Robust Self-Ensembling Network for Hyperspectral Image Classification

Yonghao Xu, Member, IEEE, Bo Du, Senior Member, IEEE, and Liangpei Zhang, Fellow, IEEE

Abstract—Recent research has shown the great potential of deep learning algorithms in the hyperspectral image (HSI) classification task. Nevertheless, training these models usually requires a large amount of labeled data. Since the collection of pixel-level annotations for HSI is laborious and time-consuming, developing algorithms that can yield good performance in the small sample size situation is of great significance. In this study, we propose a robust self-ensembling network (RSEN) to address this problem. The proposed RSEN consists of two subnetworks including a base network and an ensemble network. With the constraint of both the supervised loss from the labeled data and the unsupervised loss from the unlabeled data, the base network and the ensemble network can learn from each other, achieving the self-ensembling mechanism. To the best of our knowledge, the proposed method is the first attempt to introduce the self-ensembling technique into the HSI classification task, which provides a different view on how to utilize the unlabeled data in HSI to assist the network training. We further propose a novel consistency filter to increase the robustness of self-ensembling learning. Extensive experiments on three benchmark HSI datasets demonstrate that the proposed algorithm can yield competitive performance compared with the state-of-the-art methods.

Index Terms—Convolutional neural network (CNN), deep learning, hyperspectral image (HSI) classification, self-ensembling.

I. INTRODUCTION

RECENT years have witnessed the rapid development of hyperspectral imaging systems [1]. The abundant spectral–spatial information contained in hyperspectral images (HSIs) makes it possible to observe the Earth in a more accurate way [2]. As a result, the classification of remotely sensed HSIs has become a significant task in the remote sensing community, which aims to predict the semantic label for each pixel in the image.

Compared with traditional image classification tasks, HSI classification is more challenging because of the Hughes phenomenon [3], which is also known as the curse of dimensionality [4]. An intuitive way to address this problem is to reduce the dimension of the original hyperspectral data. This technique is known as dimension reduction [5], which can be further divided into feature selection [6] and feature extraction [7]. Besides the high dimensionality of hyperspectral data, the spectral heterogeneity also leads to poor inter-class separability [8]. To this end, recent approaches try to combine both the spatial and spectral features in the classification [9]. Some representative spatial feature extraction methods include morphological profiles [10], gray-level co-occurrence matrix [11], and Gabor filters [12]. The limitation of these methods lies in two aspects. First, handcrafted features like morphological profiles rely on expert knowledge from the designer, which may lead to the poor ability of adaptation between different scenes [13]. Second, most of these methods extract the features in a shallow manner, which may lead to unstable performance when confronted with complex imaging environments [14].

Recent advances in deep learning algorithms have provided an alternative way to address the remote sensing tasks [15], [16], [17]. Compared to shallow features, deep features are more robust, abstract, and discriminative, thus showing a stronger ability to represent the spectral, textural, and geometrical attributes of the hyperspectral data [14].

Chen et al. [18] introduced the concept of deep learning into the HSI classification task for the first time. By training a multilayer stacked autoencoder (SAE), hierarchical spectral and spatial features can be learned simultaneously. To alleviate the gradient explosion and gradient vanishing problem, the network is first pretrained with unlabeled data, where the goal in this step is to reconstruct the input signal from the hidden layer. Then, the labeled samples are utilized to fine-tune the whole network. In a like manner, the restricted Boltzmann machine (RBM) and the deep belief network (DBN) are also adopted to conduct the spectral-spatial classification of hyperspectral data [19]. Considering the contextual information among adjacent bands in HSIs, the recurrent neural network (RNN)-based approaches are proposed, which provide a different view on deep spectral feature extraction for hyperspectral data [20], [21]. Apart from the aforementioned methods, the
convolutional neural network (CNN)-based methods deserve special attention, because of their powerful feature representation ability [22], [23]. In [22], a 3D-CNN method is proposed. Different from most of the previous methods where the spectral and spatial features are extracted separately, 3D-CNN can directly learn spectral–spatial features from the input hyperspectral cube, achieving the state-of-the-art performance.

Despite the strong learning ability of deep learning algorithms, training these deep neural networks generally requires a large amount of labeled data, since there are thousands or even millions of parameters that need to be learned [14]. Once the training samples are not sufficient enough, the network may tend to be over-fitting, resulting in poor generalization ability [24]. Considering the collection of pixel-level annotations for HSIs is laborious and time-consuming, developing algorithms that can yield good performance in the small sample size situation is of great significance.

While labeled samples are usually difficult and expensive to be collected, the collection of unlabeled data is much easier. Thus, one possible solution is to use both labeled and unlabeled data during network learning. This strategy is also known as semisupervised learning [25]. Mou et al. [26] proposed to use CNN to conduct the semisupervised learning. Similar to [18], the network is first pretrained with unlabeled hyperspectral data to learn an identical mapping function. Then, a small number of training samples are utilized to fine-tune the whole network. Pseudo-label learning is another possible way to tackle this task. Wu and Prasad [25] adopted the k-means clustering algorithm to assign pseudo labels for those unlabeled data. These pseudo labels are further utilized to train the deep convolutional RNNs. Recently, there are also some novel generative adversarial networks (GANs)-based approaches [27], [28]. Zhu et al. [28] proposed to use the GAN to generate realistic virtual samples for hyperspectral data. The original training set is expanded with those generated samples to train the deep network. Nevertheless, since GAN-based methods generally assume that the generated samples share the same feature distribution with the real ones, which is difficult to be guaranteed in practice, the performance of semisupervised learning depends largely on the quality of the generated samples.

