


# Spectral–Spatial Attention Feature Extraction for Hyperspectral Image Classification Based on Generative Adversarial Network

Hongbo Liang , Wenxing Bao , Xiangfei Shen , and Xiaowu Zhang

**Abstract**—Recent research shows that generative adversarial network (GAN) based deep learning derived frameworks can improve the accuracy of hyperspectral image (HSI) classification on limited labeled samples. However, several studies point out that existing GAN-based methods are heavily affected by the complexity and inefficient description issues of HSIs. The discriminator in GAN always attempts to interpret high-dimensional nonlinear spectral knowledge of HSIs, thus resulting in the Hughes phenomenon. Another critical issue is sample generation. The generator is only used as a regularizer for the discriminator, which seriously restricts the performance for classification. In this article, we propose SSAT-GAN, a semisupervised spectral–spatial attention feature extraction approach based on the GAN that feeds raw data into a deep learning framework, in an end-to-end fashion. First, the unlabeled data is added into the discriminator to alleviate the problems of training samples and supplies a reconstructed real HSI data distribution through adversarial training. Second, to enhance the description of HSIs, we build spectral–spatial attention modules (SSAT) and extend them to the discriminator and the generator to extract discriminative characteristics from abundant spatial contexts and spectral signatures. The SSAT modules learn a three-dimensional filter bank with spectral–spatial attention weights to obtain meaningful feature maps to improve the discrimination of the feature representation. In terms of the mode collapse of GANs, the mean minimization loss is employed for unsupervised learning. Experimental results from three real datasets indicate that SSAT-GAN has certain advantages over the state-of-the-art methods.

**Index Terms**—Attention module, generative adversarial network (GAN), hyperspectral image (HSI) classification, semisupervised deep learning, spectral–spatial information.

## I. INTRODUCTION

**H**YPERSPECTRAL imagery (HSI) obtains hundreds of numerous narrow and contiguous spectral bands from

Manuscript received July 3, 2021; revised September 2, 2021; accepted September 23, 2021. Date of publication September 28, 2021; date of current version October 14, 2021. This work was supported in part by the Natural Science Foundation of Ningxia Province of China under Grant 2020AAC02028 and in part by the Natural Science Foundation of Ningxia Province of China under Grant 2021AAC03179. (Corresponding author: Wenxing Bao.)

Hongbo Liang, Wenxing Bao, and Xiaowu Zhang are with the School of Computer Science and Engineering, North Minzu University, Yinchuan 7500021, China, and also with the Key Laboratory of Images, and Graphics Intelligent Processing of State Ethnic Affairs Commission: IGIPLab, North Minzu University, Yinchuan 750021, China (e-mail: 876502548blue@gmail.com; bwx71@163.com; 2012033@nun.edu.cn).

Xiangfei Shen is with the School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China (e-mail: xfshen95@outlook.com).

Digital Object Identifier 10.1109/JSTARS.2021.3115971

the surface which provide abundant characteristics to enhance the identification ability of ground materials [1]. With high-resolution imaging technology rapidly developing, HSI becomes an ideal tool to effectively detect the surface, which spans a broad range of applications, including mineral substance [2], monitoring of plant diseases [3], anomaly detection [4], and land-cover mapping [5]. HSI classification plays a substantial role in these fields, intending to analyze discriminative characteristics of HSI and classify each pixel according to a corresponding land-cover category [6]. Therefore, two major characteristics of HSI should be considered. First, the high-dimensional nonlinear spectral signature, which originates from redundant bands of spectrums, enables the accurate distinction of homologous surface categories. Second, high spatial correlation provides spatial auxiliary contexts for accurate mapping of pixelwise classification, which derives from homogeneous regions [7].

Since the spectral information can natively reflect the characteristics of different materials, one set of traditional methods identifies the classification maps in a pixelwise way, which can be divided into two steps: 1) feature engineering, such as principal component analysis (PCA) [8], bands selection [9] and 2) classifier development, including support vector machine (SVM) [10], random forest [11]. This kind of approach is constrained by the high-dimensional nonlinear characteristics, which leads to an unsatisfactory result. To further improve the representation of HSIs, another set of approaches implements the positive effect on the spectral–spatial expression. Existing methods introduced the spatial contexts in the feature engineering step. For instance, Kang *et al.* [12] proposed the feature fusion framework combined with the edge-preserving filtering (EPF) and SVM. Jiang *et al.* [13] regarded the superpixel as a carrier to extract potential features. However, the models mentioned above consist of shallow structures which cannot provide an efficient description.

With the advancement of artificial intelligence, CNN-based approaches have attracted increased focus due to the fact that their objective functions directly aim at classification instead of two independent steps to obtain remarkable results [14], [15]. In 2016, Zhao *et al.* [16] adopted CNN to learn local spatial contexts for HSI classification. Chen *et al.* [17] designed a 3-D CNN to extract neighboring spectral cubes, which originate from HSIs instead of dimensionality-reduced data. Nonetheless, a deeper network may lead to the Hughes phenomenon, under the

conditions of both complexity of the spectral–spatial distribution and the scarcity of training samples.

Meanwhile, with the development of deep learning, a series of deep-learning-derived methods have been applied for HSI classification and proven to be successful. Many works of classification frameworks obtains superior achievements by constructing high efficiency spectral–spatial feature extraction. For instance, Zhong *et al.* [18] built a spectral–spatial residual network (SSRN) to reduce the complexity of the network design and achieved advanced performance. In [19], a dense convolutional block was employed for accurate identification. A 3D-Conv-Capsule model [20] was presented for HSI classification, which attempted to consider the pixel position attributes to enhance the spatial awareness. In addition, in Sellami's work [21], a spectral–spatial graph was constructed to fully exploit the inherent spatial distribution.

Another line of approaches accomplished spectral–spatial classification by exploiting attention mechanisms, which performs classification after aggregating features from the homogeneous regions. Xu *et al.* [22] designed a control gate attention mechanism for the quick acquisition of key features. In [23], a spectral–spatial classification framework was proposed by performing CNN with a self-attention module to enhance the correlation of features. In [24], a multiattention fusion network (MAFN) was designed to mine significant features for classification. Yu *et al.* [25] presented a dense CNN framework with a feedback attention mechanism to further improve the computation efficiency. However, the attention weight embedding was placed behind the spectral–spatial representation, which introduced the influence of interference pixels and redundant spectral bands. He *et al.* [26] designed an HSI-BERT to capture global dependence among pixels at the receptive field. However, the transformer-based method needs multiple nonlocal areas to capture global long-term dependence.

In contrast to classical optical image classification objectives in the computer vision fields, which consist of hundreds of categories, the land cover classification of HSI takes much fewer targets for identification. Therefore, the theory that deep learning takes a high amount of data for training might not apply to HSIs which lack in labeled samples. Several works focus on the semisupervised learning via both labeled and unlabeled HSI samples for training. For instance, Fang *et al.* [27] presented a resampling strategy for training CNN sufficiently. In [28], the uncertainty of unlabeled samples of HSIs are considered for classification. Although these studies have acquired significant results, they may stem from the regions of high spatial correlation context, instead of deep learning methods.

Recently, generative adversarial network (GAN) have been applied for HSI classification to alleviate the issue of limited labeled samples. Specifically, GAN-based classifiers start from semisupervised HS-GAN proposed by Zhan *et al.* [29], which used 1-D spectral vectors as the input. To exploit the benefit of spatial information, a neighborhood majority voting strategy [30] is applied to the prediction, lately. He *et al.* [31] built a 3-D bilateral filtering-based GAN framework to improve the ability of spatial awareness. A 3D-GAN is proposed for HSI classification that keeps only the first three principal components

of raw data as input. In [33], a semisupervised GAN with a conditional random field (GAN-CRF) was designed that regards the softmax prediction as conditional probabilities of HSI to refine classification maps. To enhance the meaningful semantic contexts, an adaptive DropBlock-enhanced GAN (AD-GAN) [34] was established to stabilize the training state of the model.

