling, Germany.

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.2020.2990407

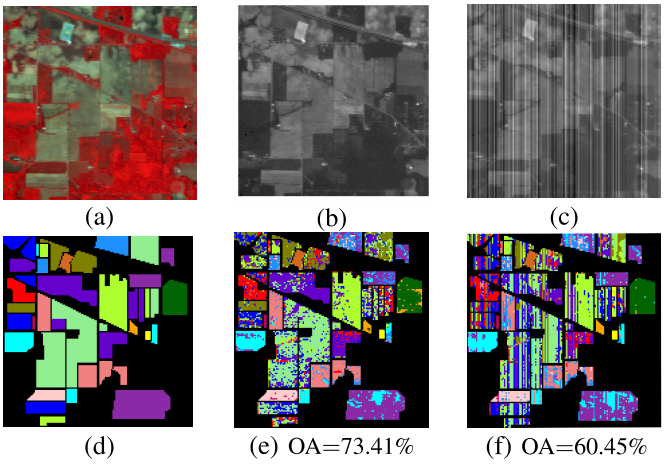

Fig. 1. Illustrated example of feature quality being affected by noise. (a) False-color image. (b) Clean image (band 10). (c) Noisy image (band 10). (d) Ground-truth map. (e) Classification map of (b) by 3DDWT. (f) Classification map of (c) by 3DDWT.

Although these methods achieve a good classification performance, they intend to be easily affected by the noise (e.g., Gaussian, stripe, and deadline noise). To illustrate this phenomenon, we provide an example in Fig. 1, from which we can observe a significant decline of classification overall accuracy (OA; from 73.41% to 60.45%) when the HSIs are contaminated by the stripe noise. This is principally because noise can destroy the image content and thus affect the feature quality. To alleviate this issue, a traditional way is to remove the noise first and then extract the feature. This letter proposes an alternative approach to simultaneously alleviate stripe noise and extract feature. This approach is built on the 3-dimensional discrete wavelet transform (3DDWT) [14] and thus is called enhanced 3DDWT (E-3DDWT). More specifically, the E-3DDWT method consists of the following steps. First, we adopt the traditional 3DDWT approach to decompose the HSI data cube, and thus obtain eight subcubes in each level. Then, we can observe that the stripe noise mainly concentrates in a few subcubes [22] (see Fig. 2) due to its spatial vertical orientation property. Finally, we discard these subcubes and get the final feature cube by stacking the remaining ones. In this way, we can mitigate the effect of noise on the feature quality to some extent.

However, the dimension of the extracted feature is very high (equals $(5L+1)d$, where L is level number and d is the original spectral dimension) while the number of annotated samples is often small, which poses a high dimension and small sample problem. To alleviate the issue, we resort to the convolutional neural network (CNN) classifier for two reasons.

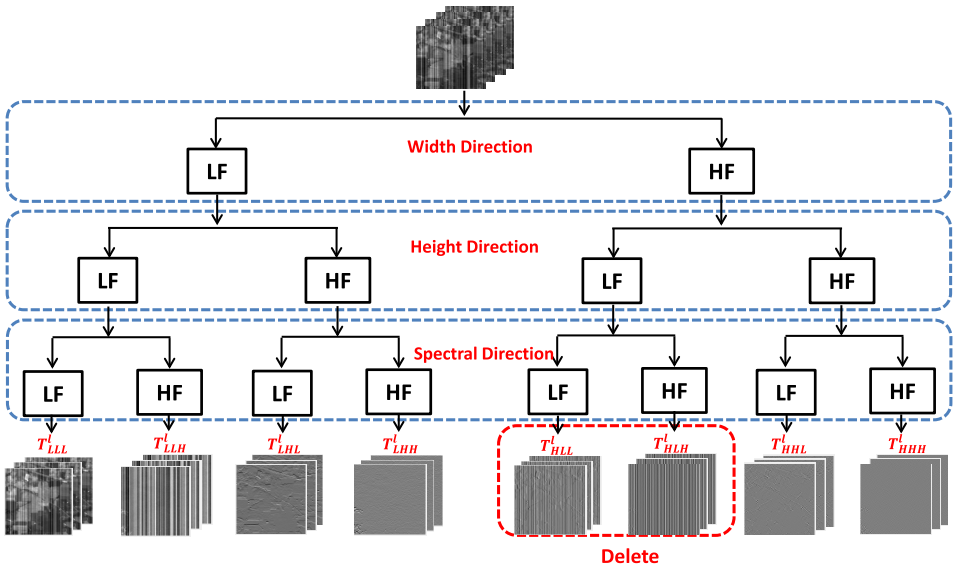

Fig. 2. L_{th} level of the proposed enhanced 3DDWT method.

