Attention-Based Domain Adaptation Using Residual Network for Hyperspectral Image Classification

Robiulhossain Mdrafi, *Graduate Student Member, IEEE*, Qian Du, *Fellow, IEEE*, Ali Cafer Gurbuz, *Senior Member, IEEE*, Bo Tang, *Member, IEEE*, Li Ma, *Member, IEEE*, and Nicolas H. Younan, *Life Senior Member, IEEE*

***Abstract*—In remote sensing images, domain adaptation (DA) deals with the regions where labeling information is unknown. Typically, hand-driven features for learning a common distribution among known and unknown regions have been extensively exploited to perform the classification task in hyperspectral images with the aid of state-of-the-art machine learning algorithms. Under limited training samples and using hand-crafted features, the classification performance degrades significantly. To overcome the engineered feature extraction process, an automatic feature extraction scheme can be seen useful to generate more complex but useful features for classification. Deep-learning-based architectures have been found to be pivotal on this regard. Deep learning algorithms are effectively used in hyperspectral domain to solve the DA problem. However, attention-based activation mappings, which are very successful for distinguishing different classes of images via transferring relevant mappings from a deep-to-shallow network is not widely explored in DA domain. In this article, we have opted to use attention-based DA through transferring different levels of attentions by means of different types of activation mappings from a deep residual teacher network to a shallow residual student network. Our goal is to provide useful but more complex features to the shallow student network for improving the overall classification in case of DA task. It has been shown that for different kinds of activation mappings, the proposed attention-based transfer improves the performance of the shallow network for the DA problem. It also outperforms the state-of-the-art DA methods based on traditional machine learning and deep learning paradigms.**

***Index Terms*—Activation mapping, attention mappings, hyperspectral image (HSI), knowledge distillation (KD), residual network, transfer learning.**

I. Introduction

DOMAIN adaptation (DA) in hyperspectral images (HSIs) helps to deduce the labels for the data where labeling information is unavailable. This can overcome the limitation of gaining information about the regions where direct human access is impossible. Due to this property of DA, one can see the classification problem in HSI domain as semisupervised, supervised, or unsupervised. By using the knowledge of the known (source) regions, DA is used to predict the unknown regions (target) whose distributions may or may not be known in advance thereby giving us much needed labeling information about the regions of interest. If the deduced labels are accurate, target region classification accuracy can improve.

The classification problem of DA can be defined in terms of the information about the target regions. If a small number of labels in the target region are available, DA can be seen as a supervised problem where the distribution about the unknown pixels in the target regions can be learned from the source region. While in the semisupervised case, the joint distribution of the unknown pixel in the target regions can be learned from the known labels of pixels of both source and target regions. Conversely, in case of unsupervised problem, no prior knowledge about the labels in the target region is given. Although classification problems in DA could be different, the main idea still remains the same where pixel distributions of the source region are needed to be matched to that of the target region for correct estimation of labels in the target region. By doing this genre of matching, the knowledge from one region is transferred to another to find the desired hidden features in a given image. The more the distribution can be matched, the more improvement on the classification accuracy can be obtained. This suggests a direct application for classification challenges in HSI. Specifically in HSI domain, challenges like atmospheric turbulence, distortion in the image acquisition process create blurry representation of pixels in the scene. These blurry representations have a detrimental effect on generalizing the pixel distributions in different regions. These representations produce different spectral signatures even for the same objects. Moreover, the high-dimensional nature of HSI makes the blur effects even more noticeable in [1] and [2]. Hence, the main theme is to overcome the effects of limitations in the image acquisition process. After overcoming these limitations, a more generalized distributions of data in both the source and target regions in the HSI image will be found. If the distributions can be generalized accurately, then the labels in the target regions can be estimated efficiently.

Different machine learning (ML) algorithms have been extensively used in the literature to achieve the generalized pixel distributions in both the source and target regions [3]–[6]. By adopting such state-of-the-art methods, the task of transferring

Manuscript received March 31, 2020; revised August 27, 2020 and October 25, 2020; accepted October 26, 2020. Date of publication November 3, 2020; date of current version November 16, 2020. This work was supported by the National Science Foundation CPS Award 1931861. *(Corresponding author: Robiulhossain Mdrafi.)*

Robiulhossain Mdrafi, Qian Du, Ali Cafer Gurbuz, Bo Tang, and Nicolas H. Younan are with the Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS 39762 USA (e-mail: rm2232@msstate.edu; du@ece.msstate.edu; gurbuz@ece.msstate.edu; tang@ece.msstate.edu; younan@ece.msstate.edu).

Li Ma is with the School of Mechanical Engineering and Electronic Information, China University of Geosciences, Wuhan 430074, China (e-mail: maryparisster@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2020.3035382

knowledge from one domain to the other has been simplified [7]. However, the hand-driven nature of the feature extraction process has become a challenge in adopting ML algorithms in DA cases. To cope with this problem, deep learning (DL) architectures are widely employed to generate automatic feature driven models from the given data itself. The design of different DL architectures also paves the way to learn different hierarchical representation of the HSI image to produce unique abstraction of the regions in HSI. To show the effectiveness of DL, the work in [8] has employed a deep neural network (DNN) on HSI to extract more informative features for the classification problem and provided improved performance measure comparing to the traditional ML-based classification methods. The trend of using different architectures of DNNs can be seen in various HSI classification tasks [2], [9]–[11]. All of these architectures show increased classification performance compared to the various state-of-the-art techniques for DA. Nonetheless, the requirement of producing useful and relevant features for DA still remains as a fundamental challenge. Recently, attention-based feature extraction using DL models opens the way to focus on specific parts of underlying objects in order to improve the classification accuracy [12]. The main idea in this study is to transfer the features from a superior deeper supervisor network with greater accuracy to a less complicated shallow student network to improve the accuracy of the shallow network based on the attention losses incurred by activation functions in the DNNs. Inspired by this, we have proposed to use attention-based domain adaption for HSI classification [13]. This prior publication reported higher classification accuracy for DA than the compared state-of-the-art methods. However, features extracted in [13] have been only done for spectral distribution of the given HSI, which can be an obstacle for improving the classification of the shallow network. In addition, only one particular activation mapping was employed to transfer the attention between the teacher (TN) and the student (SN) networks. To overcome these limitations, in this study, we have opted to evaluate different attention transfer mechanisms including taking advantage of both spatial and spectral features. It has been found that with the combination of change of attention mappings and different spatial–spectral features, we can achieve better performance measures than the earlier version of our work. In addition, for a fixed size of TN, different SN sizes are analyzed in order to determine the best attention-based DA performance for HSI.

