Hyperspectral Image Classification With Transfer Learning and Markov Random Fields

Xuefeng Jiang, Yue Zhang, Yi Li, Shuying Li, and Yanning Zhang

***Abstract*— This letter provides a brand new way of feature extraction, which can be applied in the supervised classification of hyperspectral image. The convolutional neural network (CNN) has been proven to be an effective method of image classification. However, due to its long training time, it requires a large amount of the labeled data to achieve the expected outcome. To decrease the training time and reduce the dependence on large labeled data set, we propose using the method of transfer learning by taking the advantage of Bayesian framework to integrate with spectrum and spatial information, making use of the Markov property of images to distinguish and separate the ones with class tags, and employing the CNN trained by band samples randomly selected from the data sets. The method of classification mentioned in our letter makes use of the real hyperspectral data sets to perform the experimental evaluation. The result demonstrates that our method is superior to the previous methods.**

***Index Terms*— Convolutional neural network (CNN), deep learning, image classification, Markov random fields (MRF), transfer learning (TL).**

I. Introduction

AFTER the decades of continuous improvement over the past years, the hyperspectral remote sensor has been able to distinguish the spectrographic features among diverse materials accurately. Thus, the hyperspectral analysis technique could be applied in the fields of military, agriculture, geological exploration, and environmental conservation, enhancing its significance to a great extent [1]–[3]. Nonetheless, hyperspectral imaging (HSI) possesses the characteristics of high-dimensional images and insufficient training samples, making it an extremely complicated and tough task to analyze and categorize. Even worse, it is highly likely that the process will bring about the Hughes phenomenon. Plenty of classical methods of HSI classification are based on band selection, including the improved ant colony algorithm, the improved firefly algorithm, and so on. Unfortunately, they all require mass data, along with a complex process which would cost too much time, and a stringent requirement of models' applicability. What's more, the overall accuracy may even decrease for the lack of robustness. However, the feature extraction is totally a different way of classification, which is usually combined with machine learning. For example, support vector machine, multinomial logistic regression (MLR), and so on. In recent years, with the continuous development of deep learning, HSI classification steps into an era of machine learning, arising many novel methods, such as extreme deep-learning machine [4], [5], deep-convolutional neural network (DCNN) [6], DC-CNN [7], and so on. Among the methods mentioned above, making use of the CNN would be a brilliant way to implement the classification. However, its training time is long, and the demand for the marked data is large. In this letter, we propose a Bayesian supervised classification model with combining Markov random fields (MRFs) smoothly and CNN, which reduces the training time of the CNN by transfer learning (TL). Its advantages are explained in the following two aspects: in the first, it uses TL to reduce the training time of the CNN, eliminating the drawbacks of inefficiency and insufficient samples. In the second place, we come up with an HSI classification model based on Bayesian framework, which is enabled to reduce the time. It makes further use of spatial information through the Markov clustering algorithm, along with CNN to extract spectral–spatial feature. The most outstanding advantage of MRF is that two adjacent pixels could be continuous, meaning the labels of two pixels could belong to the same class. The way of combination in this letter is rarely used among HSI classification.

II. Related Work

A. Convolutional Neural Network and Transfer Learning

1) Convolutional Neural Network: Deep learning incorporates classification and extraction of features into a single framework, the transforming raw data into more abstract form at a higher level through the nonlinear complex transformational model. It is a method that enables the system to learn feature automatically. The CNN is one type of deep learning [8].

In essence, the CNN realizes the mapping relationship between input and output. It applies supervised learning method, which enables us to learn features through the training data in an implicit way. The hierarchical architecture of CNNs is gradually proved to be efficient and successful way to learn visual representations [9]. Random forest classifier (RFC) [10] is recently proposed for hyperspectral image classification, use a multiple classifier system (MCS) based on the MLR [11],

Manuscript received March 21, 2019; revised June 1, 2019; accepted June 10, 2019. This work was supported by the National Natural Science Foundation of China under Grant 61701387. *(Corresponding author: Shuying Li.)*

X. Jiang, Y. Zhang, Y. Li, and Y. Zhang are with the School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China.

S. Li is with the School of Automation, Xi'an University of Posts and Telecommunications, Xi'an 710121, China (e-mail: angle_lisy@163.com).