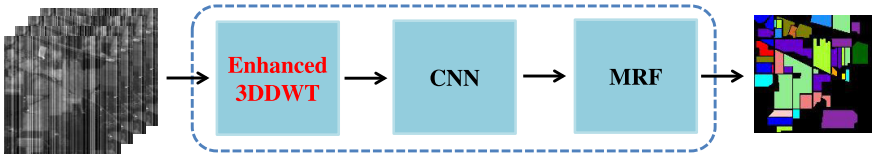

Fig. 3. Proposed framework for HSI classification.

First, to make the extracted features be more discriminative, we need to further exploit the spectral–spatial information of HSI in the classification stage while the CNN classifier is capable of extracting such knowledge. Second, the first convolutional layer of the CNN can reduce the dimension of the extracted E-3DDWT feature from $D = (5L + 1)d$ to 20 (e.g., the number of convolution filter is 20). In addition, to mitigate the insufficient labeled sample issue, we utilize the active learning technique in our CNN model. We also adopt the Markov random field (MRF), which assumes neighboring pixels to have the same label with a great probability, to help improve the final classification result.

On the whole, we propose an approach for HSI classification based on the proposed E-3DDWT feature extraction method. Specifically, this article has the following contributions.

1) We propose a novel E-3DDWT approach to mitigate stripe noise while extract the spatial–spectral feature simultaneously, which is the main novelty of this article.
2) We apply the CNN and active learning strategy to the feature cube to alleviate the high dimension problem and small samples problem, separately. Moreover, we utilize the smooth prior of the HSI labels and construct a MRF model to enhance the categorized results.
3) Experiments on two synthetically striped HSI data sets demonstrated that our approach is capable of getting better categorized results than the state-of-the-art methods.

II. Our Method

A. Notations

The HSI data are denoted as $\mathcal{H} \in \mathbb{R}^{h \times w \times d}$, where h is spatial height, w is spatial width, and d is spectral dimension. The feature cube extracted by E-3DDWT method is represented as $\mathcal{C} \in \mathbb{R}^{h \times w \times D}$, where $D = [(7L+1) - 2L]d = (5L+1)d$ and L is the level number. $\mathcal{L} = \{(c_i, y_i)\}_{i=1}^{n}$ is the labeled set and $\mathcal{U} = \{c_i\}_{i=n+1}^{N}$ denotes the unlabeled set, where n represents the labeled sample number, N (equals hw) is the total number, $c_i \in \mathbb{R}^D$ is the ith sample feature, and $y_i \in \{1, 2, \ldots, K\}$ is the corresponding label. For each c_i, the input to the CNN model is its local patch cube denoted as $x_i \in \mathbb{R}^{k \times k \times D}$, and thus the real training set for the CNN model is $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$. All the labels of the HSI are represented as $Y = \{y_i\}_{i=1}^{N}$.

B. Model

In Fig. 3, we present the framework of our method. Specifically, an enhanced 3DDWT (E-3DDWT) method is first proposed to alleviate the effect of stripe noise on the feature extraction. Then, the extracted feature cube is used for classification with a CNN model and active learning strategy. Finally, we conduct a postprocessing operation on the classification map based on the MRF model. In the following, each step of this method will be presented in detail.

1) E-3DDWT: 3DDWT is a powerful tool and has been used in various HSI applications, such as classification [14] and denoising [22]. However, the feature quality of HSI can be always contaminated by the noise, which will degrade the subsequent classification performance. A traditional way to mitigate the effect of stripe noise on the feature is to denoise first and then conduct feature extraction. Different from the traditional method, this article proposes another way where alleviating the noise and extracting the feature are done simultaneously. In this article, we propose an enhanced 3DDWT (E-3DDWT) feature extraction method shown in Fig. 2 based on this new way. More specifically, we adopt the dyadic decomposition due to its simplicity of implementation, whose one-level