Different from the existing approaches which put much effort into labels assignment [25] or samples generation [28], in this article, we introduce a robust self-ensembling network (RSEN) to address this problem. The initial inspiration of our work comes from an observation that an ensemble model of multiple base models generally yields better predictions than a single model does [29], [30]. As shown in Fig. 1, although the original three base classifiers cannot classify the dataset well, the ensemble model achieves this goal by inheriting the strength from each base classifier. Based on this characteristic, the predictions of an ensemble model on the unlabeled data are also likely to be closer to the ground-truth data than the ones of the base model. Thus, a natural idea is to adopt the predictions of the ensemble network as the pseudo labels to assist the training of the base network. As the training goes on, those base networks will be more accurate, which in turn, also helps to generate a stronger ensemble model. This technique is known as self-ensembling learning [31].

Laine and Aila [32] proposed the first self-ensembling model named as Π-model. They form consensus predictions for the unknown samples using the outputs of the network in training on different epochs. Tarvainen and Valpola [33] further designed a mean teacher model that consists of a student network and a teacher network, where both networks share the same architecture. Different from the Π-model that averages label predictions from different epochs, the mean teacher model adopts the averaged model weights from different epochs, which shows better results. French et al. [31] employed the mean teacher model to solve the cross-domain image classification problem and achieved the state-of-the-art performance. Xu et al. [34] further introduced the self-ensembling technique to cross-domain semantic segmentation task. Other related works include co-training for domain adaptation [35], [36], and self-ensembling graph convolutional networks for semisupervised learning [37].

Recall that the theoretical foundation of self-ensembling learning lies on the assumption that the ensemble model can yield more accurate predictions on those unknown data than the base model does. Under this circumstance, the predictions of the ensemble model can be regarded as the pseudo labels to assist the training of the base model, which in turn, helps to generate a more accurate ensemble model, achieving the self-ensembling mechanism. However, in practical applications, the quality of the ensemble model’s predictions can hardly be guaranteed. If the generated pseudo labels contain too many errors, the training of the whole framework would be misguided. This phenomenon may be even more serious for the HSI classification task since the labeled samples are very insufficient in the remote sensing field. Therefore, how to increase the quality of the pseudo labels used in self-ensembling learning is a critical problem.

To address the aforementioned problems, we propose a RSEN for the HSI classification task. The major contributions of this article are summarized as follows.
Fig. 2. Illustration of the proposed RSEN. There are two subnetworks, including a base network and an ensemble network. Both networks share the same architecture. In the training phase, the labeled samples are fed into the base network to calculate the cross-entropy loss, while the unlabeled samples are fed into both the base network and the ensemble network to calculate the mean squared error. At each iteration, the base network is updated with the gradients from both two losses. The ensemble network does not participate in this process. Instead, its parameters are manually updated using the exponential moving average of the historical parameters in the base network.

1) We introduce the self-ensembling learning into the HSI classification task for the first time, which provides a different view on how to utilize the unlabeled data in HSI to assist the training of deep networks with very limited labeled samples.

2) To make self-ensembling learning more efficient, a simple but effective spectral–spatial network architecture is proposed for the implementation of both the base network and the ensemble network in this study. The simplicity of the proposed network architecture enables the framework to achieve high accuracy with a relatively low time cost.

3) Since the predictions of the ensemble model may be inaccurate, directly learning from all the pseudo labels generated by the ensemble model could probably misguide the training of the whole framework. To make the ensemble learning more robust, a novel consistency filtering strategy is further proposed.

The rest of this article is organized as follows. Section II describes the proposed RSEN in detail. Section III presents the information on datasets used in this study and the experimental results. Conclusions are summarized in Section IV.

II. ROBUST SELF-ENSEMBLING NETWORKS

In this section, we will first make a brief overview of the proposed RSEN. Then, the network architecture and the optimization of the framework are presented in detail.

A. Overview of the Proposed Model

Traditional ensemble learning usually requires several base models [38]. Considering the high complexity of the deep neural networks, training multiple base networks would dramatically increase the time cost. To alleviate this issue, we adopt the self-ensembling strategy to achieve ensemble learning.

As shown in Fig. 2, there are two subnetworks in the proposed framework, including a base network and an ensemble network (which correspond to the student network and the teacher network in the mean teacher model [33]). Both networks share the same architecture to extract spectral-spatial features. During the training phase, the labeled samples are fed into the base network to calculate the supervised cross-entropy loss between the base network’s predictions and the ground-truth labels. The unlabeled samples, on the other hand, are fed into both the base network and the ensemble network. Then, the supervised mean squared error is calculated between the base network’s predictions and the ensemble network’s predictions. At each iteration, the base network is updated using the gradients from both the supervised and unsupervised losses mentioned above. The ensemble network does not participate in the back-propagation step. Instead, its parameters are updated with the parameters from the base network at the current iteration and the parameters from the ensemble network at previous iterations. In this way, the base network and the ensemble network can gradually learn from each other, achieving the purpose of ensemble learning. After the training stage, we can then feed those test data...
into the ensemble network to yield the final classification result.

B. Network Architecture

Most existing deep neural networks are very time-consuming in the training stage due to the complex network architecture [18], [22]. Considering the high-computational burden of ensemble learning, directly adopting existing deep models to implement the framework would make the training very inefficient. To tackle this problem, we propose a simple yet effective spectral–spatial network, which is used as the base model in RSEN. For simplicity, we denote this network as BaseNet.

Let $X \in \mathbb{R}^{r \times c \times n}$ be the input HSI, where $r$, $c$, $n$ are the row number, column number, and channel number of the image, respectively. We directly extract the original spectral features of each pixel in $X$ as the input spectral features $X_{\text{spectral}} \in \mathbb{R}^{r \times c \times n}$. For spatial features, the principal component analysis (PCA) is first applied to $X$ to reduce the high dimensionality of the original data, and only the first $p$ principal components are reserved (we set $p$ as 5 in this study). Denote the dimension-reduced data as $X_p \in \mathbb{R}^{r \times c \times p}$. For each pixel in $X_p$, we can extract the corresponding neighbor region with a size of $w \times w \times p$, where $w$ is the spatial size of the patch (we set $w$ as 16 in this study). In this way, we can get the spatial features for all samples in $X$ as $X_{\text{spatial}} \in \mathbb{R}^{r \times c \times w \times w \times p}$.