Although these GAN-based methods have achieved satisfying ability over the contemporaneous benchmarks, there are still two drawbacks over HSI classification to be solved.

The first challenge is the mode collapse of GAN. The generator  $G$  deceives the discriminator  $D$  through generating data from the limited labeled data distribution [35]. The restricted narrow redundant spectral signatures limit the representation ability of GAN and lead to terrible data generation. In Wang's work [34], an adaptive DropBlock is employed as a regularization method to alleviate the mode collapse. However, the supervised GANs generate the data distribution that is similar to that of labeled training ones and, thus, difficult to learn the complete real HSI distribution. In addition, the unlabeled data of HSI remains an unexploited gold mine for efficient data utilization. Recently, in response to this characteristic, Liang *et al.* [36] implemented the mean minimization loss that considers the constraint over unlabeled data of HSI and acquired superior achievement. The reason for this phenomenon is that it may minimize the values and variances of high-dimensional feature maps from  $D$ . As this point, the GAN model can hardly be subject to the impact of complex parameter calculation, which guaranteed the stability of the training state.

Another critical issue is the complexity, inefficient description of spectral–spatial characteristics. The classification performance seems to deteriorate when the extraction of spectral–spatial characteristics is affected by interference pixels. Therefore, it is hard to guarantee that the GAN always works toward the authentic HSI distribution, particularly for high-dimensional spectral signature or texture-dependent context. In Feng's work [37], the joint spatial spectral hard attention mechanism was employed in  $G$  to cooperate  $D$  discards misleading and confounding information for HSI classification. However, it only focused on a specific area of the input patches in one batch, which requires more complex technology for training. In a disparate line of work, the attention-aware block [38] was designed in ResNets to enhance the representation of HSI data. It demonstrated that the attention-aware block can learn more valuable and valid representations. However, when dealing with objects with variable spectral or irregular areas, the attentive architecture is inefficient. We argue that if the homogeneous spectrum and adaptive receptive fields are taken into account, the complexity issue of the HSI data can be alleviated.

To tackle the above-mentioned challenges of GAN-based methods, we suggest a spectral–spatial attention feature extraction approach based on GANs (SSAT-GANs) for HSI classification. The purpose of the proposal can build a significant representation for spectral–spatial characteristics and enhance the robustness and stability of GANs in the way of semisupervised learning. On the one hand, the SSAT-GAN takes the unlabeled data into account to alleviate the scarcity of labeled samples, which enables the generator  $G$  to implicitly reconstruct real HSI

cubes. Meanwhile, we adopt the mean minimization loss as an unsupervised constraint item used in the discriminator  $D$  to avoid overfitting. On the other hand, the complicated spectral–spatial characteristics of local adjacent pixels herald the redundancy and inefficiency problem, which result in more insufficient classification with more complex regions. Inspired by the fact that the attention weights can enhance the effective representation of the saliency neighborhood of an object, the spectral–spatial attention modules (SSAT) are designed separately to capture the discriminative representation in this article, in which both intraspectrum and contextual relations of HSIs participate in the attention calculation through the feedback, and the weighted feature maps are considered to enhance intraclass consistency. In this way, we extend the SSAT to consecutive feature spreading and generation blocks and pass through them to build  $D$  and  $G$ , respectively. Unlike traditional semisupervised GANs, which require a deeper convolutional architecture for feature representation, our proposal is feature-efficient because both  $D$  and  $G$  share the weights of parameters with the corresponding attention modules and further improve the feature description. To this end, the well-trained  $D$  can achieve satisfactory classification accuracy.

The main contributions of this article are listed as follows.

- 1) We design a novel semisupervised GAN-based HSI classification framework using a small number of labeled and unlabeled data for training. The mean minimization loss is employed for unsupervised learning, which boost the backpropagation of the gradient and stabilize the training of GAN.
- 2) For the purpose of alleviating the inefficient description, we integrate the spectral–spatial attributes into SSAT for representation discrimination of the HSI data.
- 3) The alternately optimized architecture design makes the SSAT-GAN a framework that generalizes well in three real HSI datasets and achieves satisfactory classification accuracy over state-of-the-art methods.

The rest of this article is organized as follows. Section II reviews the basic concepts of GANs. The scheme of the proposed SSAT-GAN and its components are introduced in Section III. Experimental results and analysis are presented in Section IV. The superiority of SSAT-GAN is discussed in Section V. Finally, the conclusion is drawn in Section VI.

## II. RELATED WORK

### A. Generative Adversarial Network

GAN is an unsupervised deep learning model proposed by Goodfellow *et al.* [39], which provides a reasonable scheme to implicitly reckon real data distribution. GAN incorporates a generator  $G$  and a discriminator  $D$  in a unified network, where  $G$  generates samples to fool  $D$  into believing it, and  $D$  distinguishes the genuineness of the samples. Contradictory results make  $G$  and  $D$  reach Nash equilibrium in the zero-sum game, which is finally expressed as a minimax optimization problem

$$\min_G \max_D \text{Loss} = E_{z \sim p_z} [\log(1 - D(G(z)))] + E_{x \sim p_{\text{data}}} [\log D(x)] \quad (1)$$

where  $z \sim p_z$  and  $x \sim p_{\text{data}}$  denote the random noise vectors and input images following real data distribution, respectively.  $E(\cdot)$  is the expectation.  $D(x)$  and  $G(z)$  represent the sigmoid output obtained from  $D$  by training on real input vectors, and synthetic data from  $G$  by random noise, respectively.  $D(G(z))$  gives the real expectations of  $D$  with the input derives from  $G(z)$ .

In the optimization process of GAN,  $G$  and  $D$  are optimized alternately. Given  $G(z)$  of  $G$ , the model will optimize  $D$  by maximizing  $E_{x \sim p_{\text{data}}} [\log D(x)] + E_{z \sim p_z} [\log(1 - D(G(z)))]$ . When  $D$  arrives at a stationary score,  $G$  is optimized by minimizing  $E_{z \sim p_z} [\log(1 - D(G(z)))]$ . Since  $D$  and  $G$  achieve the Nash equilibrium during adversarial training, GAN will learn the probability estimation of real data and produce promising results.

## III. SSAT-GAN FRAMEWORK

The SSAT-GAN flowchart is shown in Fig. 1. Suppose the raw HSI dataset  $X$  contains  $m$  pixels  $\{x_1, x_2, x_3, \dots, x_m\} \in \mathbb{R}^{1 \times 1 \times b}$ , where  $b$  is the bands of spectrum. The neighboring cubes centered at the labeled pixels form the labeled datasets  $X^1 = \{x_i^1\} \in \mathbb{R}^{w \times w \times b \times m_l}$ . Take unlabeled cubes  $X^2 = \{x_i^2\} \in \mathbb{R}^{w \times w \times b \times m_u}$ , where  $w$ ,  $m_l$ , and  $m_u$  are the spatial size of HSI cubes and the number of labeled and unlabeled HSI samples, respectively. We send these two datasets to the discriminator to learn the real distribution of HSI. The generator synthesizes HSI cube  $Z = \{z_1, z_2, z_3, \dots, z_m\}$ , with samples of size  $X^2$ . In addition, the labeled  $X^1$  has its corresponding annotation  $Y^1 = \{y_i^1\} \in \mathbb{R}^{(1+n_y) \times m_l}$ , where  $n_y$  is the number of land cover categories, and  $y_i^1[0]$  is the first item of  $y_i^1$ , which indicates the authenticity of the corresponding HSI cube. The classified prediction of HSI is carried out with a well-trained discriminator.