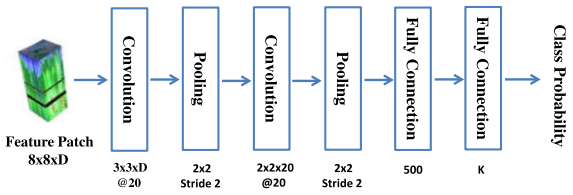

Fig. 4. Architecture of CNN [23]. The HSI patch size is $8 \times 8 \times D$. The size of convolutional kernel in the first convolutional layer is $3 \times 3 \times D$ and the number of filter is 20. The size of convolutional kernel in the second convolutional layer is $2 \times 2 \times 20$ and the number of filter is also 20. After convolution, a max-pooling layer with kernel size 2×2 and a stride of two pixels is adopted. Finally, two fully connected layers with 500 units and K units are used.

decomposition is illustrated in Fig. 2. From Fig. 2, we can see that a low-pass filter (LF) and a high-pass filter (HF) are used in the width, height, and spectral directions, respectively. In our work, we adopt the Haar wavelet. By conducting this decomposition, we can obtain eight subcubes (e.g., $\mathbf{T}^l_{LLL}$, $\mathbf{T}^l_{LLH}, \mathbf{T}^l_{LHL}, \mathbf{T}^l_{LHH}, \mathbf{T}^l_{HLL}, \mathbf{T}^l_{HLH}, \mathbf{T}^l_{HHL}, \mathbf{T}^l_{HHH}$). Since it has been verified in the work that stripe noise mainly appears on the subcubes with a vertical spatial orientation, namely, $\mathbf{T}^l_{HLL}$ and $\mathbf{T}^l_{HLH}$ (shown in the red dotted box of Fig. 2), we discard these two subcubes in the process of decomposition and thus can alleviate the stripe noise to some extent. In summary, this is the main difference between the E-3DDWT method and the 3DDWT method. Furthermore, to implement the multilevel decomposition, the approximated baseband $\mathbf{T}^l_{HHH}$ can be recursively decomposed. By stacking the remaining subcubes, we can get the final feature cube $\mathcal{C} \in \mathbb{R}^{h \times w \times (5L+1)d}$, where L is the level of decomposition.

2) CNN Model With Active Learning: After obtaining the feature cube $\mathcal{C}$ by the E-3DDWT method, we can observe that the feature dimension is very high while the labeled sample number is small, which causes a typical high-dimension and small samples problem. To alleviate the high dimension issue, we adopt the CNN model (shown in Fig. 4) since it can reduce the feature dimension (e.g., the number of convolution filter is set as 20) in the process of feature extraction while still attains a good spatial–spectral feature. Specifically, the loss function (namely, negative log-likelihood) of the CNN model is

$$L(\Theta|\mathcal{D}) = -\sum_{i=1}^{n}\sum_{k=1}^{K} 1\{y_i = k\} \log P(y_i = k|x_i, \Theta) \tag{1}$$

where $1\{\cdot\}$ is the indicator function defined as: $1\{true\} = 1$ and $1\{false\} = 0$, Θ denotes the parameter set of CNN, and $P(y_i = k|x_i, \Theta)$ represents the probability of x_i to take label k, which is defined as the output of the CNN.

Furthermore, to mitigate the small sample issue, we resort to the active learning strategy. This strategy can select the most uncertain samples, which help accelerate the CNN training and reduce the number of required label samples. For the choice of active selection criterion, we adopt the best-versus-second best (BvSB) [24], which describes the confusing extent of current model on one unlabeled sample in the candidate pool and can be computed as the difference between the biggest element and the second biggest element of the class probability vector [namely, $P(y_i = k|x_i, \Theta)$]. Therefore, the samples with higher BvSB values will be selected preferentially.

3) MRF Postprocessing: After completing the testing phase using the trained CNN, we further consider the smooth prior of the image label, which enforces neighboring pixels to share the same label with a high probability [14], [25]. Specifically, the objective function is

$$L(Y) = -\sum_{i=1}^{N}\sum_{k=1}^{K} 1\{y_i = k\} \log P(y_i = k|x_i, \Theta) + \lambda \sum_{i=1}^{N}\sum_{j \in \mathcal{N}_i} [1 - \delta(y_i, y_j)] \exp\left(-\frac{||c_i - c_j||_2^2}{2\sigma}\right) \tag{2}$$

where $\lambda \geq 0$ represents smooth parameter, $\mathcal{N}_i$ denotes the surrounding pixels of the ith pixel, $\delta(.)$ is the Kronecker function defined as: $\delta(a, b) = 1$ for $a = b$ and $\delta(a, b) = 0$ otherwise, and $\sigma > 0$ denotes scale parameter. The smooth prior of the labels is mainly reflected in the second term of (2), whose minimization can result in neighboring pixels taking the same label value. Although minimizing $L(Y)$ is an NP-hard problem, $L(Y)$ is converted into a MRF model [18], [25] and then solved by belief propagation algorithm [26].