As shown in Fig. 2, the architecture of the proposed BaseNet consists of two branches. In the upper branch, we simply use a fully connected layer to extract spectral features $h_{\text{spe}}$

$$h_{\text{spe}} = g(W_{\text{spe}}x_{\text{spe}} + b_{\text{spe}})$$

(1)

where $W_{\text{spe}}$ and $b_{\text{spe}}$ are the weight matrix and bias vector in the fully connected layer, respectively. $x_{\text{spe}}$ denotes the input spectral sample. $g(x) = \max(0, x)$ is the rectified linear unit (ReLU) function.

In the lower branch, we first adopt three convolutional layers and two average pooling layers to extract the spatial features $h_{\text{spa}}$. Formally, let $X_{\text{spa}}$ be the input spatial sample. The first convolutional layer contains 64 convolutional kernels with a kernel size of $1 \times 1$ (denoted as $64@1 \times 1$ Conv in Fig. 2), and the corresponding convolutional features $H_{\text{conv}1}$ can be calculated as

$$H_{\text{conv}1} = W_{\text{conv}1} * X_{\text{spa}} + b_{\text{conv}1}$$

(2)

where $W_{\text{conv}1}$ and $b_{\text{conv}1}$ are the weight matrix and bias vector in the first convolutional layer, respectively. $*$ denotes the convolution operation. The second convolutional layer contains 64 convolutional kernels with a kernel size of $3 \times 3$, and the corresponding convolutional features $H_{\text{conv}2}$ can be calculated as

$$H_{\text{conv}2} = W_{\text{conv}2} * H_{\text{conv}1} + b_{\text{conv}2}$$

Then, a $2 \times 2$ average pooling layer is utilized to further increase the receptive field of the convolutional features. We use the skip connection [39] to fuse both the shallow and deep features before the pooling process. The corresponding pooling features $H_{\text{pool}1}$ can be expressed as

$$H_{\text{pool}1} = D_{\text{avg}}(g(H_{\text{conv}1} + H_{\text{conv}2}))$$

(3)

where $D_{\text{avg}}$ denotes the $2 \times 2$ average pooling operation.

In a like manner, we can get $H_{\text{conv}3} = W_{\text{conv}3} * H_{\text{pool}1} + b_{\text{conv}3}$, and $H_{\text{pool}2} = D_{\text{avg}}(g(H_{\text{pool}1} + H_{\text{conv}3}))$. The final deep spatial features $h_{\text{spa}}$ can be obtained by flattening $H_{\text{pool}2}$ into a vector form.

To conduct spectral–spatial classification, $h_{\text{spe}}$ and $h_{\text{spa}}$ are further concatenated into a fusion layer $h_{\text{fusion}} = [h_{\text{spe}}; h_{\text{spa}}]$, followed by two fully-connected layers. Let $h_{\text{cls}}$ be the features in the last fully connected layer. The probability that the input sample belongs to the $i$th category can be calculated by the softmax function $\sigma$

$$\sigma(h_{\text{cls}}(i)) = \frac{e^{h_{\text{cls}}(i)}}{\sum_{j=1}^{k} e^{h_{\text{cls}}(j)}}$$

(4)

where $k$ is the number of categories.

C. Self-Ensembling Learning

Let $f_b$ and $f_e$ denote the mapping functions of base network and ensemble network, respectively. Both networks share the same architecture of BaseNet described in Section II-B. Let $X_l = \{x_{l(i)}\}_{i=1}^{nu_l}$ and $X_u = \{x_{u(i)}\}_{i=1}^{nu_u}$ be the labeled training set and unlabeled training set, respectively, where $nu_l$ and $nu_u$ are the numbers of samples in each set. Let $Y_l = \{y_{l(i)}\}_{i=1}^{nu_l}$ be the corresponding labels for $X_l$. The cross-entropy loss $L_{\text{cls}}$ for the base network can therefore be defined as

$$L_{\text{cls}}(f_b, x_l, y_l) = - \sum_{j=1}^{k} \mathbb{I}[y_{l(i)} = j] \log(f_b(N(x_l))(i))$$

(5)

where $\mathbb{I}$ is the indicator function, and $N(\cdot)$ denotes the stochastic augmentation. We implement $N(\cdot)$ by adding Gaussian noise with a mean of 0 and a standard deviation of 0.5 to each pixel in the image.

With the constraint in (5), the base network can learn supervised information from the labeled training set. To further achieve the self-ensembling mechanism, we define the consistency loss $L_{\text{con}}$ with mean-squared error as

$$L_{\text{con}}(f_b, x_u) = \sum_{j=1}^{k} \left\| f_b(N(x_u))(i) - f_e(N(x_u))(i) \right\|^2.$$  

(6)

Recall that our goal is to let the base model $f_b$ learn from the predictions of the ensemble model $f_e$ on unlabeled data. Therefore, the parameters in $f_e$ does not participate the optimization in (6). The full objective function for training the base model $f_b$ can be formulated as

$$\min_{f_b} (L_{\text{cls}}(f_b, x_l, y_l) + L_{\text{con}}(f_b, x_u)).$$ 

(7)