The main contributions of this work can be stated as follows:

1) to introduce spatial–spectral based attention mappings in HSI domain;
2) to introduce and compare different variants of attention losses in case of HSI image classification;
3) to investigate the DA performance of shallow but wider SN after using attention mappings in the given DL architecture; and
4) to compare the computational complexity of different settings for TN and SN.

The rest of this article is organized as follows. Section II gives a brief introduction of background work. Section III details the attention-based neural network architecture for DA. Section IV presents the experimental settings and results. Conclusions and future work are summarized in Section V.

II. Background and Related Work

The DA process on remote sensing images is widely exploited for regression, classification, and clustering problems. The studies in the literature about DA for remote sensing images can be mainly divided into the following categories.

A. Instance-Based DA

In this genre for DA, the samples of the source patches are iteratively reweighted for use in the target domain. For satisfactory performance, both source and target samples must share the same dimensionality and be closely related to each other [14], [15]. The reweighted parameters try to reduce the difference in marginal/conditional distributions between the source and target domain. This weighting procedure can negate the misleading instances that are not relevant in the target domain [14]. While the work in [14] deals only with a supervised example, the method in [15] studies the semisupervised classification problem. In [15], few informative labeled examples from the distribution to be labeled have been selected via actively determining the set of most informative pixels to be labeled from a set of candidates in the target domain with the help of instance weights and SVM as a classifier. Another instance of this type of DA method iteratively reweights the source data for learning a common space between the heterogeneous source and target domain [16]. From the obtained common subspace, according to the relative importance, the source data are reused and reweighted by the iterative reweighting strategy and can also be used for transferring. The major advantage of this method is that reuse of source data enables the improvement of supervised classification in case for limited labeled samples in both source and target data. The method does not support semisupervised classification. The kernel instance DA method in [17] uses multiple kernels' weights in an adaptive way to train from the available labeled samples in the source domain and adds minimum number of most informative and active samples to label the pixels in the target domain.

B. Feature-Representation-Based DA

The feature-representation-based DA intends to extract features to represent the source and target domains such that the invariant features from both the source and target pixels are selected to improve the land-cover classification [18]. Recently, kernel-based nonlinear features are extracted to align the source domain (i.e., pixels with known labels) with the target domain (i.e., pixels with unknown labels) for semisupervised remote sensing image classification [19]. In addition, DL-based architectures like denoising autoencoder (DAE) and domain adversarial neural network (DANN) have also been used in [20] to learn the invariant feature representation for DA. By proposing DL architecture, the work in [20] learns invariant features in end-to-end manner across the domains to further boosting the classification performance of HSI by focusing on the individual

snapshots of the labels in the HSI. Another variant of DL-based autoencoder, namely segmented stacked autoencoder (SAE) is proposed in [21] where authors apply individual stack autoencoders in different regions of the given HSI. It reports improved classification while producing hierarchical features in a less complex manner. However, in case of complex distributions, the method may fail to locate the important spatial features for HSI classification.

C. Parameter-Transfer-Based DA

In this DA method, the information from the target image is transferred to adjust the parameters of the classifier to improve the accuracy of land-cover classification [22]. The associated parameters are mainly learned from the source domain. The works in [23] and [24] use this approach for updating feature parameters for a maximum-likelihood classifier in a multiple cascade classifier system by retraining the source domain to improve the land-cover classification. Another variant of this type of DA method is seen in [25] where different acquisition conditions of the dataset have been considered to design the DA algorithm. It uses the parameters from the source domain to the target one by using an estimate of nonlinear deformation, which has been done by graph matching and vector quantization. The transferred maps are obtained in unsupervised manner. The obtained maps help to project the parameters from the source to the target domain to solve the classification problem in HSI. Liu *et al.* [26] use homologous component analysis (HCA) for DA where the projected data information from the source and target domain is used to align the distributions.

D. Relational Knowledge-Based DA

This DA paradigm exploits the feature formalism in relational source and target domains. These relational features help to revise the initial mapped structure of target image to yield higher classification accuracy [27]. This work can be extended to multisensor domain as depicted in [28]. The method in [28] tries to exploit the feature subspace from the multisensor data to detect the same properties in both the source and target domains. In addition, the method in [29] uses two-level cluster mapping [derived from self-organizing maps (SOM)] to match the feature maps for both the source and target domains. After finding the closest feature maps, autoencoder is used for transforming the matched source–target pairs. This method can include missing points to improve the land-cover classification in both source and target domains.

However, all the aforementioned DA methods are successful, if the data from both source and target images resemble the same characteristics. In case of supervised classification, these methods need many samples for defining source subregions. In addition to these methods, DA based on DL architectures has become very popular. In [2], different strategies of DA in case of multisource data for land-cover classification task have been presented. The results show that for a limited amount of training data, convolutional and shallow neural networks tend to outperform other competing methods. In [30], a spectral–spatial unified network (SSUN) has been proposed for DA. Here, spectral and spatial feature extraction and classifier training are incorporated to formulate a uniform cost function for optimization simultaneously. The authors use long short-term memory (LSTM) model, and a multiscale convolutional neural network to extract spectral and spatial features, respectively, showing an improvement of overall accuracy (OA) in the case of multiple HSI datasets. For attention, transfer-based approaches mainly inspired from the unsupervised image saliency detection with Gestalt-laws guided optimization and attention in [31], a 3-D convolutional neural network (3-DCNN)-based residual channel and space attention network (RGSCA) is used for HSI classification [32]. It uses residual connection in both bottom-up and top-down manner to optimize the attentions of the channel and spatialwise features in the training process. It has the advantage to boost up the spatial features even if we have a limited amount of training samples. The works in [33] and [34] add a new dimension of feature extraction by including fusions of spectral and spatial features in HSI domain. Mu *et al.* [33] use a multiscale and multilevel spectral–spatial network to fuse the features effectively in different domains. The method uses a combination of 3-D and 2-D CNN to achieve the fusion of spectral and spatial features. In [34], local spectral features are obtained by applying 1-D CNN to each band in the HSI, where multiscale spatial features are obtained via hierarchical spatial pyramid pooling. Then, these features are concatenated to get the spectral–spatial fused features for HSI classification.