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.2019.2923647

and CNN is utilized to automatically find spatial-related features at high levels [12]. Pan *et al.* [13] proposed a novel simplified deep-learning model, rolling guidance filter (RGF), and vertex component analysis network (R-VCANet).

For hyperspectral image classification, the sample patch should be a 3-D patch. The sample is first input into the first convolution layer followed by a max pooling operation. Then obtains a set of pooled feature maps of a specific value and passes the combination through the second convolution layer filters to again obtain feature maps and then input into the max pooling layer. Finally, the pooled feature maps input into the fully connected layer, the connection layer can be defined as

$$h^{l+1} = g(W^l h^l + b^l), \quad l = 2, 3, 4 \tag{1}$$

where h^l is the fully connected layer input, h^{l+1} is the input, W^l and b^l are the weight and offset, respectively; $g(\cdot)$ is the activation function. The fully connected layer is mainly used to extract more abstract features to improve classification accuracy.

2) Transfer Learning: TL refers to a learning pattern that a system applies the knowledge of one specific field to another [14], [15]. Several kinds of TL methods are applied to hyperspectral classification, such as deep mapping-based heterogenous TL model (DTLM) [16] and kernel-based feature selection method [17]. According to the literature survey, TL has three types: instance-based, parameter-based, and feature-based methods. We borrowed the idea of TL, generate (pseudo) labels in the uncalibrated position, reduced the number of samples needed for classification, and shorten the classification time.

B. Markov Random Fields

Clustering is an effective method to use neighborhood correlation and spatial features. It can divide the image in the spectral dimension into related regional objects according to certain criteria. Using these regional objects to represent spatial information, it can better to make up for deficiencies based solely on the spectral information classification. MRF can improve the segmentation performance. The MRF theory is based on spatial correlation. The Markov property of the image defines the conditional probability of the pixel only related to its neighboring pixels but has nothing to do with all other factors. Schweizer and Moura [18] proposed an adaptive hyperspectral image analysis method based on the MRF theory, which effectively realized the classification and recognition of the features of spectral–spatial information. The statistical properties of each pixel on a Markov property definition image are only related to its neighborhood. Being the epsilon neighborhood of S, if random fields satisfy the following conditions:

$$\begin{aligned} P &= \{X = x\} > 0 \quad \forall x \in A \\ P\{X_s &= x_s | X_r = x_r, r \neq s, \forall r \in \delta(s)\} \\ &= P\{X_s = x_s | X_r = x_r, \forall r \in \delta(s)\}. \end{aligned} \tag{2}$$

Then, X is called the MRFs with epsilon neighborhood δ. The equations written above are called local conditions of MRFs. We use MRFs based on the maximum posterior probability to combine spectrum and spatial information. The structure knowledge of the data could be preserved in the target domain through applying MRF in order to optimize the classification.

III. Proposed Approach

In this section, we introduce the CNN based on TL at first, and then we transform classification tasks into the labeled questions of MRF through equations.

A. Transfer Learning and Convolutional Neural Network

CNN consists of convolutional layer and fully connected layer and has been proven to be a powerful tool for image classification [19]. CNN is forced to use local connection mode between neurons of adjacent layers to make the most of spatial correlation. Comparing with other multilayer perceptrons, CNN has a better performance when dealing with image processing, including image classification [20], image denoising, image deblurring, and so on.

However, training an excellent CNN not only requires massive labeled data but also a long processing time, particularly, when it is needed to deal with complex model and mass data. In order to train hyperspectral image classifier as quickly as possible, we used TL. Next, let us give the details of how we manage to propose the algorithm.

The input hyperspectral image $I_{R\times N}$, R is the number of pixels which is $m \times n$, N is the number of bands. Input the labeled sample X_i and marked tag Y_i of $I_{R\times N}$, X_i is a feature vector of the ith sample, Y_i is the labeled sample corresponding to X_i, $Y_i \in \{1, 2, \ldots, k\}$, k is the number of categories contained in the hyperspectral image. Considering the speed problem, we chose to randomly select ten bands as D_t at random, denoting the target domain sample, and also randomly select ten bands data from other bands as D_s, denoting the source domain sample. By k-means clustering [21], to implement clustering toward D_s, we get the clustering result of source domain sample D_s, and mark it as D_o, the total spatial information. Combining labeled sample X_i and the spatial information of D_s, and transferring the samples into labeled samples under the constraint of total spatial information D_o, we get the new labeled sample X_j and the new marked tag Y_j. Now, we could describe total spatial information D_o as: $D_o = \{D_z\}, z = 1, 2, \ldots, t$, D_z is the spatial information of the zth clustering block, also is the area whose clustering tags are the same and connected. t being the number of clustering block in the clustering result. Next, we calculate the site of each sample in the labeled sample X_i. If $i \in D_z$, then we add source domain sample X_r and labeled sample X_i into the new labeled sample X_j simultaneously under the conditions that $r \in D_z$, $r \neq i$. We can reconstruct the data from the target domain by the neighboring data on the source domain. Next, we mark tag of X_r as $Y_r = Y_i$.