We optimize (7) by mini-batch stochastic gradient descent algorithm. At each iteration, we first update the parameters in $f_b$ with back-propagation algorithm. Then, the parameters in $f_e$ are manually updated using the exponential moving average of the historical parameters in $f_b$. Let $\theta_{b}^{t}$ and $\theta_{e}^{t}$ denote the parameters of $f_b$ and $f_e$ at the $t$th iteration, respectively. Then, $\theta_{e}^{t}$ can be updated by

$$\theta_{e}^{t+1} = \alpha \theta_{e}^{t-1} + (1 - \alpha) \theta_{b}^{t}$$

(8)

where $\alpha$ is a smoothing coefficient.
D. Filtering Strategy

The theoretical foundation of self-ensembling learning lies in the assumption that the ensemble model can yield more accurate predictions on those unlabeled data than the base model does. Under this circumstance, the predictions of the ensemble model can be regarded as the pseudo labels to assist the training of the base model, which in turn, helps to generate a more accurate ensemble model, achieving the self-ensembling mechanism. Thus, the quality of the predictions generated by \( f_e \) determines the performance of the whole framework. However, in practical applications, it is hard to ensure that \( f_e \) can always yield accurate pseudo labels for all unlabeled samples, and the inaccuracy of \( f_e \) may mislead the training of \( f_b \). To address this challenge, in this section, we propose the consistency filter to improve the robustness of self-ensembling learning. An illustration of the proposed filtering strategy is shown in Fig. 3.

Given an unlabeled sample \( x_{u(i)} \), we first conduct \( m \) times stochastic augmentations so that we can get \( m \) copies of \( x_{u(i)} \) under random augmentations (we set \( m = 5 \) in this study). Let \( x_{u(i)}^t \) be the \( t \)th augmented sample of \( x_{u(i)} \), where \( t = 1, 2, \ldots, m \). For each \( x_{u(i)}^t \), its corresponding probability vector \( \hat{p}^t = (\hat{p}_1^t, \hat{p}_2^t, \ldots, \hat{p}_j^t, \ldots, \hat{p}_N^t) \) can be generated by feeding \( x_{u(i)}^t \) into the ensemble network \( f_e \), where \( 0 \leq \hat{p}_j^t \leq 1 \) is the probability that \( x_{u(i)}^t \) belongs to the \( j \)th category. The consistency of the predictions on \( x_{u(i)} \) by \( f_e \) can be measured by the standard deviation of these \( m \) probability vectors \( [\hat{p}^1, \hat{p}^2, \ldots, \hat{p}^m] \)

\[
\text{cons} = -\sum_{j=1}^{k} \text{var}(\hat{p}_j^1, \hat{p}_j^2, \ldots, \hat{p}_j^m)
\]  

where \( \text{var}(\cdot) \) denotes the variance calculation. Note that there is a negative sign before the summation operation in (9). Thus, a larger standard deviation on the probability vector corresponds to a smaller cons value.

In the training phase, each batch of unlabeled samples can be sorted by the cons value in a descending order. Let \( L = [l_1, l_2, \ldots, l_b] \) be the sorted index list, where \( l_i \) denotes the index of the sample that has the \( i \)th largest cons value in the batch, and \( b \) is the batch size of the unlabeled samples.

The consistency \( F_{\text{cons}(i)} \) of \( x_{u(i)} \) can thereby be defined as

\[
F_{\text{cons}(i)} = \begin{cases} 
1, & \text{if } i \in [l_1, l_2, \ldots, l_q] \\
0, & \text{if } i \notin [l_1, l_2, \ldots, l_q] 
\end{cases}
\]  

where \( 1 \leq q \leq b \) is a threshold that controls how many samples can be filtered in the batch. Considering that the reliability of the ensemble model is poor at the beginning of the training, and would get improved as the training goes on, \( q \) is determined with a ramp-up function

\[
q = \text{int}\left(b \times e^{-(\text{iter} / \text{iter}_\text{max})^2}\right)
\]

where \( \text{iter} \) and \( \text{iter}_\text{max} \) denote the number of current iteration step and the number of total iteration steps, respectively. \( \text{int}(\cdot) \) is the rounding operation.

The design of the consistency filter lies in the intuition that the ensemble model should give consistent predictions for \( x_u \) regardless of the stochastic augmentations. Otherwise, the quality of the generated pseudo label may not be guaranteed.

With the constraint of (10), only samples with the top \( q \) largest cons values in the batch will remain, and the consistency loss in Section II-C can be modified as

\[
L_{\text{con}}(f_b, x_u) = F_{\text{cons}} \sum_{j=1}^{k} \left\| f_b(N(x_u))_{(j)} - \hat{f}_e(N(x_u))_{(j)} \right\|^2
\]

where \( \hat{f}_e(N(x_u)) = \text{mean}([\hat{p}^1, \hat{p}^2, \ldots, \hat{p}^m]) \) is the mean predictive probability vector of the ensemble model \( f_e \) on \( m \) copies of \( x_u \) under random augmentations.

The complete optimization procedure for the whole framework is shown in Algorithm 1.

III. EXPERIMENTS

In this section, we first introduce the datasets used in this study. Then, the experimental results are presented and analyzed in detail.

A. Data Descriptions

In this study, three benchmark hyperspectral datasets are utilized to evaluate the performance of the proposed method.
Algorithm 1 RSEN

Input: (1)
1) The labeled training set $X_l$ and $Y_l$, the unlabeled training set $X_u$, and the test set $X_t$.
2) A base network $f_b$, and an ensemble network $f_e$.
3) The smoothing coefficient $\alpha$ in the exponential moving average, and the number of total training iterations $T$.