SSAT-GAN incorporates the spectral and spatial attention modules in both discriminator and generator to extract discriminative features, where the discriminator and generator are, respectively, composed of convolutional and transposed layers.

### A. Spectral and Spatial Attention Modules

The purpose of SSAT is to enhance the feature analysis of a salient and effective domain, which is inspired by CBAM [40]. Given an intermediate feature map, it sequentially calculates attention weights along spectral and spatial dimensions, separately.

1) *Spectral Attention Module*: The spectral attention module aims at exploring the intraclass consistency of spectrums. As each band of spectral energy is considered as a class feature detector, our spectral attention focuses on “which” bands are meaningful given an input cube. To highlight discriminative signature from spectral knowledge while retaining uniform characteristics, we use both depthwise separable convolution (Depth\_CONV) [41] and 3-D convolution (3D\_CONV) operations. For aggregating intraspectrum information, average-pooling and max-pooling have been commonly adopted so far. In addition, we employ a spectral squeeze mechanism to assign an independent weight to each element along spectral dimension. Finally, the spectral attention weights will be generated by a dynamic activation function.

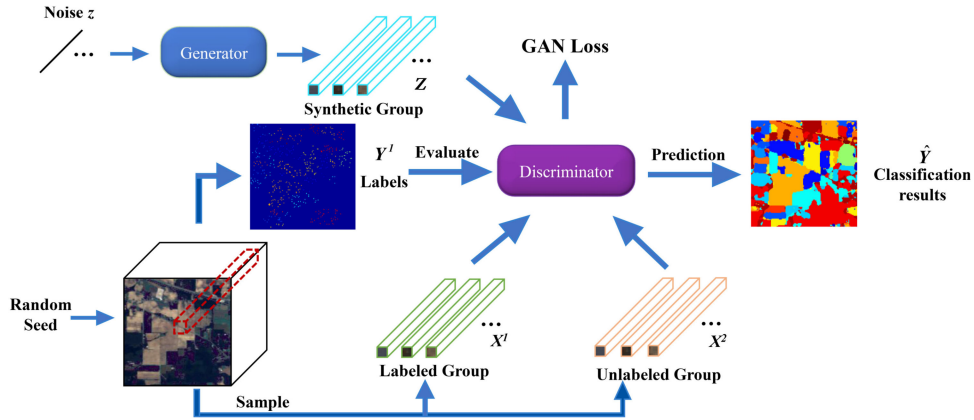


Fig. 1. Flowchart of SSAT-GAN framework for HSI classification. First, the unlabeled group  $X^2$  is established to initialize the parameters of a discriminator, and a generator transforms the noise vectors  $z$  to a set of fake HSI cubes  $Z$ , which implicitly learns the real HSI distribution. Then, the discriminator attempts to identify the authenticity of the input HSI cubes that derive from  $X^2$  or  $Z$ . Finally, the categorical information  $\hat{Y}$  is predicted by the discriminator that feeds labeled  $X^1$  during training. The corresponding annotation  $Y^1$  is adopted for the evaluation and acquire supervised partial loss of the GAN.

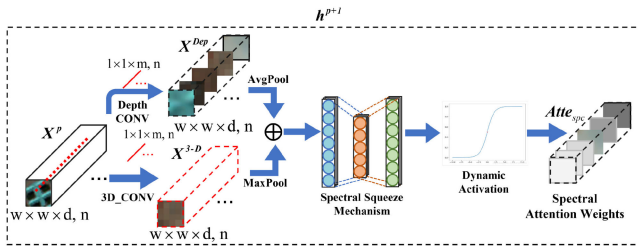


Fig. 2. Spectral attention module utilizes both Depth\_CONV and 3D\_CONV descriptors with pooling operations, followed by a spectral squeeze mechanism to predict spectral attention weights.

As demonstrated in Fig. 2, take  $n$  HSI cubes  $X^p$  of size  $w \times w \times d$  as the  $p + 1$ th input feature map. It first captures the available homogeneous area using both Depth\_CONV and 3D\_CONV operations with  $n$  spectral kernels of size  $1 \times 1 \times m$ , generating two different spatial context descriptors:  $X^{\text{Dep}}$  and  $X^{3-D}$ . Both descriptors are then forwarded to average-pooling and max-pooling operations, which denote the salient and effective features, respectively. After an *elementwise addition* strategy, the feature vectors are passed through a squeeze mechanism to extract spectral energy relationship, producing our spectral attention weights  $\text{Atte}_{\text{spc}}$ . It can enlarge the weights of HSI pixels with discriminative signatures in the spectral distribution and suppress those of adverse pixels for identification. The squeeze mechanism is composed of fully connected layers (FCs) with one embedding layer. To optimize the parameter efficiency, the embedding units are set to  $\mathbb{N}^{d/1 \times 1 \times r}$ , where the  $r$  is the optimization ratio. The  $\text{Atte}_{\text{spc}}$  can be formulated as

$$\begin{aligned} \text{Atte}_{\text{spc}} &= \sigma \left( \text{FCs} \left( \text{AvgPool} \left( X^{\text{Dep}} \right) \right. \right. \\ &\quad \left. \left. + \text{MaxPool} \left( X^{3-D} \right) \right) \right) \\ &= \sigma \left( \mathbf{W}_1 \left( \mathbf{W}_0 \left( X_{\text{avg}}^{\text{Dep}} + X_{\text{max}}^{3-D} \right) \right) \right), \end{aligned} \quad (2)$$

where  $\sigma(\cdot)$  is a sigmoid activation, which constrains the probabilities in the range of  $[0, 1]$ .  $\mathbf{W}_0 \in \mathbb{N}^{d/r \times d}$  and  $\mathbf{W}_1 \in \mathbb{N}^{d \times d/r}$  note that the FCs squeeze weights along spectral dimension.

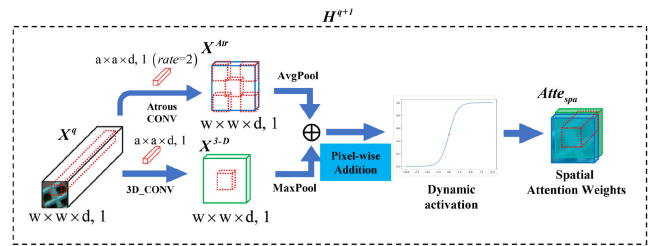


Fig. 3. Spatial attention module combines two similar feature maps that are convolved with Atrous\_CONV and 3D\_CONV, and pooled along the spectral axis, and then feeds them to a convolution layer.

It can be considered as the signal-to-noise ratio (SNR) enhancement from the physical level, that is, the ratio of the validity spectral energy considered as signal energy specified by  $\mathbf{W}_1$  to the squeezed features considered as noise energy  $\mathbf{W}_0$ .

2) *Spatial Attention Module*: For exploiting the interclass differences of spatial contexts, we build a spatial attention module to generate spatial attention map. As a reasonable theory, a pixel does not always share the category of its neighbors, and the spatial attention providing “where” are interesting areas, which is complementary to the spectral attention. As illustrated in Fig. 3, we apply the atrous convolution (Atrous\_CONV) and 3D\_CONV operations along the spectral axis, generating two different intermediate feature maps, which aim at extending receptive field and reducing the interference of abnormal pixels. Atrous\_CONV has been proved effectively for learning intraclass consistency homogeneous areas of HSIs [42], which has a shared kernel with multiple dilations for learning spatial contexts. For the “gridding problem” [43], both intermediate maps are then forwarded to average-pooling and max-pooling to enhance the spatial representation and generate a contextual descriptor with *pixelwise addition* strategy. The spatial attention weights will be predicted after activating the neural parameters of the contextual descriptor.