For increasing the tagged data in the target domain, we make Y_r and Y_i joined, and get the new marked tag Y_j.

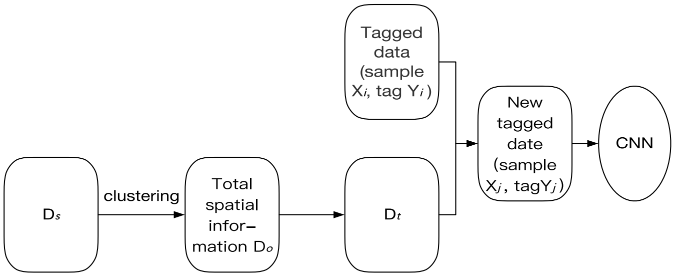
Tagged data (sample X_i, tag Y_i)
Dₛ
clustering
Total spatial infor– mation Dₒ
Dₜ
New tagged date (sample X_j, tagY_j)
CNN

Fig. 1. Flowchart of the proposed method.

The CNN is trained after learning through X_j and Y_j. Then, we use CNN to do the classification of D_T, thus achieve the label. Since we used a smaller number of hyperspectral image bands, high efficiency is achieved. Therefore, comparing with the previous way of using all the bands to realize analyzation, the process is greatly optimized. Fig. 1 illustrates the flowchart of the proposed method. Meanwhile, by combining with spatial information of the transferred samples, the accuracy of classification is improved as well.

In this experiment, the parameters are set as follows.

1) The network depth is seven: two pairs of convolution, max pooling layers, and three fully connected layers.
2) The convolution kernel sizes of these two layers are set to 5 and 3 seperately, the number of cores 100 and 300 are provided.
3) The kernel size of the max pooling layer is 2×2.
4) Three fully connected layers have 200, 100, and 16 nodes, respectively.
5) For the training patch, the patch size k is set to 9.

B. Priority of MRF

After achieving the pseudo-label through trained CNN, we begin to introduce our core framework. To be specific, we need to solve the following questions in order to gain the classification results:

$$\begin{aligned} Y &= \arg\max_{Y\in\kappa^n} \log P(Y|\widetilde{Y}) \\ &= \arg\max_{Y\in\kappa^n} \log P(\widetilde{Y}|Y) + \log P(Y) \\ &= \arg\max_{Y\in\kappa^n} \sum_{i=1}^{n}\sum_{K=1}^{K} 1\{Y_i = k\} \log \widetilde{y}_i k \\ &\quad + \log \frac{1}{Z} e^{\mu \sum_{i=1}^{n} \sum_{j\in N(i)} \delta(y_i - y_j)}. \end{aligned} \tag{3}$$

In the process of image segmentation, the adjacent pixels probably have the same label. Usually, taking the advantage of this kind of prior information could improve the performance. In this letter, we applied the smoothness prior approach on the tags. Z mentioned in the above equation is a constant of normalized distribution, k is the parameter of label smoothness, N refers to adjacent pixels of pixel i, μ is the label smoothness parameter. As we can see, when the adjacent tags are equal, the possibility to gain paired interaction terms is higher than the possibility of not to gain paired interaction terms. According to this, smoothness prior approach is beneficial to segmentation. The following is the ultimate classification model:

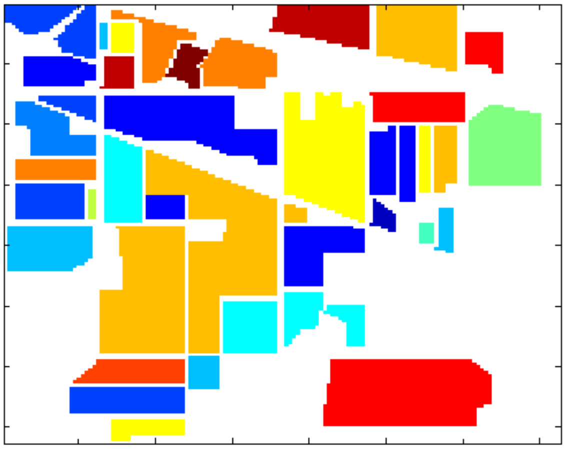

Fig. 2. Segmentation model of Indian pines data.