Initialize the parameters in $f_b$ and $f_e$ with random Gaussian values.

for $\text{iter}$ in range$(0, T)$ do
3: Get mini-batch samples $x_l \in X_l, y_l \in Y_l, x_u \in X_u$.
4: Compute the consistency filter $F_{\text{cons}}$ via (9) and (10).
5: Compute the cross-entropy loss $L_{\text{cls}}(f_b, x_l, y_l)$ and consistency loss $L_{\text{con}}(f_b, x_u)$ via (5) and (12).
6: Update the parameters in $f_b$ by descending its stochastic gradients via $\nabla_{f_b}(L_{\text{cls}}(f_b, x_l, y_l) + L_{\text{con}}(f_b, x_u))$.
7: Update the parameters in $f_e$ by the exponential moving average via (8).

end for
9: Feed the samples in the test set $X_t$ to $f_e$ to accomplish the classification.

Output: The predictions on the test set $X_t$.

TABLE I
NUMBERS OF TRAINING AND TEST SAMPLES USED IN THE PAVIA UNIVERSITY DATASET

| Class number | Class name      | Training | Test   |
|--------------|----------------|----------|--------|
| 1            | Asphalt        | 30       | 6601   |
| 2            | Meadows        | 30       | 18619  |
| 3            | Gravel         | 30       | 2069   |
| 4            | Trees          | 30       | 3034   |
| 5            | Metal sheets   | 30       | 1315   |
| 6            | Bare soil      | 30       | 49999  |
| 7            | Bitumen        | 30       | 1300   |
| 8            | Bricks         | 30       | 3652   |
| 9            | Shadows        | 30       | 917    |
| Total        |                | 270      | 42506  |

The first dataset is acquired by the reflective optics system imaging spectrometer (ROSIS) sensor over the Pavia University, northern Italy. This image consists of 103 spectral bands with 610 by 340 pixels and it has a spectral coverage from 430 to 860 nm and a spatial resolution of 1.3 m. The false-color composition picture of the Pavia University dataset and the corresponding ground-truth map are shown in Fig. 4. The training and test sets are listed in Table I.

The second dataset is acquired by an ITRES-CASI 1500 sensor over the University of Houston campus and its neighboring urban area on June 23, 2012, between 17:37:10 and 17:39:50 UTC. The average altitude of the sensor is about 1676 m, which results in 2.5 m spatial resolution data consisting of 349 by 1905 pixels. The hyperspectral imagery consists of 144 spectral bands ranging from 380 to 1050 nm and is processed with radiometric correction, attitude processing, GPS processing, and geo-correction to yield the final hyperspectral cube. The false-color composition picture of the Houston dataset and the corresponding ground-truth map are shown in Fig. 5. The training and test sets are listed in Table II.

The third dataset is collected by the AVIRIS sensor over the Salinas Valley, California, with a spatial resolution of 3.7 m. After the removal of the water absorption bands, the image consists of 204 spectral bands ranging from 400 to 2500 nm with 512 by 217 pixels. It includes vegetables, bare soils, and vineyard fields. The false-color composite of the Salinas dataset and the corresponding ground-truth map are shown in Fig. 6. The training and test sets are listed in Table III.

B. Implementation Details

The proposed method is implemented with the PyTorch package. We use Adam optimizer to train the network with a learning rate of $5e-4$ and a batch size of 256. The training epochs are set as 20. Each batch consists of 128 labeled samples and 128 unlabeled samples. The number of principal components $p$ in dimension reduction is set as 5. The window size $w$ in the patch extraction is set as 16. We implement the stochastic augmentation $\mathcal{N}(\cdot)$ by adding Gaussian noise with
a mean of 0 and a standard deviation of 0.5 to each pixel in the image. The smoothing coefficient \( \alpha \) in (8) is set to 0.95.

To ensure the fairness of the experiments, we randomly repeat all methods reported in this study with 30 repetitions. In each repetition, we first generate the training set by randomly selecting 30 samples from each category in the reference data. Then, we use the remaining reference data to make up the test set. In this way, the test data would also vary in different repetitions. For the unlabeled training set used in the self-ensembling mechanism, we randomly select 10000 samples from the reference data without annotation. The overall accuracy (OA), kappa coefficient (\( \kappa \)), average accuracy (AA), and the producer accuracy for each class are utilized to quantitatively estimate different methods. Both the average value and the standard deviation are reported. The experiments are carried out with two Intel Xeon Silver 4114 2.20-GHz CPUs with 128 GB of RAM, and one NVIDIA GeForce RTX 2080 Ti GPU.

C. Classification Results

In this section, we report the classification results of the proposed method along with other methods. A brief introduction for each method included in the experiments is summarized below.

1) Raw: Classification with original spectral features via the RBF-SVM classifier.
2) PCA: Classification with the first ten principal components via the RBF-SVM classifier.
3) 2D-CNN: Spatial classification with 2D-CNN. We follow the implementations in [22]. For each pixel in the image, a \( 27 \times 27 \) patch is extracted on the first principal component as the input of the network.
4) SAE: Spectral–spatial classification with SAE. We follow the implementations in [18].
5) 3D-CNN: Spectral–spatial classification with 3D-CNN. We follow the implementations in [22]. For each pixel in the image, a \( 27 \times 27 \times n \) cube is extracted as the input of the network, where \( n \) is the number of bands.
6) SSFCN: Spectral–spatial classification with fully convolutional networks. We follow the implementations in [24].
7) SSUN: Spectral–spatial classification with RNN and CNN. We follow the implementations in [21].
8) SACNet: Spectral–spatial classification with self-attention learning and context encoding. We follow the implementations in [40].
9) SpectralFormer: Spectral–spatial classification by sequential learning from neighboring bands and patches using a transformer architecture. We follow the implementations in [41].
10) BaseNet: The proposed spectral–spatial network without using the self-ensembling mechanism.
11) RSEN: The proposed RSEN.