III. Experiments

A. Data Sets

Two data sets[1] [12], [14] are used to assess our approach. The first one is collected using the AVIRIS sensor and is called Indian Pines, whose size is $145 \times 145 \times 200$. The second one is Pavia University, whose size is $610 \times 340 \times 103$. Before adding the stripe noise to the two data sets, each data set is first normalized into the range [0–1]. Then, we randomly add stripe noise with the range [−0.3 to 0.3] to 70 columns of each band for both data sets, and obtain the HSI data sets with stripe noise. A sample image is shown in Fig. 1(c).

B. Experimental Settings and Parameter Settings

Three experimental settings are shown as follows.

1) *Ablation Study:* This experiment aims to assess the performance of each module in our method, such as E-3DDWT feature, CNN classifier with active learning (CNN-AL), and MRF. Specifically, for the feature, we compare the proposed E-3DDWT feature with original data and 3DDWT feature. For the classifier, we compare the CNN-AL with support vector machine (SVM) and CNN. For the postprocessing, we compare the method with MRF and without MRF. For fair comparison, when one module is being verified, the other ones should be fixed. Specifically, the Indian Pines data are used for this experiment. The training set is constructed by randomly choosing 5% of the samples from each class.
2) *Comparison With the State-of-the-Art Methods:* We assess our approach on two synthetically striped HSI data sets in comparison with eight advanced approaches, such as low-rank (LR) decomposition method [8], 3-dimensional Gabor wavelet (3DGW) [15], orthogonal total variation component analysis (OTVCA) [27], joint progressive learning (JPlay) [28], local block multilayer sparse extreme learning machine (LBMS-ELM) [12], stacked auto-encoder (SAE) [17], 3-D-CNN [20], and CNN with MRF (CNN-MRF) [18]. In addition, the training sets of Indian Pines data set are constructed by

[1]The two data sets are available at: http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes.

TABLE I

MEAN OA (%) WITH DIFFERENT FEATURES

Feature	Original Data	3DDWT	E-3DDWT
OA (%)	87.16(1.08)	90.12(1.43)	**91.98(1.17)**

TABLE II

MEAN OA (%) WITH DIFFERENT CLASSIFIERS

Classifier	SVM	CNN	CNN-AL
OA (%)	74.11(1.81)	88.23(1.64)	**91.98(1.17)**

TABLE III

MEAN OA (%) WITH/WITHOUT MRF

Post-processing	Without MRF	MRF
OA (%)	85.51(1.62)	**91.98(1.17)**

randomly selecting 5% and 3% of the samples in each class, while the training sets of Pavia University are built by randomly choosing 1% and 0.75% of the samples from each class.

3) *Comparison With Traditional Pipeline:* This experiment aims to compare the performance of the two pipelines, namely, 1) first denoising and then extracting feature using 3DDWT method and 2) simultaneously extracting feature and denoising (using E-3DDWT method). Traditional methods belong to the first pipeline and ours belongs to the second one. This experiment is conducted on two data sets, and the training sets are built as follows: randomly selecting 5% and 1% samples from each class for Indian Pines and Pavia University, respectively. For the first pipeline, we use two state-of-the-art mixed noise removal methods, namely, non-independent identically distributed Mixture of Gaussian (NMoG) [29] and mixed Gaussian and sparse noise reduction (MGSNR) [22] as the destriping approaches.

The parameter settings of our method are shown in the following. The number of decomposition level L is fixed as 1 since the offline experiment implies that the classification performance decreases as the level number increases. In addition, we follow the settings of the article [23] to select the round number, the epochs in each round, and the number of added samples in each round for the active learning strategy. Also, as suggested by Cao *et al.* [23], the smooth parameter λ is set as 5. All the experiments are repeated five times, and mean results are reported.