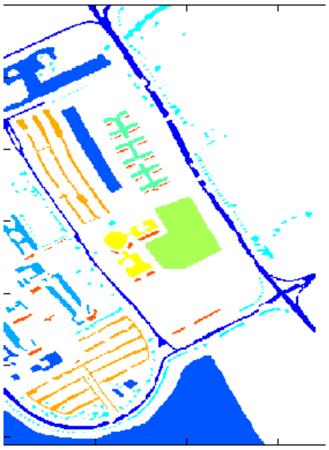

Fig. 3. Pavia University image.

$$\begin{aligned} Y = \arg\max_{y\in k^n} \Bigg\{ &\sum_{i=1}^{n}\sum_{K=1}^{K} 1\{Y_i = k\} \log Y_\mu \\ &+ \mu \sum_{i=1}^{n} \sum_{j\in N(i)} \delta(Y_i - Y_j) \Bigg\}. \end{aligned} \tag{4}$$

The object function contains multiple paired interaction terms, making it a challenging combinatorial optimization problem. We use random variables everywhere in the process, leading to the label's forming into MRFs. Thus, it could also be seen as a type of MRF model with the object function as its energy function, and the first term being the type of pixel's cost [22]. The more likely a pixel belongs to one certain type, the possibility of allocating corresponding tags to the pixel becomes greater.

IV. Experimental Results

We evaluated the effectiveness of the TL-MRF using two benchmark data sets made up of hyperspectral images. The effect of TL to the target detection performance is investigated. The latest compared methods we used are CNN [9], RF [10], MLR [11], and TL + MRF. Some experiments were performed

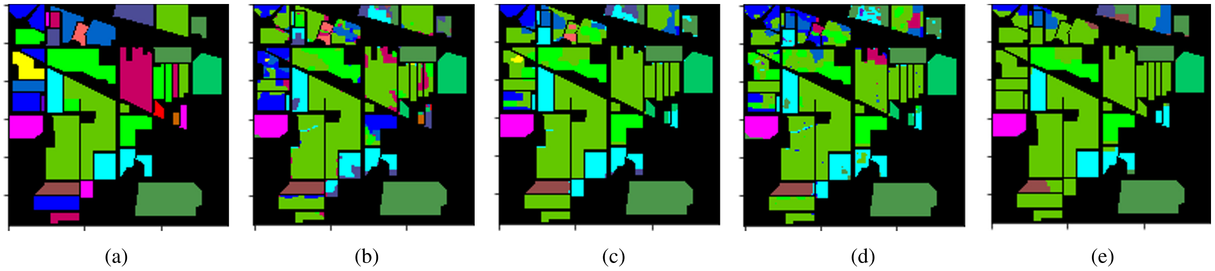

Fig. 4. Indian pines. (a) Ground-truth map. (b) MRF. (c) RF. (d) CNN. (e) TL-MRF.

implemented in the Python language and the Tensorflow library, and the others are run in MATLAB R2014b. The remainder of this section is organized as follows.

1) The data sets are displayed in Section IV-A.
2) In Section IV-B, the state-of-the-art comparison methods, evaluation indexes, and some details on the implementation of the proposed method are explicitly described.
3) Finally, the benefits of the proposed method are demonstrated in Section IV-C.

A. Data Sets

Introduced below are the two real data sets.

The first data set is the Indian Pines, which is widely used in hyperspectral classification. This scene was gathered by AVIRIS sensor over the Indian Pines test site in North-western Indiana and consists of 145 × 145 pixels and 224 spectral reflectance bands. The second remote sensing data is Pavia University, the number of spectral bands is 103 for Pavia University. Pavia University is 610 × 610 pixels, the geometric resolution is 1.3 m.