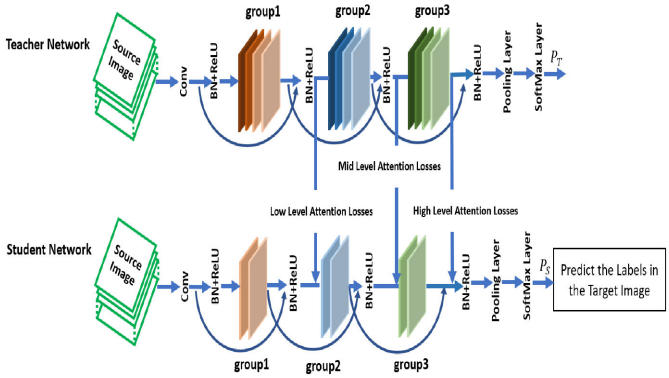

Fig. 1. Block diagram of the proposed method.

III. Proposed Method

A. Overview of the Architecture

Motivated by the recent success of the attention-based transfer learning for object classification [12], we proposed a DL-based attention transfer architecture in [13] for DA to classify hyperspectral imagery. The overview has been presented in Fig. 1. The proposed architecture mainly consists two vital components: the TN and the SN (elaborated in Fig. 1). The TN, which is a deep residual network, transfers knowledge to improve the performance of a shallow pretrained network, namely, SN. These components are triggered by the central element of these networks, the DL architecture, basically, a series of convolution layers to extract hierarchical features at different levels

automatically from the raw input image. These hierarchical features help to produce different levels of mappings, which are critical to make architecture to focus on the specific parts of the image. These levels of mappings are known as attention mappings. For different levels of convolutional layers, different levels of attention mappings can be realized. In this work, we have exploited the attention mappings for the TN to transfer them to the SN for classifying the test regions for the given HSI. Since it would be computationally expensive to feed the whole attention mappings at different levels from the TN to the SN, we have opted to use attention losses as the passing parameters from the TN to the SN to reduce the computational cost of the proposed architecture.

B. DL Architecture

As already mentioned, the state-of-the-art DL architecture like CNN and its different variants are widely used and demonstrated to be very successful in many complex computer vision tasks. The fundamental blueprint of CNN follows a multiple convolutional filters to give us a hierarchical representation of the given image. The formulation of such blueprint can be constructed in two ways: shallow and wide. A shallow CNN is the initial architecture of CNN to be used for many complex computer vision tasks [35]. However, as the network size grows large, a shallow CNN has the problem of gradient explosion [36]. To overcome this limitation, a much deeper and wider blueprint of CNN has been proposed in [36]. This genre of CNN is found to be more successful than a shallower one for many computer vision tasks, widely known as a wide residual network (WRN). The main difference is that it allows to use convolutional layers with different width and depth by using the residual blocks of convolutional filters in the architecture. This residual blocks help to overcome the problem of gradient explosion after including series of convolutional layers. In [12], it is also shown that a wider and deeper residual network achieves better classification accuracy than a less wider and deeper residual network. In this work, both the TN and SN are WRNs. The number of feature maps in each WRN has been determined by two hyperparameters, namely, width and depth. These two parameters decide how deep and wide the WRN can be in terms of generating feature maps. In our case, the TN is deeper in terms of number of the components than that of the SN. The feature map size per each layer in TN is higher than that of the SN. Both the TN and SN initially pass through a convolutional layer with eight filters as the output. Then, as shown in Fig. 1, the resultant feature maps are fed into three groups of convolutional layers. Each group is initially fed by batch normalization, and ReLU layers. These groups produce hierarchical features in a residual manner, which is indicated by skipping connection of the input to the output layer in each group. The first group of convolutional layer generates an output with dimension of four filters multiplied with the width of the layer. The second group of convolutional filter outputs eight filters multiplied with the width of the layer to produce the residual features. The final group of convolutional filters uses 16 filters multiplied with the width of the layer to produce its output. The output of the final group of convolution is passed through batch normalization, ReLU, the pooling (average pooling is used), and the softmax layer to predict the label of the given set of pixels in HSI domain. This whole operation is repeated for the given number of depth. Different combinations of width and depth have been considered for the SN for a given TN, which have been discussed detailed in Section IV. The three blocks represent the generated low-, mid-, and high-level of features, respectively. These levels of features represent different mappings of attentions in the network for the given source regions of the HSI. However, transferring all the attention mappings from low-, mid-, and high-level features from TN would be computationally expensive. Therefore, we have opted to transfer attention losses.

C. Attention Losses

The attention loss parameters mainly share the loss of each residual block level in a WRN to give low-, mid-, and high-level losses. Basically, these attention loss parameters act as the hints for the SN parameters to improve classification. Initially, a WRN with a certain depth and width is trained to get the parameters for the TN with some extra nontrainable attention loss parameters. Then, with the aid of the TN, the SN is trained in the same model architecture but these some extra nontrainable attention loss parameters from the TN is added as extra parameters for training. These losses mainly give the summary of features at various stages of the WRN while training the input image. In this work, the following attention losses are considered.

1) Knowledge Distillation (KD)-Based Attention Loss: KD is the base line transfer learning method, which is introduced in [37]. In our work, attention takes the form of the knowledge to be transferred from the TN to the SN. Initially, the TN is trained with the cross entropy as the function between the ground truth and the predicted labels. Then, at the first step of KD, the probability of each class p_i is calculated from the softmax distribution of the classes, i.e.,

$$p_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}. \tag{1}$$

Here, z_i is the presoftmax logit for class i. T is the temperature, which controls the amount of attention to be distilled in the SN. If T is high, then we will have a softer probability distribution, i.e., all classes have almost the same probability. Therefore, more incorrect decisions can be made while classification, which enables the hidden attention from the incorrect classes to become more prominent to be distilled. The attention learned from training the TN with the normal softmax (T=1) can be distilled and partially transferred to the SN by minimizing the following loss function, which is known as the KD-based attention loss:

$$L_{\text{KD}} = \alpha T^2 \text{CE}(P_S, P_T) + (1-\alpha)\text{CE}(P_S, y_{\text{true}}). \tag{2}$$

In (2), CE stands for cross entropy, y_{true} is the true label of the class. P_S, P_T are the softened probability of the SN and the TN for the same temperature T, respectively, and α is the tuning parameter to tune the weighted average between the two components in the loss function. The first component in the loss function helps to focus on the similarity between the soften

distributions of both the TN and the SN, whereas the second one tries to optimize the soften distribution of the SN with respect to the ground truth label of the data. However, using KD, only for attention transferring has the limitations of difficulty to optimize the pararmeters in very deep networks [12], [38]. Therefore, we opt to use activation-based attention loss.