C. Experimental Result Analysis

1) Ablation Study: The results of this experiment are demonstrated in Tables I–III. First, it can be observed from Table I that the 3DDWT feature is more discriminative than the original data, which is reasonable since the 3DDWT method is capable of fully utilizing the spatial–spectral information. In addition, we can also see that the E-3DDWT feature outperforms the 3DDWT feature. This is because E-3DDWT can mitigate the stripe noise, whereas it extracts the feature and thus obtains better features than 3DDWT. From Table II, we can first see that the CNN outperforms SVM. Also, it can be found that CNN-AL obtains the best classification performance since more uncertain and informative samples are selected for training, thus helping to train the CNN in a better way. Finally, the utilization of MRF can enhance the categorized results due to the consideration of the smooth prior of labels, which can be seen in Table III. In summary, this experiment verifies that each module of our method can achieve performance gain.

TABLE IV

CLASSIFICATION RESULTS (%) FOR INDIAN PINES

Method	5%			3%		
	AA	OA	Average Time (s)	AA	OA	Average Time (s)
LR	69.34(3.56)	88.61(2.05)	762.32	63.78(0.61)	83.23(4.14)	674.85
3DGW	67.35(2.30)	86.43(1.50)	240.24	62.15(3.21)	77.85(2.26)	202.36
OTVCA	72.88(2.60)	78.25(2.03)	59.96	65.79(3.11)	74.55(2.12)	53.64
JPlay	74.32(2.55)	80.09(1.67)	70.69	66.72(2.42)	74.23(1.98)	61.07
LBMS-ELM	79.83(1.79)	87.32(2.15)	138.26	**76.06(1.84)**	83.46(2.01)	122.14
SAE	62.87(3.05)	83.12(1.75)	532.64	59.43(1.81)	77.35(5.64)	503.25
3DCNN	79.18(2.74)	85.05(1.81)	2203.18	71.35(2.49)	84.26(1.92)	1474.63
CNN-MRF	75.31(1.25)	86.05(1.12)	2693.58	70.81(1.56)	82.45(1.66)	2103.28
Ours	**80.34(1.78)**	**91.98(1.17)**	1741.02	72.96(1.53)	**85.90(1.05)**	983.60

TABLE V

CLASSIFICATION RESULTS (%) FOR PAVIA UNIVERSITY

Method	1%			0.75%		
	AA	OA	Average Time (s)	AA	OA	Average Time (s)
LR	81.82(1.48)	90.31(0.66)	1015.32	73.46(3.12)	87.29(2.88)	846.74
3DGW	71.32(3.06)	90.22(1.54)	456.18	70.05(2.87)	88.84(0.32)	325.74
OTVCA	73.49(2.21)	81.52(1.44)	230.97	69.37(2.89)	80.47(1.99)	202.71
JPlay	76.97(1.91)	83.63(1.25)	151.83	70.21(3.03)	79.89(2.17)	125.27
LBMS-ELM	82.30(1.56)	89.68(1.03)	3658.00	75.01(1.90)	86.74(2.39)	3572.19
SAE	74.42(2.65)	87.73(1.09)	569.38	72.25(2.40)	85.62(0.95)	407.63
3DCNN	81.65(1.31)	88.27(0.96)	1974.35	74.35(1.79)	86.12(1.90)	1677.90
CNN-MRF	77.24(1.48)	85.60(1.37)	2314.70	72.55(2.10)	83.28(1.49)	1948.13
Ours	**82.37(2.13)**	**91.27(1.48)**	1365.79	**75.19(1.23)**	**89.80(1.76)**	1023.58

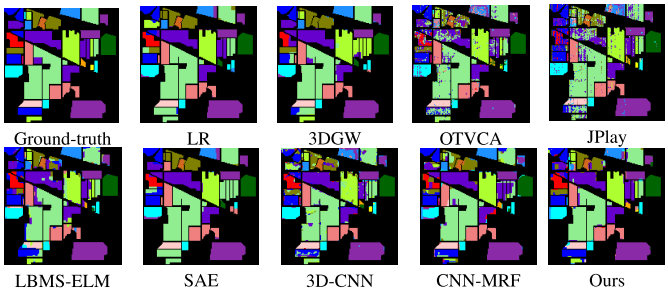

Fig. 5. Visual classification maps of Indian Pines data set.