The quantitative results are reported in Tables IV–VI. In most cases, deep learning-based methods demonstrate their superiority compared with traditional spectral classification methods like Raw and PCA. Take the classification results of the Pavia University dataset, for example. While the Raw method can only achieve an OA of around 76%, SAE and 3D-CNN can obtain an OA of more than 82%. However,
since only 30 labeled samples for each class are utilized in the training, we can find that the superiority of the existing deep learning-based methods is limited in some cases. For example, it can be observed from Table V, while the Raw method can yield an OA of around 85% on the Houston dataset, SAE can only obtain an OA of around 73%, which is lower than the previous one with 12%. These results show that the insufficiency of the training samples may limit the capability of deep learning-based methods in some cases since there are so many parameters that need to be tuned in a deep neural network. By contrast, the proposed BaseNet can achieve competitive performance compared to the existing deep learning-based methods despite its simplicity. Take the results in the Pavia University dataset, for example. BaseNet can yield an OA of around 86%, which is comparable with the OA of SSUN. The main reason here lies in the simple, yet effective network architecture adopted in BaseNet, which enables the model to be lightweight so that the framework can be powerful even with limited training samples. With the help of the self-ensembling mechanism, the proposed RSEN...
can significantly improve the OA to around 94%, achieving the highest accuracy. Similar performances can be observed in the Houston dataset and Salinas dataset.

Although an ensemble model of multiple base models generally yields better predictions than a single model [29], [30], we find there also exist some failure cases in our experiments. Take the sixth category on the Pavia University dataset, for example. While BaseNet can achieve an accuracy of around 85% on this category, the proposed RSEN can only obtain an accuracy of around 80%, which is lower by 5% points. This phenomenon indicates that the ensemble model cannot always guarantee more accurate predictions compared to the base classifier. Nevertheless, since ensemble learning involves multiple base classifiers, its robustness is generally much better. While the standard deviation of the BaseNet on the sixth category on the Pavia University dataset is around 8%, RSEN can decrease the standard deviation to 5%. Similar phenomena can be observed in other datasets.

The time cost of each method is also reported in Tables IV–VI. It can be observed that deep learning-based methods generally cost more time compared with SVM-based methods. For all three datasets used in our experiments, 3D-CNN requires the longest running time, while the time cost of PCA is always the least. Compared with other deep learning-based methods, the proposed BaseNet is very efficient. For example, it only costs around 10 s to run the BaseNet in the Pavia University dataset, while SAE requires around 449 s, and 3D-CNN even requires around 832 s. Due to the high-computational burden of ensemble learning, the proposed RSEN costs much more time than the BaseNet. Nevertheless, compared with other existing deep learning-based methods, the runtime of RSEN is still competitive, owing to the efficiency of the proposed BaseNet.

To evaluate the qualitative results of these methods, the classification maps are presented in Figs. 4–6. It can be observed that the maps generated by spectral classification
methods like Raw and PCA generally contain more salt and pepper noise, while spectral-spatial classification methods like SSUN can effectively alleviate this phenomenon. However, due to the insufficiency of the training samples, existing state-of-the-art deep learning-based methods like 3D-CNN tend to misclassify the image, especially in the boundary regions. The proposed BaseNet also suffers from this problem, as we can see from highway regions in Fig. 5(j). Many building objects are misclassified as the highway category by BaseNet. By contrast, owing to the self-ensembling mechanism and the proposed filtering strategy, RSEN can yield much better classification maps, where both the regions inside an object and object boundaries are well interpreted.

We further compare the proposed RSEN with the state-of-the-art semisupervised methods, including LapSVM [42], LFDA [43], RLDE [43], Breaking Ties [44], K-means-PLCSU [44], BT-SVM [45], KNN-SNI [46], MLR-SNI [46], PL-SSDL (DPMM) [25], PL-SSDL (C-DPMM) [25], and SAE [18] in Table VII. The results of the comparing methods (except the LapSVM and SAE) are directly duplicated from the original articles.

Note that only ten labeled training samples per class are utilized for training in these experiments, which would be very challenging especially for deep learning-based methods. It can be observed from Table VII that the proposed RSEN achieves the highest accuracy in both OA and $\kappa$ metrics for all datasets.
In the Pavia University dataset, our method outperforms SAE by more than 12% in the OA metric. These results demonstrate that the proposed RSEN can yield good performance even with extremely limited labeled samples.

### D. Ablation Study and Parameter Analysis

To evaluate how each module in RSEN influences the classification performance, we further conduct an ablation study in Table VIII. Here, “SE” denotes the self-ensembling mechanism. $F_{cons}$ represents the proposed consistency filtering strategy. It can be observed that the self-ensembling mechanism can significantly improve the classification performance, especially for the Pavia University dataset (86.86%–91.81%). With the proposed filtering strategies, RSEN further improves the classification performance of self-ensembling learning, achieving the highest OAs for all three datasets.

The foundation of the self-ensembling learning lies in the pseudo labels generated by the ensemble network $f_e$. Thus, a natural idea is to investigate whether $f_e$ can yield more accurate predictions than the base network $f_b$. To this end, we plot the classification accuracy on the test set for both $f_e$ and $f_b$ at different training epochs. As shown in Fig. 7, it can be observed that at the early training stage, the performance of $f_e$ is close to $f_b$ or even worse than $f_b$ [e.g., Fig. 7(c)]. After some epochs, $f_e$ gradually outperforms $f_b$ and the orange curves are higher than the blue ones in Fig. 7 in most cases. Eventually, $f_b$ can also yield good performance and the OA gap between $f_b$ and $f_e$ is very little at the end of the training, which demonstrates the effectiveness of the self-ensembling mechanism.