B. Experimental Method and Metrics

About AVIRIS Indian Pines Data, We randomly selected ten labeled band samples from 220 bands and clustered them to get spatial information. Then, we transferred the samples to target samples by calculating the locations of the labeled samples. Since then, we implemented the classification of few bands which had been selected at random as well and used the remaining samples in each class for testing. As for the methods mentioned above, we made charts for each of them. We got the segmentation model of hyperspectral images, of which the colors represent different aggregative centers.

We perform similar experiments on a second real HSI data set. Here, we show a sample band and the corresponding ground truth class map. Fig. 2 shows the color graphic of the image, while Fig. 3 shows the real label.

The classification performance metrics used in this letter are as follows.

1) *Overall Classification Accuracy (OA):* Refers to the ratio of the number of category pixels that are correctly classified to the total number of categories.
2) *Average Accuracy (AA):* The ratio between each type of prediction correctly and the total number of each type of class, and finally, the average of the accuracy of each class.

TABLE I
INDIAN PINES

class	CNN	RF	MLR	TL+MRF
OA	92.71	95.67	67.57	93.89
AA	90.04	93.64	51.50	89.77
k	91.32	94.80	62.16	92.93
Time(s)	329.346	492.509	50.383	330.449

TABLE II
PAVIA UNIVERSITY

class	CNN	RF	MLR	TL+MRF
OA	92.53	81.33	89.63	91.79
AA	89.44	80.62	81.44	88.67
k	87.26	90.54	85.34	91.64
Time(s)	262.870	706.395	57.085	294.581

3) *Kappa Coefficient (Kappa):* The Kappa coefficient is a ratio that represents the proportion of the errors that are reduced by classification and completely random classification.

C. Results and Discussion

The most vital influence of TL lies in the fusion time. For data set of Pavia Centre and University (Fig. 5), when only a little training data exists, the training process requires a numerous amount of time to optimize. In addition, the time could decrease quickly when the training data increases. Fig. 4(d) shows the classification result of the former training samples, while Fig. 4(e) shows the result after transferring the samples. By making a comparison, we found that our algorithm has a better classification effect when the number of training samples is limited.

From Tables I and II, we can see that our method obtains the best classification results on all data sets compared with all competing methods, which demonstrates the effectiveness of the proposed method. Based on the experimental results, some observations are achieved as follows.

1) Comparing with CNN, OA of TL-MRF has a better performance which indicates that the superiority of TL-MRF lies in the speed and the accuracy. To sum up, our algorithm has a satisfactory performance on Indian pine.

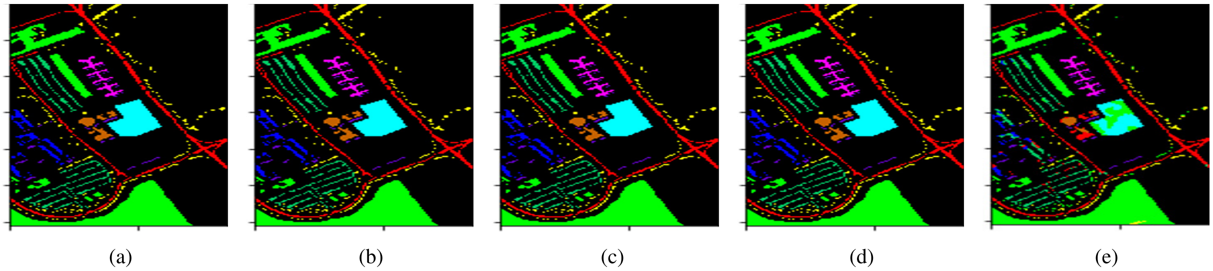

Fig. 5. Pavia University. (a) Ground-truth map. (b) MRF. (c) RF. (d) CNN. (e) TL-MRF.

2) Compared with other methods, the proposed method improves the overall accuracy dramatically and the time. This method is more advantageous in time than RF and in date k than CNN and MRF for the Indiana Pine data set and Pavia University. Considering the difference between the proposed TL-MRF and other methods, we attribute the improvement mainly from the combination of transformational learning and MRF.

The experiment proves the effectiveness of our algorithm, which takes the advantage of the similarity of hyperspectral bands, gets spatial information through few bands, samples into new labeled sample space with constrained spatial information and TL, and reaches a positive effect through training labeled samples of few bands.