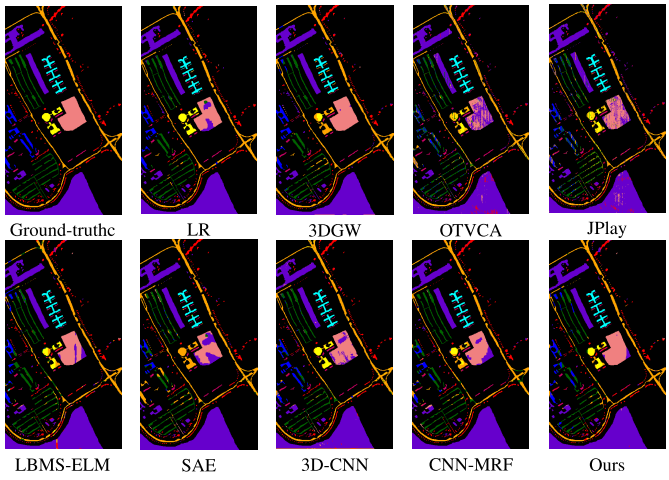

Fig. 6. Visual classification maps of Pavia University data set.

2) Comparison With the State-of-the-Art Methods: The average categorized results are demonstrated in Tables IV and V, and the final visual classification maps are illustrated in Figs. 5 and 6. From Tables IV and V, we can observe that our approach outperforms the competing

TABLE VI
MEAN OA (%) WITH DIFFERENT PIPELINES

Method	pipeline 1 (NMoG)	pipeline 1 (MGSNR)	pipeline 2 (Ours)
Indian Pines	91.89(1.36)	91.95(1.74)	**91.98(1.17)**
Pavia University	**91.42(1.85)**	91.38(1.93)	91.27(1.48)

approaches on both data sets. This is mainly because our method alleviates the noise in the feature extraction process and thus gets better features than the traditional methods (e.g., LR, 3DGW, OTVCA, JPlay, and LBMS-ELM). In addition, the factors, based on which the proposed method performs better than the three deep learning methods (e.g., SAE, 3-D-CNN, and CNN-MRF), are attributed to not only the noise compression while conducting feature extraction but also the utilization of active learning strategy, which can help select more uncertain and informative samples for the training. By comparing the visual classification results from Figs. 5 and 6, we can see that our method has a more precise classification map. In addition, the average running time is also reported in Tables IV and V.

3) Comparison With Traditional Pipeline: From the results recorded in Table VI, we can observe that our approach obtains comparable classification results compared with the traditional pipeline, which further verifies that our method provides a comparable pipeline with the current popular one. Therefore, this result implies that more research along this proposed pipeline can be conducted in the future.

IV. CONCLUSION

This letter proposes a novel enhanced 3DDWT approach to simultaneously extract the feature and alleviate the stripe noise. The extracted feature has higher quality than those methods which do not consider the noise effect. Moreover, we adopt a CNN model with an active learning strategy and MRF in the classification stage, which can alleviate the small samples issue and also improve some performance to some extent. Experimental results confirm that our approach has an advantage over the state-of-the-art methods in extracting features from a noisy HSI data set.

REFERENCES

[1] 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.

[2] J. Yao, D. Meng, Q. Zhao, W. Cao, and Z. Xu, "Nonconvex-sparsity and nonlocal-smoothness-based blind hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2991–3006, Jun. 2019.

[3] 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.*, early access, Jan. 16, 2020, doi: 10.1109/TGRS.2019.2957251.

[4] P. Ghamisi *et al.*, "New frontiers in spectral-spatial hyperspectral image classification: The latest advances based on mathematical morphology, Markov random fields, segmentation, sparse representation, and deep learning," *IEEE Geosci. Remote Sens. Mag.*, vol. 6, no. 3, pp. 10–43, Sep. 2018.

[5] 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.

[6] X. Wu, D. Hong, J. Tian, J. Chanussot, W. Li, and R. Tao, "ORSIm detector: A novel object detection framework in optical remote sensing imagery using spatial-frequency channel features," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 5146–5158, Jul. 2019.

[7] 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.

[8] Y. Xu, Z. Wu, and Z. Wei, "Spectral–spatial classification of hyperspectral image based on low-rank decomposition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2370–2380, Jun. 2015.