One important factor that influences the performance of the proposed RSEN is the number of unlabeled samples used in the experiments. Here, we run the proposed RSEN with different numbers of unlabeled samples. For each dataset, we conduct 30 repetitions, and both the mean value ($\mu$) and the standard deviation value ($\sigma$) are plotted in Fig. 8. We can find that the OA of the proposed RSEN is relatively low if only 1000 unlabeled samples are utilized in the training for all datasets. As the number of unlabeled samples gets larger, the performance of the framework also gets better. To balance the accuracy and the computation burden, the number of unlabeled samples is empirically set as 10000 for all datasets in our experiments. The parameter $q$ is also an important parameter in the proposed method, which controls the number of samples filtered in the batch. Fig. 9 presents how different values of $q$ would influence the performance of the proposed method. Here, we use “ramp-up” to represent the ramp-up function described in (11). It can be observed that a larger $q$ generally helps to increase the performance, especially for
Fig. 9. Overall accuracies obtained by the proposed RSEN with different values of $q$.

Fig. 10. Data distributions of the reference samples in the original spectral feature space (first column), and the RSEN feature space (second column), respectively. (a) and (b) Pavia University dataset. (c) and (d) Houston dataset. (e) and (f) Salinas dataset.

IV. CONCLUSION

Deep learning is a powerful tool to address the HSI classification task. Nevertheless, training a deep neural network usually requires a large amount of labeled data, which is very hard to be collected. To tackle this challenge, in this study, we propose a RSEN which can yield good performance with very limited labeled samples. To the best of our knowledge, the proposed method is the first attempt to introduce the self-ensembling technique into the HSI classification task, which provides a different view on how to utilize the unlabeled data in HSI to assist the network training. We further propose a novel consistency filtering strategy to increase the robustness of self-ensembling learning. Extensive experiments on three benchmark HSI datasets demonstrate that the proposed method can yield competitive performance compared with the state-of-the-art methods.

Since the insufficiency of labeled data is a common challenge in many remote sensing tasks, whether the proposed robust self-ensembling mechanism can yield good performance on other remote sensing scenarios is also worth studying. We will try to explore this issue in future work.

TABLE IX

|           | 0.1  | 0.3  | 0.5  | 0.7  | 0.9  | 0.95 | 0.99 | 0.999 |
|-----------|------|------|------|------|------|------|------|-------|
| Pavia University | 93.56 | 93.89 | 94.13 | 94.37 | 94.82 | 94.65 | 91.53 | 87.14 |
| Houston    | 93.70 | 94.26 | 94.16 | 94.84 | 94.69 | 94.74 | 93.41 | 91.59 |
| Salinas    | 93.47 | 93.63 | 93.57 | 93.85 | 94.02 | 94.32 | 94.13 | 90.79 |

We further present the data distribution of the reference samples in the original spectral feature space and the RSEN feature space. The $t$-SNE is utilized to project the data from the original high-dimension space into a 2-D space [47]. It can be observed from Fig. 10 that in the original spectral space, samples from different categories are highly overlapped, while in the RSEN feature space, the inter-class distances get much larger and samples from different categories become more separable.

REFERENCES

[1] P. Zhong and R. Wang, “Jointly learning the hybrid CRF and MLR model for simultaneous denoising and classification of hyperspectral imagery,” IEEE Trans. Neural Netw. Learn. Syst., vol. 25, no. 7, pp. 1319–1334, Jul. 2014.
[2] S. Jia, Z. Lin, B. Deng, J. Zhu, and Q. Li, “Cascade superpixel regularized Gabor feature fusion for hyperspectral image classification,” IEEE Trans. Neural Netw. Learn. Syst., vol. 31, no. 5, pp. 1638–1652, May 2019.
[3] G. Hughes, “On the mean accuracy of statistical pattern recognizers,” IEEE Trans. Inf. Theory, vol. IT-14, no. 1, pp. 55–63, Jan. 1968.
[4] D. L. Donoho et al., “High-dimensional data analysis: The curses and blessings of dimensionality,” AMS Math Challenges Lect., vol. 1, p. 32, Aug. 2000.
[5] S. Yang, Z. Feng, M. Wang, and K. Zhang, “Self-paced learning-based probability subspace projection for hyperspectral image classification,” IEEE Trans. Neural Netw. Learn. Syst., vol. 30, no. 2, pp. 630–635, Jun. 2018.
[6] Q. Wang, J. Lin, and Y. Yuan, “Salient band selection for hyperspectral image classification via manifold ranking,” IEEE Trans. Neural Netw. Learn. Syst., vol. 27, no. 6, pp. 1279–1289, Jun. 2018.
[7] J. Peng, L. Li, and Y. Y. Tang, “Maximum likelihood estimation-based joint sparse representation for the classification of hyperspectral remote sensing images,” IEEE Trans. Neural Netw. Learn. Syst., vol. 30, no. 6, pp. 1790–1802, Jun. 2019.
Y. Xu, B. Du, and L. Zhang, “Beyond the patchwise classification: Spectral-spatial fully convolutional networks for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 391–406, Jan. 2018.

Y. Xu, B. Du, and L. Zhang, “Can we generate good samples for hyperspectral classification?—A generative adversarial network based method,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Valencia, Spain, Jul. 2018, pp. 5752–5755.

L. Zhu, Y. Chen, P. Ghamisi, and J. A. Benediktsson, “Generative adversarial networks for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5046–5063, Sep. 2018.

S. Singh, D. Hoiem, and D. Forsyth, “Swapout: Learning an ensemble of deep architectures,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 28–36.

B. Krawczyk, L. L. Minkub, J. Gama, J. Stefanowski, and M. Woźniak, “Ensemble learning for data stream analysis: A survey,” *Inf. Fusion*, vol. 37, pp. 132–156, Sep. 2017.

G. French, M. Mackiewicz, and M. Fisher, “Self-ensembling for visual domain adaptation,” in *Proc. ICLR*, 2018, pp. 1–5.

S. Laine and T. Aila, “Temporal ensembling for semi-supervised learning,” 2016, arXiv:1610.02247.

A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1195–1204.