2) Activation-Based Attention Loss: Basically, we can approximate the dimension of any active layer of a CNN as $M \times N \times P$, where M is the number of feature planes with the spatial dimension of $N \times P$. An activation map outputs a spatial attention map, i.e., a flattened 2-D tensor defined over the spatial dimensions $N \times P$. It means it reduces the dimension of the active layer by the means of a statistic of feature plane to get the reduced map aka attention map [12]. In this work, the activation-based spatial attention maps are generated as

$$C = \sum_{i=1}^{M} |T_i|^w \tag{3}$$

where T_i is the ith feature plane of tensor T $[M \times N \times P]$, which denotes the shape of current active layer, and C is the obtained activation-based spatial attention maps. The absolute value of the tensor T is taken because it helps to indicate about the importance of a particular hidden neuron activation with respect to the specific input. Therefore, by computing the sum of the absolute values raised to the power of w over the M feature planes, a spatial-activation-based attention map can be generated for the given tensor T. From (3), it is seen that the spatial map C can control the discriminative property by changing the weights to spatial location, which is related with the neurons for high activation. The more power w is increased, the more attention is given toward the spatial location that corresponds to the neurons with the highest activation. We can also get attention mapping over spatial locations that can carry multiple neurons with high activation. Once the activation mapping is obtained, the activation-based attention loss is calculated as

$$L_{\text{AT}} = \text{CE}(P_S, y_{\text{true}}) + \frac{\beta}{2} \sum_{j \epsilon I} || \frac{Y_S^j}{||Y_S^j||_2} - \frac{Y_T^j}{||Y_T^j||_2} ||_p. \tag{4}$$

Here, $Y_S^j = \text{vec}(C_S^j)$, $Y_T^j = \text{vec}(C_T^j)$ are the vectorized jth activation-based attention maps for the SN and TN, respectively, and p define the norm type. In this work, we have selected $p = 2$ to make the attention mapping as the l-2 normalized. β is the hyperparameter to control how much attention can be transferred from the TN to the SN.

3) Combined Activation and KD-Based Attention Losses: We use the combined attention loss including both the activation and KD-based attention losses, which is defined as

$$\begin{aligned} L_{\text{AT+KD}} &= L_{\text{AT}} + L_{\text{KD}} \\ &= \frac{\beta}{2} \sum_{j \epsilon I} || \frac{Y_S^j}{||Y_S^j||_2} - \frac{Y_T^j}{||Y_T^j||_2} ||_p \\ &\quad + \alpha T^2 \text{CE}(P_S, P_T) + (1-\alpha)\text{CE}(P_S, y_{\text{true}}). \end{aligned} \tag{5}$$

The common term between (2) and (5) is incorporated as a single term to compute the total loss due to activation mapping and KD.

4) Gradient-Based Attention Loss: In this work, we consider to transfer the gradient-based attention loss from the deeper TN to shallow SN. The gradient-based attention loss is given as

$$\begin{aligned} L_{\text{GD}} &= \alpha T^2 \text{CE}(P_S, P_T) + (1-\alpha)\text{CE}(P_S, y_{\text{true}}) \\ &\quad + \frac{\beta}{2} ||J_S - J_T||_2 \end{aligned} \tag{6}$$

where J_S and J_T is the gradient of the trainable activation mappings of the SN and TN, respectively, i.e., $J_S = \frac{\partial Y_S}{\partial W_S}$ and $J_T = \frac{\partial Y_T}{\partial W_T}$.

Overall steps of the proposed method can be summarized as follows.

1) Decide for a selected number of patches with different patch sizes in source and target HSI regions.
2) Feed the source region into the wide residual TN for a particular width and depth.
3) Pass the image through the three residual blocks of 4, 8, and 16 filters with each block being repeated several times.
4) Use (3) to extract the spatial–spectral attention maps for each block with $w = 2$ to get the high-, mid-, and low-levels of attention maps.
5) Store these maps as nontrainable parameters for the TN while training it.
6) Repeat steps 2) to 5) for the same image in case for residual SN with the different combinations of depth and width to get the spatial attention maps at different levels.
7) Load the nontrainable attention maps from the already trained TN obtained in 5) and compare with the attention maps of the SN in 6) while training the SN using (5) or (6). The parameters of the pretrained TN are fixed in this process.
8) For testing phase, find the performance measures by feeding the randomly selected samples from the target regions to the trained SN.
9) Find land-cover classification map, by using the whole target region with the trained SN.

IV. Experiments

In this work, we have employed Pavia city dataset [39], which can resemble the shifting of pixel intensities between the source and the target domains. This dataset uses the ROSIS-03 hyperspectral sensor with a spatial resolution of 1.3 m over the city of Pavia, Italy. A total of 102 bands has been obtained within the spectrum region between 430 and 860 nm. In this work, for this urban setting dataset, four thematic classes have been considered. The selected labels are buildings, roads, shadows, and vegetation. For land-cover classification task, the natural variation of vegetation produces a significant difference of the spectral signatures of the given classes across the whole image. Therefore, to account this difference, we have opted to split the whole image into different source and target subregions. The selected source subregion contains the patch of 172×123 pixels, whereas the target one takes a larger patch of 350×350. Since, the size of the source subregion is very small compared to the target one; therefore, variation of the spectral signatures over all the classes in the entire image cannot be fully represented. To get

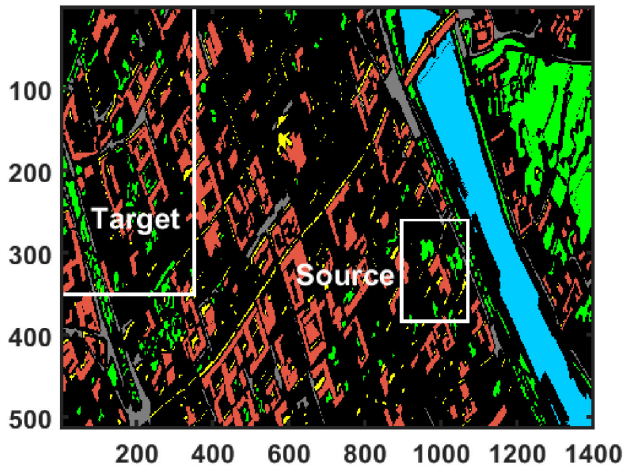

Fig. 2. Sample source and target subregions for the ROSIS image of the city of PAVIA.

the information of the target subregions from such less number of source subregions, domain adaption is required. The sample source and target subregions of the ROSIS image is shown in Fig. 2, which are almost exactly the same source and target regions as shown in [19] and [20] and different from the source and target regions in [13].