Supposing an HSI cube  $X^q$  of size  $w \times w \times d$  is the  $q + 1$ th input of the spatial attention module, we first aggregate spectral

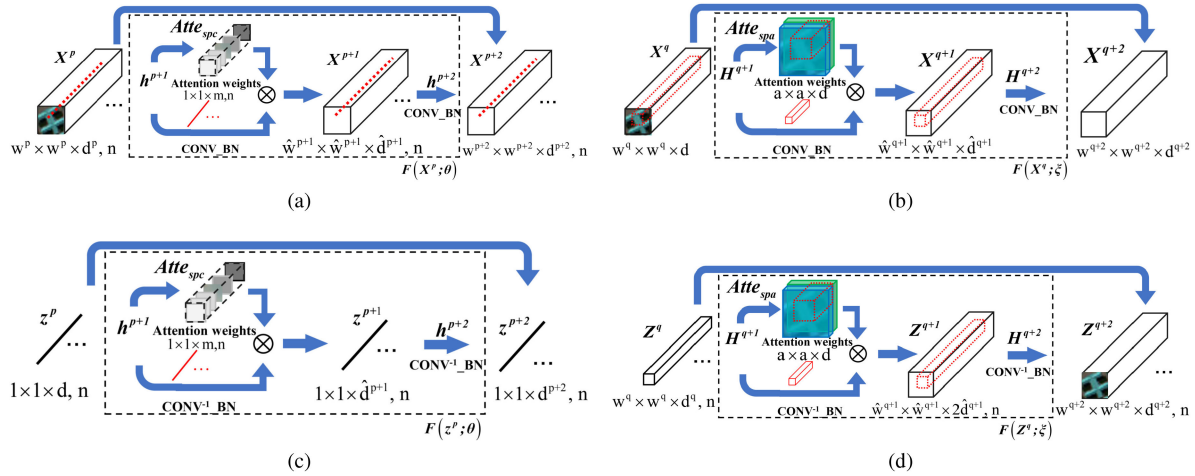


Fig. 4. Four feature spreading blocks with lightweight spectral–spatial attention modules aiming for HSI feature extraction and generation in SSAT-GAN. (a) and (b) Spectral and spatial attention feature spread blocks in discriminators; (c) and (d) Spectral and spatial attention feature generation blocks in generators.

information by two convolutional operations with kernel size  $a \times a \times d$ , generating  $X^{\text{Attr}}$  and  $X^{\text{3-D}}$ . Then they passed through both pooling operations and generate two maps:  $X_{\text{avg}}^{\text{Attr}}$  and  $X_{\text{max}}^{\text{3-D}}$ . Each denotes local effective information and uniform contexts across the spectral knowledge. The spatial attention weights  $\text{Atte}_{\text{spa}}$  are then predicted by a standard convolution after the *pixelwise addition*, which is computed as

$$\begin{aligned} \text{Atte}_{\text{spa}} &= \sigma \left( \left[ \text{AvgPool} \left( X^{\text{Attr}} \right); \right. \right. \\ &\quad \left. \left. \text{MaxPool} \left( X^{\text{3-D}} \right) \right] \right) * H^{q+1} + b^{q+1} \\ &= \sigma \left( \left[ X_{\text{avg}}^{\text{Attr}}, X_{\text{max}}^{\text{3-D}} \right] \right) * H^{q+1} + b^{q+1}. \end{aligned} \quad (3)$$

where  $*$  denotes the convolutional operation and  $H^{q+1} \in \mathbb{R}^{a \times a \times 2}$  is the spatial convolutional kernel, in which  $a$  denotes the spatial sampling size and  $b^{q+1}$  denotes the bias. Note that their spatial sizes are fixed at  $w \times w$  under the padding strategy, which means that the spatial attention module can explore the adaptive neighboring correlation at the dilated receptive regions. Therefore, the spatial attention module can provide supplementary information for accurate spectral feature mapping.

### B. Spectral–Spatial Attention Discriminator and Generator

We incorporate our SSAT in the generator and discriminator and extend them to four spectral–spatial attention spread learning and generation blocks. Fig. 4 shows the architecture of four attention blocks, each of which can be regarded as an extension of successive convolution and transposed convolution.

1) *Spectral Attention Feature Spread Block*: For the redundant spectral bands, as shown in Fig. 4(a), the spectral attention module is introduced in the  $p+1$ th layer to assign an attention weight to the spectral tensor of HSI and aggregates the intra-class correlation of the narrow spectrum. Next, the  $p+2$ th layer utilizes a 3-D convolution layer with batch normalization [44] (CONV\_BN) to update the parameters according to the spectral attention feature. The skip connection is applied instead of directly mapping between the  $p+1$ th and the  $p+2$ th layers and builds the spectral attention feature extraction function  $F(X^p; \theta)$ .

If  $X^p$  and  $X^{p+1}$ , respectively, represent the input intermediate feature cube of the  $p$ th layer and the output feature cube of the  $p+1$ th spectral convolutional layer, then the architecture of  $F(X^p; \theta)$  can be formulated as

$$X^{p+2} = X^p + F(X^p; \theta), \quad (4)$$

$$F(X^p; \theta) = \left( \text{Atte}_{\text{spc}} \otimes \text{R} \left( \hat{X}^{p+1} \right) \right) * h^{p+2} + b^{p+2}, \quad (5)$$

$$X = \text{R} \left( \hat{X}^p \right) * h^{p+1} + b^{p+1}, \quad (6)$$

$$\hat{X}^p = \frac{X^p - E(X^p)}{\text{Var}(X^p)} \quad (7)$$

where  $\theta = \{h^{p+1}, h^{p+2}, b^{p+1}, b^{p+2}\} \in \mathbb{R}^{1 \times 1 \times m, n}$ . Note that  $\theta$  is the weights and biases of the spectral convolutional kernels, which sharing their parameters for the whole training.  $\text{Atte}_{\text{spc}}$  is the spectral attention weights proposed by (2).  $\text{R}(\cdot)$  is the ReLU activation function, which sets negative values to zero.  $E(\cdot)$  and  $\text{Var}(\cdot)$  indicate the expectation and variance functions of the input HSI cubes, which is applied in BN, respectively.  $*$  represents the convolution operation, and  $\otimes$  is the elementwise multiplication. Furthermore,  $\text{Atte}_{\text{spc}}$  retains the weights among spatial dimensions to the same, under aggregating intraspectrum information to improve radiant energy efficiency from each band.