Y. Xu, B. Du, L. Zhang, Q. Zhang, G. Wang, and L. Zhang, “Self-ensembling attention networks: Addressing domain shift for semantic segmentation,” in *Proc. AAAI Conf. Artif. Intel.*, vol. 33, pp. 5581–5588, Jul. 2019.

Y. Luo, P. Liu, L. Zheng, T. Guan, J. Yu, and Y. Yang, “Category-level adversarial adaptation for semantic segmentation using purified features,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 8, pp. 3940–3956, Aug. 2021.

Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, “Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 2507–2516.

Y. Luo, R. Ji, T. Guan, J. Yu, P. Liu, and Y. Yang, “Every node counts: Self-ensembling graph convolutional networks for semi-supervised learning,” *Pattern Recognit.*, vol. 106, Oct. 2020, Art. no. 107451.

J. Shi, T. Shao, X. Liu, X. Zhang, Z. Zhang, and Y. Lei, “Evolutionary multitask ensemble learning model for hyperspectral image classification,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 936–950, 2021.

K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

Y. Xu, B. Du, and L. Zhang, “Self-attention context network: Addressing the threat of adversarial attacks for hyperspectral image classification,” *IEEE Trans. Image Process.*, vol. 30, pp. 8671–8685, 2021.

D. Hong et al., “SpectralFormer: Rethinking hyperspectral image classification with transformers,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

L. Gómez-Chova, G. Camps-Valls, J. Muñoz-Mari, and J. Calpe, “Semi-supervised image classification with Laplacian support vector machines,” *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 3, pp. 336–340, Jul. 2008.

D. Ou, K. Tan, Q. Du, J. Zhu, X. Wang, and Y. Chen, “A novel tri-training technique for the semi-supervised classification of hyperspectral images based on regularized local discriminant embedding feature extraction,” *Remote Sens.*, vol. 11, no. 6, p. 654, Mar. 2019.

I. Dópido, J. Li, P. Gamba, and A. Plaza, “A new hybrid strategy combining semisupervised classification and unmixing of hyperspectral data,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 8, pp. 3619–3629, Aug. 2014.

I. Dópido, J. Li, P. Marpa, A. Plaza, J. Dias, and J. A. Benediktsson, “Semisupervised self-learning for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 7, pp. 4032–4044, Jan. 2013.

K. Tan, J. Hu, J. Li, and P. Du, “A novel semi-supervised hyperspectral image classification approach based on spatial neighborhood information and classifier combination,” *ISPRS J. Photogramm. Remote Sens.*, vol. 105, pp. 19–29, Jul. 2015.

L. van der Maaten and G. Hinton, “Visualizing data using t-SNE,” *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
Bo Du (Senior Member, IEEE) received the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2010.

He is currently a Professor with the School of Computer Science and the Institute of Artificial Intelligence, Wuhan University, where he is also the Director of the National Engineering Research Center for Multimedia Software. He has published more than 80 research articles in the IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), the IEEE TRANSACTIONS ON CYBERNETICS (TCYB), the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS), the IEEE JOURNAL OF SELECTED TOPICS IN EARTH OBSERVATIONS AND APPLIED REMOTE SENSING (JSTARS), and the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS (GRSL). Thirteen of them are ESI hot papers or highly cited papers.

His major research interests include pattern recognition, HSI processing, and signal processing.

Dr. Du regularly serves as a Senior Program Committee (PC) Member for the International Joint Conferences on Artificial Intelligence (IJCAI) and the Association for the Advancement of Artificial Intelligence (AAAI). He received the Highly Cited Researcher by the Web of Science Group in 2019 and 2020, the IEEE Geoscience and Remote Sensing Society (GRSS) 2020 Transactions Prize Paper Award, the IJCAI Distinguished Paper Prize, and the IEEE Workshop on Hyperspectral Image and Signal Processing Best Paper Award in 2018. He was the IEEE Data Fusion Contest Champion. He has served as the Area Chair for the International Conference on Pattern Recognition (ICPR). He also serves as a Reviewer for 20 Science Citation Index (SCI) magazines, including the IEEE TPAMI, IEEE TCYB, IEEE TGRS, IEEE TIP, IEEE JSTARS, and IEEE GRSL. He also serves as an Associate Editor for Neural Networks, Pattern Recognition, and Neurocomputing.

Liangpei Zhang (Fellow, IEEE) received the B.S. degree in physics from Hunan Normal University, Changsha, China, in 1982, the M.S. degree in optics from the Xi’an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi’an, China, in 1988, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1998.

He was a Principal Scientist with the China State Key Basic Research Project from 2011 to 2016 appointed by the Ministry of National Science and Technology of China to lead the Remote Sensing Program in China. He is currently a Chair Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing (LIESMARS), Wuhan University. He has published more than 700 research articles and five books. He also holds 30 patents. He is the Institut for Scientific Information (ISI) highly cited author. His research interests include hyperspectral remote sensing, high-resolution remote sensing, image processing, and artificial intelligence.

Dr. Zhang is a fellow of the Institute of Electrical and Electronic Engineers (IEEE) and the Institution of Engineering and Technology (IET). He was a recipient of the 2010 Best Paper Boeing Award, the 2013 Best Paper ERDAS Award from the American Society of Photogrammetry and Remote Sensing (ASPRS), and the 2016 Best Paper Theoretical Innovation Award from the International Society for Optics and Photonics (SPIE). His research teams won the top three prizes of the IEEE GRSS 2014 Data Fusion Contest, and his students have been selected as the winners or finalists of the IEEE Geoscience and Remote Sensing Symposium (IGARSS) Student Paper Contest in recent years. He is the Founding Chair of the IEEE Geoscience and Remote Sensing Society (GRSS) Wuhan Chapter. He also serves as an associate editor or editor for more than ten international journals.

He is currently serving as an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.