A. Experimental Settings

Each pixel in both target and source regions is normalized to zero mean and unit variance. In total, 200 pixels are selected randomly from each class in the source domain for training the TN. Out of the 200 training pixels from each class, 85% of them are used for training the model, and rest of them are used for validation. Then, $200 \times c$ pixels are selected randomly in the target region for training and testing the SN, where c represents the total number of classes. After testing the network, the unknown regions are presented to the SN to generate the knowledge about the unknown label in the image. The depth and width of the TN are fixed as 28 and 10, respectively, whereas the SNs depth and width are 16 and 2, respectively. We have selected these size based on the typical WRN settings [36]. The number of filter maps for the three groups of convolutional layers for both the WRNs of TN and SNs are 4, 8, and 16. For optimization process, stochastic gradient descent is used with exponential learning rate decay starting from 0.1. We set the batch size to 16 with 200 epochs, weight decay of 0.05 in each trainable layer, and all the weights are initialized from Gaussian distribution with zero mean and unit variance. The values of T, α, and β contributing to the attention losses are set to 2, 0.1, and 0.5, respectively. First, the TN is pretrained with the source data. Then, while the training the SN, the parameters of the TN are fixed. The baseline is created after testing the TN, i.e., WRN on the target region.

B. Results and Analysis

The performance of the proposed method is reported with respect to the quantitative measures including OA, average accuracy (AA), and kappa statistics [40]. OA shows the number of correctly classified testing samples to the total number of testing samples in the target regions. AA is the ratio of the sum of the accuracy figures for each class to the number of classes in the target regions. Kappa statistic estimates the agreement between the classification results and the ground truth where full agreement shows the complete alignment of the predicted results with the ground truth while complete misalignment of the results indicates the randomness is dominant in the label prediction. The performance measures with respect to the compared methods with different patch sizes are reported in Table I. The patch size of 1×1 denotes the spectral features of both the source and target regions as in [13]. The label in this case are the labels associated with each given pixel. In other patch size cases, the label is decided based on the central pixel label of the given patch. Since the DL network can extract features on a particular width and depth, using convolution can perform spatial–spectral learning altogether from a given region. It has been seen that the proposed method improves the performance if we adopt learning features jointly from both spatial and spectral domains. Table I shows the results of the performance measures used in this work with different testing regions for the TN or SN. Src-TN represents training and testing data from the source region, whereas Tgt-TN shows the performance of the TN network trained in the source region but tested on the target region. These results are included for creating the baseline for the proposed method. Rest of the results demonstrates the performance of the SN on the target region. The bold faced measures in Table I shows the performance of the shallow SN after training the transferred parameters from the deeper TN. We have used the loss-activation-based mappings in (5) as the loss parameters to produce this result. It has seen that the architecture deriving features and labels from the 3×3 patch size provides the best performance measures as highlighted with underline in Table I. It is obvious that after introducing the attention in the SN, the performance of the proposed method has improved compared to the case than without using the attention in the SN. It can be seen that our proposed method with spatial–spectral features outperforms the compared state-of-the-art methods in [19] and [20]. These methods exploit traditional ML [19] and DL [20] entities for DA, based on in-variance and representational nature of the features, respectively. These methods are very useful for extracting complex features; however, our proposed method helps to produce not only more complex but also useful features for the classification task. The attention mechanism nature of the proposed method helps to locate and produce such kind of features. Due to this structure of the proposed architecture, better performances can be provided. In addition to the result reported in Table I, we have also evaluated the performance of shallow but wider SNs to see whether the introduction of the attention helps improve their classification accuracy. The results of different wider shallower SN with respect to the given TN has been reported in Table II. It can be seen that as the network becomes wider, the shallow network tends to perform better and yield better estimates of performance measures than the original shallow one (16×2 SN). Similar to the previous cases, in this case, we also utilize the attention loss in (5). However, obtained

TABLE I
DA CLASSIFICATION RESULTS FOR PAVIA DATASET (SRC:SOURCE, TGT:TARGET, PM:PROPOSED METHOD, W/ATT:WITH ATTENTION, W/O ATT:WITHOUT ATTENTION) (TN: 28 × 10, SN:16 × 2)

Method	Kappa		OA%		AA%	
	Avg.	Std.	Avg.	Std.	Avg.	Std.
Src-TN (1×1) [13]	0.8820	0.0033	96.47	0.0032	96.87	0.0043
Tgt-TN (1×1) [13]	0.8740	0.0036	95.39	0.0036	96.13	0.0048
PM-W/O Att (1×1) [13]	0.8684	0.0040	93.19	0.0039	94.71	0.0049
PM-W/Att (1×1) [13]	**0.8730**	**0.0037**	**94.35**	**0.0037**	**95.17**	**0.0047**
Src-TN (3×3)	0.9002	0.0029	98.61	0.0027	98.39	0.0037
Tgt-TN (3×3)	0.8959	0.0034	98.23	0.0031	98.15	0.0035
PM-W/O Att (3×3)	0.8772	0.0039	95.62	0.0029	96.31	0.0032
PM-W/Att (3×3)	**0.8828**	**0.0041**	**96.23**	**0.0035**	**98.84**	**0.0036**
Src-TN (5×5)	0.8779	0.0038	97.50	0.0034	97.91	0.0041
Tgt-TN (5×5)	0.8834	0.0032	97.08	0.0029	98.03	0.0037
PM-W/O Att (5×5)	0.8739	0.0046	95.45	0.0032	96.32	0.0034
PM-W/Att (5×5)	**0.8798**	**0.0033**	**95.89**	**0.0039**	**96.42**	**0.0036**
Src-TN (7×7)	0.8764	0.0033	97.61	0.0036	97.80	0.0038
Tgt-TN (7×7)	0.8810	0.0035	96.87	0.0040	97.91	0.0047
PM-W/O Att (7×7)	0.8761	0.0036	95.37	0.0030	96.23	0.0032
PM-W/Att (7×7)	**0.8776**	**0.0039**	**95.71**	**0.0038**	**96.37**	**0.0046**
DAE [20]	0.8680	0.0050	92.40	0.0030	93.20	0.0070
DANN [20]	0.8680	0.0050	92.60	0.0030	85.40	0.0100
SSTCA [19]	0.8470	0.0100	91.10	0.0060	92.90	0.0040