2) *Spatial Attention Feature Spread Block*: The spatial attention feature spread block aims to explore neighboring correlation and intraclass consistency of central pixels in high-spatial regions. Fig. 4(b) shows the detail of the spatial block. The depths along the spectrum of kernels are in an identical size with that of the input cubes  $X^q$ , which means the block extracts adaptive spatial context while maintaining the spectral attention feature. The architecture of the block can be formulated as

$$X^{q+2} = X^q + F(X^q; \xi), \quad (8)$$

$$F(X^q; \xi) = \left( \text{Atte}_{\text{spa}} \otimes \text{R} \left( \hat{X}^{q+1} \right) \right) * H^{q+2} + b^{q+2}, \quad (9)$$

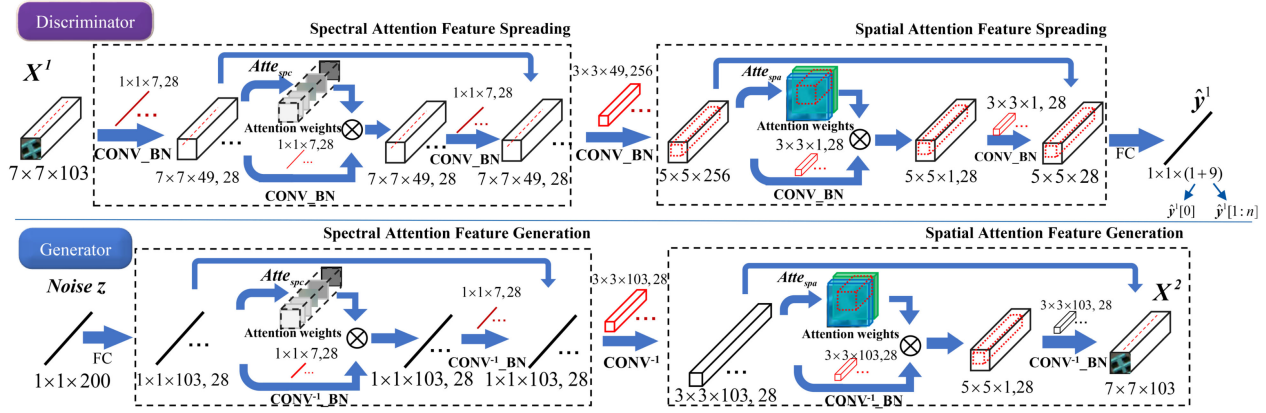


Fig. 5. Spectral–spatial discriminator (top), which contains successive spectral and spatial attention feature spread blocks and outputs a vector consisting of an indicative entry of real or fake data and categorical probabilities; spectral–spatial generator (bottom), which contains successive spectral and spatial attention feature generation blocks and transforms a vector from random noise to a synthetic HSI cube.

$$X = \mathbf{R} \left( \hat{X}^q \right) * H^{q+1} + b^{q+1}, \quad (10)$$

$$\hat{X}^q = \frac{X^q - E(X^q)}{\text{Var}(X^q)} \quad (11)$$

where  $\xi = \{H^{q+1}, H^{q+2}, b^{q+1}, b^{q+2}\} \in \mathbb{R}^{a \times a \times d, 1}$ .  $\mathbf{Atte}_{\text{spa}}$  is the spectral attention weights proposed by (3). In contrast to spectral attention, spatial attention can be also regarded as image denoising widely used in computer vision, that is,  $\mathbf{Atte}_{\text{spa}}$  searches an adaptive relationship from local spatial, and feedback to the input feature tensor.

3) *Spectral–Spatial Attention Feature Generation Blocks*: To overcome the challenge of a small-sample scenario, the idea of spectral–spatial attention is extended to the generator to improve the variety of generation. Fig. 4(c) and (d) shows the details of spectral–spatial attention generation blocks; they embed both attention modules to spread feature generation, which contains successive transposed 3-D convolution ( $\text{CONV}^{-1}_{\text{BN}}$ ) and generates HSI cubes with spectral–spatial distributions. The architecture of the spectral attention generation block takes the form

$$z^{p+2} = z^p + F(z^p; \theta), \quad (12)$$

$$F(z^p; \theta) = (\mathbf{Atte}_{\text{spc}} \otimes \mathbf{R}(\hat{z}^{p+1})) *^T h^{p+2} + b^{p+2}, \quad (13)$$

$$z = \mathbf{R}(\hat{z}^p) *^T h^{p+1} + b^{p+1} \quad (14)$$

where each element of  $\theta$  indicates parameters of spectral transposed convolutional layers,  $\mathbf{Atte}_{\text{spc}}$  is the continuation of (2), and  $*^T$  denotes the transposed convolution operation.  $\hat{z}^p$  is the normalization result of batch feature cubes  $z^p$ , whose calculation refers to (7). Similarly, the spatial attention generation block takes the form

$$Z^{q+2} = Z^q + F(Z^q; \xi), \quad (15)$$

$$F(Z^q; \xi) = \mathbf{Atte}_{\text{spa}} \otimes \mathbf{R}(\hat{Z}^{q+1}) *^T H^{q+2} + b^{q+2}, \quad (16)$$

$$Z = \mathbf{R}(\hat{Z}^q) *^T H^{q+1} + b^{q+1} \quad (17)$$

where  $\xi$  denotes parameters of spatial transposed convolutions, and  $\mathbf{Atte}_{\text{spa}}$  is obtained from (3). Furthermore,  $\hat{Z}^q$  is also the BN results of batch feature input  $Z^q$ , which is computed as (11).

Unlike traditional feature representation blocks, which perform the attention mechanism after the feature extraction for HSI data characterization, the proposed spectral–spatial attention feature spread and generation blocks are feature-efficient, i.e., the attention maps are executed during the feature extraction. It can be described from two aspects. 1) The consecutive spectral–spatial attention feature spread blocks of the discriminator draw the SSAT into the architecture for training, which provides learnable spectral and spatial attributes. On the one hand, being similar to SNR enhancement, the spectral attention weight  $\mathbf{Atte}_{\text{spc}}$  retains the high-frequency details of the HSI data and improves the discrimination of high-level semantic description. On the other hand, the spatial attention weights  $\mathbf{Atte}_{\text{spa}}$  with the denoising theory can emphasize a broader receptive field to learn adaptive neighborhood relations. Under the guidance of SSAT, the discriminator can always obtain excellent interpretation ability for the HSI data, whether in high-purity spectral domains or high texture local regions. 2) The spectral–spatial attention feature generation blocks of the generator share both  $\mathbf{Atte}_{\text{spc}}$  and  $\mathbf{Atte}_{\text{spa}}$  with that of feature spread blocks. It means that the implicit synthetic HSI cubes produced by generator help the discriminator learn more robust and efficient characteristics.

### C. Semisupervised SSAT-GAN

Taking the Pavia University (UP) dataset as raw input HSI cubes, Fig. 5 details the SSAT-GAN algorithm stream. The discriminator  $D$  contains a spectral attention feature spread block, spatial attention feature spread block, and one FC, and it outputs the vectors with the softmax layer. The generator  $G$  includes one FC, spectral attention generation block, and spatial attention generation block to generate HSI cubes. In addition, we extend the SSAT-GAN to semisupervised classification, which adopts unlabeled training samples of a raw HSI cube to improve HSI classification.

In contrast to original GANs, semisupervised SSAT-GAN leads a supervised item into the GAN loss to achieve the HSI classification. The labeled HSI cube  $X^1 = \{x_i^1\} \in \mathbb{R}^{7 \times 7 \times 103}$  has its corresponding annotation labels  $Y^1 = \{y_i^1\} \in \mathbb{R}^{1 \times (1+n_y)}$ , where  $n_y$  is the total number of ground truth category, and the extra “1” category denotes whether the HSI cube is from synthetic or real data. Therefore, the prediction of the well-trained  $D$  can take the form

$$\hat{Y}^1 = D(X^1; \theta_D) \quad (18)$$

where  $\theta_D$  denotes parameters for training  $D$  for each element of  $y_i^1$ , which includes  $(1 + n_y)$  entries. In particular,  $y_i^1[0]$  is the authenticity of  $x_i$ , and  $y_i^1[1 : n_y]$  denotes the output vectors of softmax, which contain probabilities that  $y_i^1$  belongs to each category.