V. Conclusion

In this letter, we propose a novel method of hyperspectral image classification, of which the main ideas are to combine TL with MRFs, to take the advantage of similarities among the bands, and to transfer CNN to hyperspectral training, thus achieving a better result with less training time. The method mainly contains two parts. On the one hand, we make use of TL, randomly select bands as the source domain and the target domain, and achieve the initial HSI classification result after training. On the other hand, we optimize the initial result through MRF. According to the classification result of the data sets known to all, our method has its superiority in accuracy.

TL is trending, and our method still has much to improve in further study, for a higher level of hyperspectral image classification network could achieve better performance. In addition, the calculation amount of the method could go lower with more efforts.

References

[1] Q. Wang, M. Chen, F. Nie, and X. Li, "Detecting coherent groups in crowd scenes by multiview clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published. doi: 10.1109/TPAMI.2018.2875002.

[2] Q. Wang, X. He, and X. Li, "Locality and structure regularized low rank representation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 911–923, Feb. 2019.

[3] X. Huang and L. Zhang, "An SVM ensemble approach combining spectral, structural, and semantic features for the classification of high-resolution remotely sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 257–272, Jan. 2012.

[4] F. A. Mianji and Y. Zhang, "Robust hyperspectral classification using relevance vector machine," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 6, pp. 2100–2112, Jun. 2011.

[5] J. Yue, S. Mao, and M. Li, "A deep learning framework for hyperspectral image classification using spatial pyramid pooling," *Remote Sens. Lett.*, vol. 7, no. 9, pp. 875–884, 2016.

[6] H. Zhang, Y. Li, Y. Zhang, and Q. Shen, "Spectral-spatial classification of hyperspectral imagery using a dual-channel convolutional neural network," *Remote Sens. Lett.*, vol. 8, no. 5, pp. 438–447, 2017.

[7] S. Derrode, C. Carincotte, and S. Bourennane, "Unsupervised image segmentation based on high-order hidden Markov chains [radar imaging examples]," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2004, p. V-769.

[8] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 3320–3328.

[9] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sensors*, vol. 2015, Jan. 2015, Art. no. 258619.

[10] S. Amini, S. Homayouni, and A. Safari, "Semi-supervised classification of hyperspectral image using random forest algorithm," in *Proc. IEEE Geosci. Remote Sens. Symp.*, Jul. 2014, pp. 2866–2869.

[11] L. Yi, J. Li, A. Plaza, J. Bioucas-Dias, A. Cuartero, and P. G. Rodríguez, "Spectral partitioning for hyperspectral remote sensing image classification," in *Proc. IEEE Geosci. Remote Sens. Symp.*, Jul. 2014, pp. 3434–3437.

[12] W. Zhao and S. Du, "Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4544–4554, Aug. 2016.

[13] B. Pan, Z. Shi, and X. Xu, "R-VCANet: A new deep-learning-based hyperspectral image classification method," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 5, pp. 1975–1986, May 2017.

[14] S. Si, D. Tao, and B. Geng, "Bregman divergence-based regularization for transfer subspace learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 7, pp. 929–942, Jul. 2010.

[15] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Spectral–spatial hyperspectral image segmentation using subspace multinomial logistic regression and Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 3, pp. 809–823, Mar. 2012.

[16] J. Lin, R. Ward, and Z. J. Wang, "Deep transfer learning for hyperspectral image classification," in *Proc. IEEE 20th Int. Workshop Multimedia Signal Process. (MMSP)*, Aug. 2018, pp. 1–5.

[17] C. Persello and L. Bruzzone, "Kernel-based domain-invariant feature selection in hyperspectral images for transfer learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 5, pp. 2615–2626, May 2016.

[18] S. M. Schweizer and J. M. F. Moura, "Efficient detection in hyperspectral imagery," *IEEE Trans. Image Process*, vol. 10, no. 4, pp. 584–597, Apr. 2001.

[19] Q. Wang, F. Zhang, and X. Li, "Optimal clustering framework for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 5910–5922, Oct. 2018.

[20] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning (Adaptive Computation and Machine Learning series)*. Cambridge, MA, USA: MIT Press, 2016.

[21] S. Ranjan, D. R. Nayak, K. S. Kumar, R. Dash, and B. Majhi, "Hyperspectral image classification: A k-means clustering based approach," in *Proc. 4th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, Jan. 2017, pp. 1–7.

[22] 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," 2017, *arXiv:1705.00727*. [Online]. Available: https://arxiv.org/abs/1705.00727