[9] X. Cao, Z. Xu, and D. Meng, "Spectral-spatial hyperspectral image classification via robust low-rank feature extraction and Markov random field," *Remote Sens.*, vol. 11, no. 13, p. 1565, 2019.

[10] F. Cao *et al.*, "Sparse representation-based augmented multinomial logistic extreme learning machine with weighted composite features for spectral–spatial classification of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6263–6279, Nov. 2018.

[11] Z. Wu, W. Zhu, J. Chanussot, Y. Xu, and S. Osher, "Hyperspectral anomaly detection via global and local joint modeling of background," *IEEE Trans. Signal Process.*, vol. 67, no. 14, pp. 3858–3869, Jul. 2019.

[12] F. Cao, Z. Yang, J. Ren, W. Chen, G. Han, and Y. Shen, "Local block multilayer sparse extreme learning machine for effective feature extraction and classification of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5580–5594, Aug. 2019.

[13] D. Hong, N. Yokoya, J. Chanussot, J. Xu, and X. X. Zhu, "Learning to propagate labels on graphs: An iterative multitask regression framework for semi-supervised hyperspectral dimensionality reduction," *ISPRS J. Photogramm. Remote Sens.*, vol. 158, pp. 35–49, Dec. 2019.

[14] X. Cao, L. Xu, D. Meng, Q. Zhao, and Z. Xu, "Integration of 3-dimensional discrete wavelet transform and Markov random field for hyperspectral image classification," *Neurocomputing*, vol. 226, pp. 90–100, Feb. 2017.

[15] L. Shen and S. Jia, "Three-dimensional Gabor wavelets for pixel-based hyperspectral imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 12, pp. 5039–5046, Dec. 2011.

[16] H. Bi, L. Xu, X. Cao, and Z. Xu, "Polsar image classification based on three-dimensional wavelet texture features and Markov random field," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 3921–3928.

[17] Z. Lin, Y. Chen, X. Zhao, and G. Wang, "Spectral-spatial classification of hyperspectral image using autoencoders," in *Proc. 9th Int. Conf. Inf., Commun. Signal Process.*, Dec. 2013, pp. 1–5.

[18] X. Cao, F. Zhou, L. Xu, D. Meng, Z. Xu, and J. Paisley, "Hyperspectral image classification with Markov random fields and a convolutional neural network," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2354–2367, May 2018.

[19] R. Hang, Q. Liu, D. Hong, and P. Ghamisi, "Cascaded recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5384–5394, Aug. 2019.

[20] A. B. Hamida, A. Benoit, P. Lambert, and C. Ben Amar, "3-D deep learning approach for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4420–4434, Aug. 2018.

[21] B. Rasti *et al.*, "Feature extraction for hyperspectral imagery: The evolution from shallow to deep," *IEEE Geosci. Remote Sens. Mag.*, to be published, doi: 10.1109/MGRS.2020.2979764.

[22] B. Rasti, P. Ghamisi, and J. A. Benediktsson, "Hyperspectral mixed Gaussian and sparse noise reduction," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 3, pp. 474–478, Mar. 2020.

[23] X. Cao, J. Yao, Z. Xu, and D. Meng, "Hyperspectral image classification with convolutional neural network and active learning," *IEEE Trans. Geosci. Remote Sens.*, early access, Feb. 13, 2020, doi: 10.1109/TGRS.2020.2964627.

[24] A. J. Joshi, F. Porikli, and N. Papanikolopoulos, "Multi-class active learning for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2372–2379.

[25] H. Bi, J. Sun, and Z. Xu, "A graph-based semisupervised deep learning model for PolSAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2116–2132, Apr. 2019.

[26] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2282–2312, Jul. 2005.

[27] B. Rasti, M. O. Ulfarsson, and J. R. Sveinsson, "Hyperspectral feature extraction using total variation component analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 6976–6985, Dec. 2016.

[28] D. Hong, N. Yokoya, J. Xu, and X. Zhu, "Joint & progressive learning from high-dimensional data for multi-label classification," in *Proc. ECCV*, Sep. 2018, pp. 469–484.

[29] Y. Chen, X. Cao, Q. Zhao, D. Meng, and Z. Xu, "Denoising hyperspectral image with non-iid noise structure," *IEEE Trans. Cybern.*, vol. 48, no. 3, pp. 1054–1066, Mar. 2018.