Semisupervised GAN aims to alleviate the issue of small samples by labeled and unlabeled data of HSI. The point of view referred in [33] illustrated that  $D$  needs a bad  $G$  as a regularizer for training GANs. An opposite theory cited in [34] has pointed out that high-quality synthetic samples help  $D$  improve generalization ability for HSIs. In our proposal, we extend our spectral–spatial attention weights to  $G$ , reconstructing HSI cubes, implicitly. It can be divided into two phases. First,  $G$  is considered as the regularizer of  $D$  to improve HSI classification, and it updates the penalty factor with the discriminative loss. Thus, the optimized loss function of  $D$  takes the form

$$\begin{aligned} L_{\text{SEMI}}(\theta_D, \theta_G) &= L_{\text{SUP}}(\theta_D, \theta_G) + L_{\text{UNSUP}}(\theta_D, \theta_G) \\ &= L_{\text{SUP}}(\theta_D) + L_{D1}(\theta_D) \\ &\quad + L_{D2}(\theta_D, \theta_G) \end{aligned} \quad (19)$$

where  $\theta_D$  and  $\theta_G$  are the optimization parameters of the  $D$  and  $G$ , respectively.  $L_{\text{SEMI}}$  is the total objective loss for optimizing SSAT-GAN.  $L_{\text{SUP}}$ ,  $L_{D1}$ , and  $L_{D2}$  are, respectively, the unsupervised and supervised items of  $D$ , and the unsupervised item of  $G$ . These items are all formulated as

$$\begin{aligned} L_{\text{SUP}}(\theta_D) &= -E_{X^1 \sim p_{\text{data}}} \log D(X^1; \theta_D)[1 : n] \\ &= -E_{X^1 \sim P_{\text{data}}} \log \hat{Y}^1[1 : n], \end{aligned} \quad (20)$$

$$\begin{aligned} L_{D1}(\theta_D) &= -E_{X^1 \sim p_{\text{data}}} (1 - \log D(X^1; \theta_D)[0]) \\ &= -E_{X^1 \sim p_{\text{data}}} \log (1 - \hat{Y}^1[0]), \end{aligned} \quad (21)$$

$$\begin{aligned} L_{D2}(\theta_D, \theta_G) &= -E_{z \sim p_z} \log D(G(z; \theta_D))[0] \\ &= -E_{z \sim p_z} \log D(Z; \theta_D)[0] \\ &= -E_{z \sim p_z} \log \hat{Y}^1[0] \end{aligned} \quad (22)$$

where  $L_{\text{SUP}}$  is applied for optimizing the real HSI predictions of softmax vectors, which corresponds to  $y_i^1[1 : n_y]$  from (18).  $L_{D1}$  aims at updating the recognition degree by unlabeled HSI cubes, and  $L_{D2}$  focuses on increasing the authenticity of generated samples, which both correspond to  $y_i^1[0]$  from (18).

It is to be observed that the optimization of a semisupervised GAN focuses on exploring a real HSI data distribution by limited labeled samples, which often causes overfitting. As the high

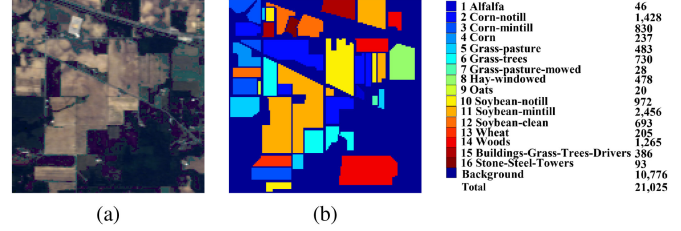


Fig. 6. Indian Pines dataset. (a) False-color image. (b) Ground-truth labels.

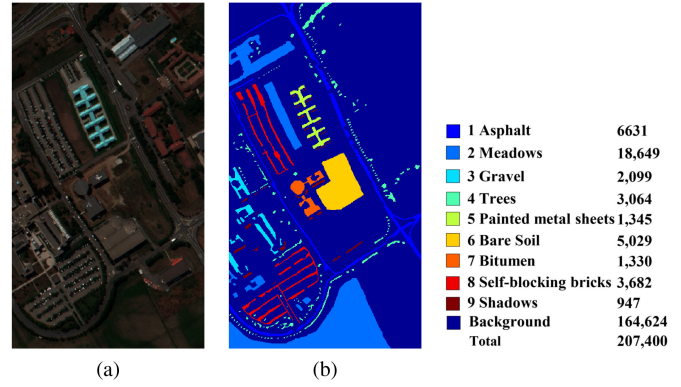


Fig. 7. Pavia University dataset. (a) False-color image. (b) Ground-truth labels.

dimensional feature learning of LD1 is not constrained, it will contribute little to and even jeopardize the discriminator to enhance the capability of HSI classification. Thus, we minimize the high-dimensional output of (21) to update the gradient in reverse and decrease the value and variance to inhibit overfitting, which is available in another work [36] called mean minimization loss. The function takes the form

$$\theta^* = \operatorname{argmin}_{\theta} \left( \frac{1}{N} \sum_{i=1}^N \operatorname{average}(f(x_i; \theta)) \right) \quad (23)$$

where  $N$  is the total entities of batch samples,  $x_i$  is the training sample, and  $f(x_i; \theta)$  indicates the high-dimensional output of a model, which, in this article, is the output before the FC. Second, we employ the predictive spectral–spatial attention weights for generating high-quality samples. Furthermore,  $L_{D1} + L_{D2}$  is also part of the GAN loss for training  $G$ , whose corresponding loss function  $L_G$  can be formulated as

$$\begin{aligned} L_G(\theta_D, \theta_G) &= -E_{z \sim p_z} \log (1 - D(G(z; \theta_D))[0]) \\ &= -E_{z \sim p_z} \log (1 - D(Z; \theta_D)[0]) \\ &= -E_{z \sim p_z} \log (1 - \hat{Y}^1[0]). \end{aligned} \quad (24)$$

The training of SSAT-GAN involves two alternating steps through *rms* or adjacent optimization fashions in every epoch. First, the gradients of the discriminator  $-\nabla_{\theta_D} L_{\text{SEMI}}$  are employed to adjust  $\theta_D$  to capture discriminative spectral–spatial features of HSI. Second, the gradients of  $-\nabla_{\theta_D} L_G$  are applied to adjust  $\theta_G$  to ameliorate the adversarial training. The detailed training process of SSAT-GAN is described in Algorithm 1.

**Algorithm 1:** Training Process of SSAT-GAN.

**Input:** The labeled training data:  $\mathbf{X}_{\text{train}}^l$ , unlabeled training data  $\mathbf{X}_{\text{train}}^u$ , and the test data  $\mathbf{X}_{\text{test}}$  from  $n_y$  classes, corresponding annotation of training data  $Y^l$ , the batch size  $bt$ , and the number of training epochs  $e$ .