TABLE II
DA CLASSIFICATION RESULTS FOR PAVIA DATASET WITH DIFFERENT WIDTH SIZE OF THE SN (SRC:SOURCE, TGT:TARGET, PM:PROPOSED METHOD, W/ATT:WITH ATTENTION, W/O ATT:WITHOUT ATTENTION) (TN: 28 × 10, PATCH SIZE: 3 × 3)

Method	Network Size	Kappa		OA%		AA%	
		Avg	Std	Avg	Std	Avg	Std
Src-TN	16×2	0.9002	0.0029	98.61	0.0027	98.39	0.0037
Tgt-TN		0.8959	0.0034	98.23	0.0031	98.15	0.0035
PM-W/O Att		0.8772	0.0039	95.62	0.0029	96.31	0.0032
PM-W/Att		**0.8828**	**0.0041**	**96.23**	**0.0035**	**96.84**	**0.0036**
PM-W/O Att	16×4	0.8831	0.0042	95.97	0.0034	96.59	0.0036
PM-W/Att		**0.8880**	**0.0036**	**96.54**	**0.0031**	**97.13**	**0.0039**
PM-W/O Att	16×8	0.8842	0.0049	96.19	0.0034	96.59	0.0036
PM-W/Att		**0.8851**	**0.0044**	**96.83**	**0.0031**	**97.49**	**0.0039**
PM-W/O Att	16×16	0.8882	0.0040	96.35	0.0055	96.77	0.0043
PM-W/Att		**0.8893**	**0.0049**	**97.24**	**0.0046**	**97.78**	**0.0051**
PM-W/O Att	16×32	0.8925	0.0052	96.73	0.0036	97.18	0.0035
PM-W/Att		**0.8914**	**0.0043**	**97.53**	**0.0042**	**98.07**	**0.0037**

TABLE III
COMPUTATIONAL COMPARISON FOR DIFFERENT CONFIGURATIONS OF THE SN AND TN

Network Size	Number of Parameters	Testing time (in Sec)
TN (28 ×10)	9.2M	0.24
SN (16 ×2)	215K	0.03
SN (16 ×4)	733k	0.10
SN (16 ×8)	2.8M	0.15
SN (16 ×16)	7M	0.21
SN (16 ×22)	23M	0.53

improvement comes with some computational overhead. The computational cost in terms of number of parameters and testing time has been reported in Table III. All simulations are run on a DL machine with 2 NVIDIA Titan RTX GPUs. It can be seen that as the shallow network becomes wider the computational complexity increases. However, we can still get a very high accuracy comparing to state-of-the-art techniques by using fewer parameters in the SN while getting attention from a very wide and deep TN.

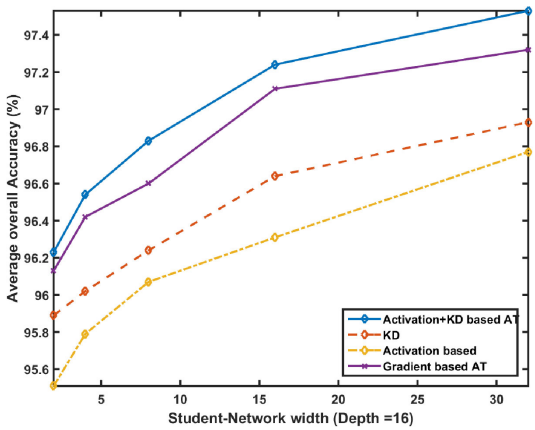

Fig. 3. Performance comparison of different attention losses as a function of SN width (for SN depth of 16, and TN size of 28 × 10).

Moreover, we have also exploited attention losses due to gradient activation equation in (6) and KD in (2) to show the performance of the proposed attention transferring DL architecture.

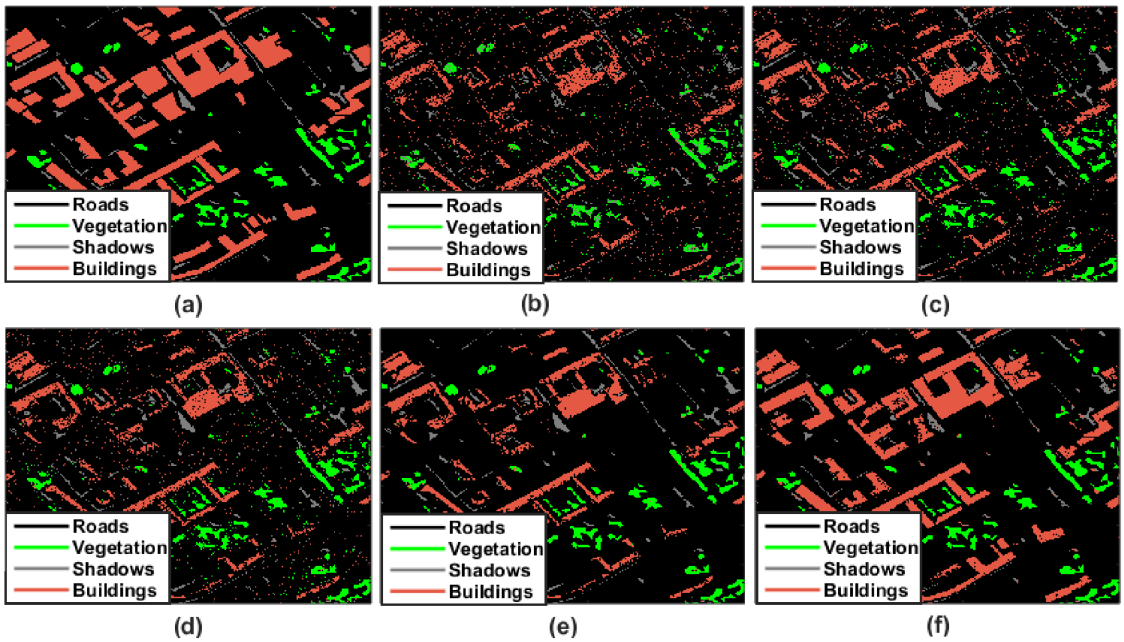

Fig. 4. Labels prediction of a sample target region. (a) Ground Truth. (b) Without attention transfer. (c) Activation-based attention. (d) KD-based attention. (e) Gradient-based attention. (f) Activation+KD based attention.

As shown in Fig. 3, the activation-based attention outperforms other comparing attention mechanisms, but gradient-based attention performs very close to it. It is seen that introducing various forms of attentions in the HSI domain can lead to produce different attention maps for DA case. It also shows that activation and gradient-based attentions carry more useful and important information than that of full activations. Furthermore, this result indicates the importance of transferring attention maps in cases where spatial information is more important than that of spectral domain. Intuitively from Fig. 3, we can also say that activation and gradient-based attention mappings preserve the most important spatial information of the source domain captured by the neurons in the TN to enhance the abilities of the SN to extract more useful features from the source domain, and contribute to the improved performance of DA on the target region.