**Output:** The labels of the test samples  $\mathbf{X}_{\text{test}}$

```

1: Begin
2: Initialize: Randomly initialize the parameters  $\theta_D$  and  $\theta_G$  of the discriminator  $D$  and the generator  $G$ ;
3: for  $i = 0$  to epoch  $e$  do
4:   for  $bt$  training samples of each batch do
5:     Generate  $bt$  noises  $\{z_1, z_2, \dots, z_{bt}\}$  from the Gaussian distribution  $\mu(-1, 1)$ ;
6:     Concatenate noises with labels  $\{y_1, y_2, \dots, y_{bt}\}$ ;
7:     Input  $\mathbf{X}_{\text{train}}^l$  into  $D$  to obtain real HSI features via (4) and (8);
8:     Calculate  $\text{Atte}_{\text{spc}}$  via (2);
9:     Calculate  $\text{Atte}_{\text{spa}}$  via (3);
10:    Predict classification vectors  $D(\mathbf{x}_i^l; \theta_D)[1 : n_y]$ ;
11:    Compute  $L_{\text{SUP}}$  via (20);
12:    Input noises  $\{z_1, z_2, \dots, z_{bt}\}$ , class labels  $\{y_1, y_2, \dots, y_{bt}\}$ ,  $\text{Atte}_{\text{spc}}$ , and  $\text{Atte}_{\text{spa}}$  to  $G$ ;
13:    Generate samples  $\mathbf{Z}$  via (13) and (16);
14:    Input  $\mathbf{X}_{\text{train}}^u$  and  $\mathbf{Z}$  to  $D$ ;
15:    Predict authentic vectors  $D(\mathbf{x}_i^u; \theta_D)[0]$ ;
16:    Compute  $L_{D1}$  and  $L_{D2}$  via (21) and (22);
17:    Compute  $L_G$  via (24)
18:    Update  $\theta_D$  by minimizing  $L_{\text{SUP}} + L_G$ 
19:    Update  $\theta_G$  by minimizing  $1 - D(z_i; \theta_D)[0]$ ;
20:     $bt = bt + 1$ ;
21:   end for
22:    $i = i + 1$ ;
23: end for
24: Classify  $\mathbf{X}_{\text{test}}$  by the well-trained  $D$ ;

```

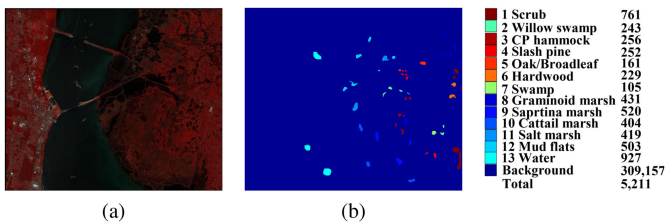


Fig. 8. Kennedy Space Center dataset. (a) False-color image. (b) Ground-truth labels.

#### IV. EXPERIMENTAL ANALYSIS

We detail the experimental results from three real hyperspectral datasets, including the Indian Pines (IN), the University of Pavia (UP), and the Kennedy Space Center (KSC). Each of them is standardized by mean variance operation. Three classification evaluation metrics, including overall accuracy (OA), average accuracy (AA), and kappa coefficient ( $\kappa$ ), are employed to validate the experimental performance of SSAT-GAN and the comparison algorithms. In particular, OA considers the total percentage of correctly classified pixels; AA details the average

percentage to the sum of correctly classified pixels in each category; the kappa coefficient provides the corrected percentage of correctly classified pixels as expected purely by chance based on confusion matrix. All experiments are implemented with an NVIDIA TITAN V GPU with 12-GB graphic memory, TensorFlow GPU 1.8.0 with CUDA 9.0, and Python 3.5.

#### A. Experimental Datasets

1) *Indian Pines*: IN was acquired by airborne visible/infrared imaging spectrometer (AVIRIS) from Northwest Indiana in 1992 and includes 16 vegetation categories, with an imbalance in pixel numbers over categories. It contains  $145 \times 145$  spectral pixels with a spatial resolution of 20 m per pixel, retaining 200 bands of spectrum from 400 to 2500 nm after removing corrupted water-absorption effects.

2) *University of Pavia*: UP was gathered by reflective optics system imaging spectrometer (ROSIS) in 2001 from Northern Italy, consisting of  $610 \times 340$  spectral pixels with nine urban land-cover classes, and 1.3 m spatial resolution per pixel, employing 103 bands of spectrum from 430 to 860 nm after abandoning 20 noisy bands.

3) *Kennedy Space Center*: KSC was obtained by AVIRIS in 1996 from Florida and includes 13 upland and wetland land-cover types, with  $512 \times 614$  spectral pixels and 176 bands of spectrums to assess the classification capacity, after discarding information with a low-SNR, with a range from 400 to 2500 nm.

Figs. 6–8 illustrate the dataset, the corresponding ground reference maps, and category information. All labeled samples are split into two groups: the training group and the test group. For the unlabeled group, the unlabeled training samples are randomly selected from the background. GANs contain relatively higher computational complexity, which is often guided to the mode collapse. Thus, we refer to Monte Carlo sampling [45] which is mentioned in [33] to marginalize noise during training.

#### B. Parameter Tuning

Fig. 5 takes the UP neighboring cube as an instance to show the detail of the discriminator  $D$  and generator  $G$ . The  $7 \times 7 \times 103$  HSI cubes are randomly directly extracted from raw 3-D HSI data as the real input, followed by feeding them into  $D$ .  $G$  utilizes  $1 \times 1 \times 200$  noise vectors as the input and outputs  $7 \times 7 \times 103$  fake HSI cubes. We alternately update the parameters of the SSAT-GAN through backpropagation of the gradients. For the efficiency of the grid search, we set the learning rate to 0.0005 and the batch size to 16 and employ the RMSProp optimizer [46] to alternately optimize them. Once the hyperparameters of SSAT-GAN are configured, we analyze four factors that avoid model collapse and influence the HSI interpretation performance of SSAT-GAN.

1) *Evaluation of Different Depths of Spectral–Spatial Attention Block*: We assessed the impact of different depths of the spatial–spectral attention feature spreading blocks on classification results. For SSAT-GAN, the depths of blocks were validated from four convolutional layers to eight convolutional layers on all datasets. To maintain the stability of the model, the depth of the generator was symmetric to that of the discriminator. As



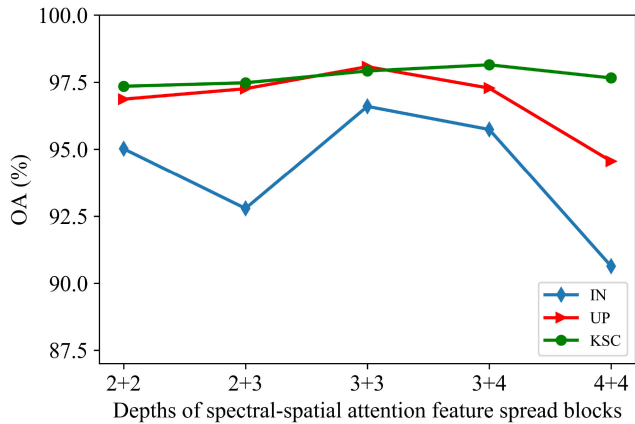


Fig. 9. OAs of SSAT-GAN with different depths of convolutional layers in their spectral–spatial attention feature spread blocks using 500 labeled samples on IN and UP, and 250 on KSC for training. The  $x + y$  formation on the abscissa indicates  $x$  spectral and  $y$  spatial convolutional layers in discriminator.