The learning of the labels by the SN for a sample region before and after introducing the attention mechanism is shown in Fig. 4. The ground truth of the target region is presented in Fig. 4(a). Following figures from Fig. 4(b)–(f) show the labels learned by the different configuration of SN. Fig. 4(b) shows the land-cover map obtained without using attention, and Fig. 4(c)–(f) shows the results using the activation-based, KD, gradient, and activation+KD-based attention, respectively. All the results are produced with 3×3 patches. It can be seen that the result from using activation-based attention in Fig. 4(f) is closer to the ground truth.

V. Conclusion and Future Work

In this work, an attention-based DA for HSI classification has been proposed. It has been shown that the proposed deep transfer learning architecture outperforms state-of-the-art methods adopted for DA. In particular, a simpler student network with fewer parameters and faster testing time can yield the performance comparative to a deeper and wider teacher network. The future work mainly involves investigating transferred parameters to increase the overall classification performance in DA.

References

[1] M. Simões, J. Bioucas-Dias, L. B. Almeida, and J. Chanussot, "A convex formulation for hyperspectral image superresolution via subspace-based regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3373–3388, Jun. 2015.

[2] Y. Xu *et al.*, "Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 IEEE GRSS data fusion contest," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1709–1724, Jun. 2019.

[3] D. Tuia, C. Persello, and L. Bruzzone, "Domain adaptation for the classification of remote sensing data: An overview of recent advances," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 41–57, Jun. 2016.

[4] L. Bruzzone and M. Marconcini, "Toward the automatic updating of land-cover maps by a domain-adaptation SVM classifier and a circular validation strategy," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 4, pp. 1108–1122, Apr. 2009.

[5] K. Bahirat, F. Bovolo, L. Bruzzone, and S. Chaudhuri, "A novel domain adaptation Bayesian classifier for updating land-cover maps with class differences in source and target domains," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 7, pp. 2810–2826, Jul. 2012.

[6] Y. Liu and X. Li, "Domain adaptation for land use classification: A spatio-temporal knowledge reusing method," *Int. J. Geo-Inf. J. Photogrammetry Remote Sens.*, vol. 98, pp. 133–144, 2014.

[7] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson, "Advances in hyperspectral image classification: Earth monitoring with statistical learning methods," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 45–54, Jan. 2014.

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

[9] 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.
[10] W. Li, G. Wu, F. Zhang, and Q. Du, "Hyperspectral image classification using deep pixel-pair features," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 844–853, Feb. 2017.
[11] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Oct. 2016.
[12] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *CoRR*, abs/1612.03928, 2016.
[13] R. H. Md. Rafi, B. Tang, Q. Du, and N. H. Younan, "Attention-based domain adaptation for hyperspectral image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 67–70.
[14] L. Bruzzone and M. Marconcini, "Domain adaptation problems: A DASVM classification technique and a circular validation strategy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 770–787, May 2010.
[15] G. Matasci, D. Tuia, and M. Kanevski, "SVM-based boosting of active learning strategies for efficient domain adaptation," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 5, no. 5, pp. 1335–1343, Oct. 2012.
[16] X. Li, L. Zhang, B. Du, L. Zhang, and Q. Shi, "Iterative reweighting heterogeneous transfer learning framework for supervised remote sensing image classification," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 10, no. 5, pp. 2022–2035, May 2017.
[17] C. Deng, X. Liu, C. Li, and D. Tao, "Active multi-kernel domain adaptation for hyperspectral image classification," *Pattern Recognit.*, vol. 77, pp. 306–315, 2018.
[18] L. Bruzzone and C. Persello, "A novel approach to the selection of spatially invariant features for the classification of hyperspectral images with improved generalization capability," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 9, pp. 3180–3191, Sep. 2009.
[19] G. Matasci, M. Volpi, M. Kanevski, L. Bruzzone, and D. Tuia, "Semisupervised transfer component analysis for domain adaptation in remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3550–3564, Jul. 2015.
[20] A. Elshamli, G. W. Taylor, A. Berg, and S. Areibi, "Domain adaptation using representation learning for the classification of remote sensing images," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 10, no. 9, pp. 4198–4209, Sep. 2017.
[21] J. Zabalza *et al.*, "Novel segmented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging," *Neurocomputing*, vol. 185, pp. 1–10, 2016.
[22] J. Gao, W. Fan, J. Jiang, and J. Han, "Knowledge transfer via multiple model local structure mapping," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2008, pp. 283–291.
[23] A. Khosla, T. Zhou, T. Malisiewicz, A. A. Efros, and A. Torralba, "Undoing the damage of dataset bias," in *Proc. Eur. Conf. Comput. Vision*, 2012, pp. 158–171.
[24] C. E. Woodcock, S. A. Macomber, M. Pax-Lenney, and W. B. Cohen, "Monitoring large areas for forest change using landsat: Generalization across space, time and landsat sensors," *Remote Sens. Environ.*, vol. 78, no. 1–2, pp. 194–203, 2001.
[25] D. Tuia, J. Munoz-Mari, L. Gomez-Chova, and J. Malo, "Graph matching for adaptation in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 329–341, Jan. 2013.
[26] Y. Liu, W. Tu, B. Du, L. Zhang, and D. Tao, "Homologous component analysis for domain adaptation," *IEEE Trans. Image Process.*, vol. 29, pp. 1074–1089, 2020.
[27] L. Mihalkova, T. Huynh, and R. J. Mooney, "Mapping and revising Markov logic networks for transfer learning," in *Proc. 22nd Nat. Conf. Artif. Intell.*, 2007, vol. 7, pp. 608–614.
[28] C. Paris and L. Bruzzone, "A sensor-driven hierarchical method for domain adaptation in classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1308–1324, Mar. 2018.
[29] S. Chakraborty and M. Roy, "A neural approach under transfer learning for domain adaptation in land-cover classification using two-level cluster mapping," *Appl. Soft Comput.*, vol. 64, pp. 508–525, 2018.
[30] Y. Xu, L. Zhang, B. Du, and F. Zhang, "Spectral-spatial unified networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 5893–5909, Oct. 2018.
[31] Y. Yan *et al.*, "Unsupervised image saliency detection with Gestalt-laws guided optimization and visual attention based refinement," *Pattern Recognit.*, vol. 79, pp. 65–78, 2018.
[32] P. Wu, Z. Cui, Z. Gan, and F. Liu, "Residual group channel and space attention network for hyperspectral image classification," *Remote Sens.*, vol. 12, no. 12, 2020, Art. no. 2035.
[33] C. Mu, Z. Guo, and Y. Liu, "A multi-scale and multi-level spectral-spatial feature fusion network for hyperspectral image classification," *Remote Sens.*, vol. 12, no. 1, 2020, Art. no. 125.
[34] G. Sun *et al.*, "Deep fusion of localized spectral features and multi-scale spatial features for effective classification of hyperspectral images," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 91, 2020, Art. no. 102157.
[35] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[36] S. Zagoruyko and N. Komodakis, "Wide residual networks," *CoRR*, abs/1605.07146, 2016.
[37] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
[38] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7130–7138.
[39] G. Licciardi *et al.*, "Decision fusion for the classification of hyperspectral data: Outcome of the 2008 GRS-S data fusion contest," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 11, pp. 3857–3865, Nov. 2009.
[40] R. G. Congalton, "A review of assessing the accuracy of classifications of remotely sensed data," *Remote Sens. Environ.*, vol. 37, no. 1, pp. 35–46, 1991.

Robiulhossain Mdrafi (Graduate Student Member, IEEE) received the B.Sc. degree in electrical, electronic, and communication engineering from the Bangladesh University of Professionals, Dhaka, Bangladesh, in 2010. He is currently working toward the Ph.D. degree in electrical and computer engineering with Mississippi State University (MS-State), Starkville, MS, USA.

He is currently a Research Assistant with the Information Processing and Sensing Laboratory, MS-State. His research interests include deep-learning-based inverse problems, such as sparse signal/image reconstruction, physics aware deep learning, and learning in radar and remote sensing applications.

Qian Du (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of Maryland at Baltimore, Baltimore, MD, USA, in 2000.

She is currently the Bobby Shackouls Professor with the Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS, USA. Her research interests include hyperspectral remote sensing image analysis and applications, and machine learning.

Dr. Du is a Fellow of the SPIE—International Society for Optics and Photonics. She was the Chair of the Remote Sensing and Mapping Technical Committee of the International Association for Pattern Recognition from 2010 to 2014. She was the Co-Chair for the Data Fusion Technical Committee of the IEEE GRSS from 2009 to 2013. She was the General Chair for the fourth IEEE GRSS Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing held at Shanghai, in 2012. She was an Associate Editor for the IEEE JOURNAL OF TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING from 2011 to 2015, the *Journal of Applied Remote Sensing* from 2014 to 2015, and the IEEE SIGNAL PROCESSING LETTERS from 2012 to 2015. She is also the Editor-in-Chief of the IEEE JOURNAL OF TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING from 2016 to 2020.

Ali Cafer Gurbuz (Senior Member, IEEE) received B.S. degree in electrical engineering from Bilkent University, Ankara, Turkey, in 2003, and the M.S. and Ph.D. degrees in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2005 and 2008, respectively.

From 2003 to 2009, he researched compressive sensing-based computational imaging problems with Georgia Tech, Atlanta, GA, USA. Between 2009 and 2017, he held faculty positions with TOBB University, Ankara, and The University of Alabama, Tuscaloosa, AL, USA, where he pursued an active research program on the development of sparse signal representations, compressive sensing theory and applications, radar and sensor array signal processing, and machine learning. He is currently an Assistant Professor with Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS, USA, where he is a Co-Director of Information Processing and Sensing (IMPRESS) Laboratory.

Dr. Gurbuz is the recipient of The Best Paper Award for Signal Processing Journal in 2013 and the Turkish Academy of Sciences Best Young Scholar Award in Electrical Engineering, in 2014. He was an Associate Editor for several journals, such as *Digital Signal Processing*, *EURASIP Journal on Advances in Signal Processing*, and *Physical Communications*.

Bo Tang (Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Rhode Island, Kingstown, RI, USA, in 2016.

He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS, USA. From 2016 to 2017, he was an Assistant Professor with the Department of Computer Science, Hofstra University, Hempstead, NY, USA. His research interests include the general areas of statistical machine learning and data mining, as well as their various applications in cyber-physical systems, including robotics, autonomous driving, and remote sensing.

Li Ma (Member, IEEE) received the B.S. and M.S. degree in pattern recognition and intelligent system from Shandong University, Jinan, China, in 2004 and 2006, respectively, and the Ph.D. degree in pattern recognition and intelligent system from Huazhong University of Science and Technology, Wuhan, China, in 2011.

From 2008 to 2010, she was a Visiting Scholar with Purdue University, West Lafayette, IN, USA. She also visited Mississippi State University, Starkville, MS, USA, in 2018. She is currently an Associate Professor with the School of Mechanical Engineering and Electronic Information, China University of Geosciences, Wuhan. Her research interests include hyperspectral data analysis, pattern recognition, and remote sensing applications.

Nicolas H. Younan (Life Senior Member, IEEE) received the B.S. and M.S. degrees in electrical and computer engineering from Mississippi State University, in 1982 and 1984, respectively, and the Ph.D. degree from Ohio University in 1988.

He is currently a Department Head Emeritus and a Professor Emeritus of Electrical and Computer Engineering with Mississippi State University, where he was the Department Head and the James Worth Bagley Chair from 2009 to 2019. He has authored/coauthored more than 300 papers in journals and refereed conference proceedings. He has been involved in the development of advanced signal processing and pattern recognition algorithms for data mining, data fusion, feature extraction and classification, and automatic target recognition/identification. His research interests include signal processing and pattern recognition with applications to smart technologies.

Dr. Younan was the General Chair and Editor for the 4th IASTED International Conference on Signal and Image Processing, the Co-Editor for the 3rd International Workshop on the Analysis of Multi-Temporal Remote Sensing Images, the Guest Editor for *Sensors Journal*, *Remote Sensing*, *Electronics*, *Pattern Recognition Letters*, and IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, and a Co-Chair for the Workshop on Pattern Recognition for Remote sensing from 2008 to 2010. He is a member of the IEEE Geoscience and Remote Sensing Society, serving on two technical committees: Image Analysis and Data Fusion, and Earth Science Informatics (previously Data Archive and Distribution). He was also the Vice-Chair of the International Association on Pattern Recognition Technical Committee 7 on Remote Sensing from 2008 to 2010.