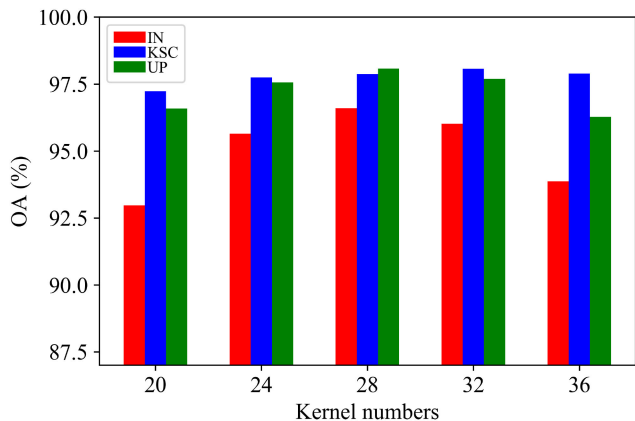


Fig. 10. OAs of SSAT-GAN for varying kernel numbers in their spectral–spatial attention spreading blocks using 500 labeled samples on IN and UP, and 250 on KSC for training.

illustrated in Fig. 9, it achieved the highest evaluation results on both IN and UP datasets, when set the depths of spectral-spatial attention feature spread blocks to “3 + 3”, i.e. the discriminator which consists of 3 spectral and 3 spatial convolutional layers, compared with other settings of convolutions. As for the KSC, the differences of OAs between deeper SSAT-GANs and their corresponding shallow depth get a small value. Meanwhile, in contrast to the obvious overfit deeper layers under limited training sample effects reviewed in [33], the quantitative HSI classification performance of SSAT-GANs with varying depths illustrated that our attention modules mitigate the overfitting effects to other GANs.

2) *Evaluation of Different Numbers of Kernels for SSAT-GAN*: Kernel numbers of each layer from feature spreading blocks greatly affects computation consumption and expressiveness of SSAT-GANs. We evaluated the impact of different numbers of kernels of the spectral–spatial attention feature spreading blocks on the results. In Fig. 10, the discriminator and the generator of SSAT-GAN set the same kernel number in their convolution and transposed convolution layers, with the number of kernels verified from {20, 24, 28, 32, 36}. As can be

seen from Fig. 10, when the kernel numbers were fixed at 28 and 32, it achieved the highest classification results on all three datasets.

3) *Influence of Unlabeled Real HSI Cubes*: To evaluate the influence of unlabeled real HSI cubes, we tested SSAT-GAN and its three extensions using different numbers of unlabeled HSI samples on the IN, UP, and KSC datasets. The three extensions of SSAT-GAN are denoted as Spa-AT-GAN (the ones that only contain the spatial attention feature spreading part), the Spc-AT-GAN (the ones that only contain the spectral attention feature spreading part), and the Spa-Spc-AT-GAN (the ones that contain both spreading blocks, where the spatial attention module is set before the spectral attention module). Table I recorded the classification results of SSAT-GANs. Each experiment randomly selected 0, 300, 1000, and 5000 unlabeled samples for training.

For IN and KSC, the classification of SPA-SPC-AT-GAN did not efficiently improve with the increase of unlabeled samples. Among the four methods, SSAT-GAN had the best evaluation on each dataset with various unlabeled samples due to the spectral–spatial attentive feature learning guidance. Moreover, models with 300 unlabeled samples had the most accurate evaluation on all three datasets, and the improvement of 1000 unlabeled samples was not obvious. When the number increased to 5000, the results showed a downward trend in all extensions of SSAT-GAN on three datasets. This proves that adding too many real samples does not greatly improve the classification, which is caused by the abnormal distribution of unlabeled pixels. In addition, it can be seen that the HSI classification has been significantly improved if the unlabeled samples are set equal to the labeled samples. This conclusion is consistent with the opinions reported by Zhong *et al.* [33] and Liang *et al.* [36].

4) *Evaluation of Different Spatial Data Sizes*: To assess the impact of spatial size on the experimental results, we tested SSAT-GAN with spatial data sizes of  $\{5 \times 5, 7 \times 7, 9 \times 9, 11 \times 11, 13 \times 13\}$ . Fig. 11 shows that SSAT-GAN could capture relatively high and stable results while the spatial size was equal to or greater than  $7 \times 7$ . This is mainly because a larger spatial size has more abundant spatial information. These experimental results also indicate that spatial contexts gradually gain an important role in HSI classification.

TABLE I  
OAS (%) OF SSAT-GANs USING VARIOUS NUMBER OF UNLABELED SAMPLES AND 300 LABELED SAMPLES IN THE IN, UP, AND KSC DATASETS

Datasets	Models	0	300	1000	5000
IN	Spa-AT-GAN	80.88	<b>82.58</b>	74.33	67.31
	Spc-AT-GAN	77.77	79.97	<b>81.64</b>	70.00
	Spa-Spc-AT-GAN	79.71	<b>80.36</b>	79.97	70.65
	SSAT-GAN	92.89	<b>93.00</b>	92.52	87.89
UP	Spa-AT-GAN	89.98	<b>91.01</b>	82.74	80.08
	Spc-AT-GAN	93.26	95.39	<b>96.03</b>	92.71
	Spa-Spc-AT-GAN	89.84	<b>92.42</b>	92.33	87.65
	SSAT-GAN	95.75	<b>97.80</b>	97.23	95.95
KSC	Spa-AT-GAN	92.77	<b>94.64</b>	91.65	87.11
	Spc-AT-GAN	95.26	<b>95.79</b>	94.98	93.24
	Spa-Spc-AT-GAN	88.67	91.92	<b>91.98</b>	85.45
	SSAT-GAN	96.98	<b>97.97</b>	97.45	93.29

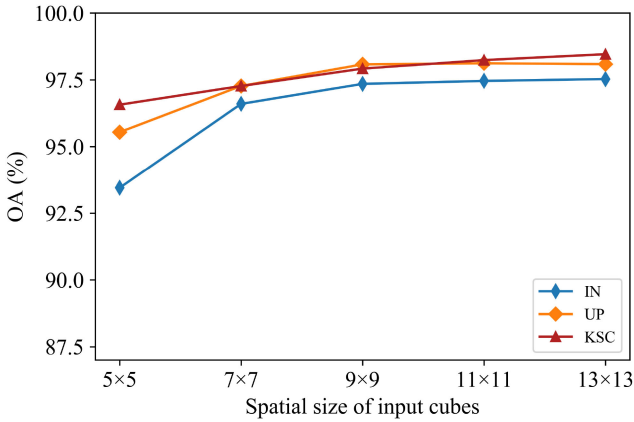


Fig. 11. OAs of SSAT-GAN containing various spatial sizes of input cubes on three datasets.

### C. Comparison With Various Algorithms

This experiment aimed to compare the performance of the proposed SSAT-GAN with the EPF-SVM [12] (EPF-based SVM) and the state-of-the-art deep learning derived methods, such as SSRN [18], 3D-Conv-Capsule [20], and HSI-BERT [26]. To verify the improvement of GANs, we exploited three GAN-based methods for comparison, including 3D-GAN [32], GAN-CRF [33], and AD-GAN [34]. Moreover, to demonstrate the effectiveness of the SSAT module, we also introduced the extensions of SSAT-GAN: Spa-AT-GAN (only comprises one spatial attention feature spreading block), Spc-AT-GAN (only comprises one spectral attention feature spreading block), and Spa-Spc-AT-GAN (comprises one spatial attention feature spreading block and one spectral attention feature spreading block). To make a fair comparison, all the competitive algorithms were tuned to their optimal settings.

Regarding the EPF-SVM, the two parameters of the joint bilateral filter were set as follows:  $\delta_s = 4$  and  $\delta_r = 0.2$ . Meanwhile, the hyperparameters of SVM were set as follows:  $\gamma = 4$  and  $\epsilon = 0.01$ . For SSRN and HSI-BERT, we set the input HSI cubes with the same spatial size of  $7 \times 7$ . For 3D-Conv-Capsule, the routing interaction was set to three times to determine its coupling coefficients. For 3D-GAN, the first three principal components of HSIs were applied for channel input, and the spatial size was set to  $64 \times 64$ . For GAN-CRF, the neighborhood of  $9 \times 9$  pixels was employed and configured three spectral and spatial convolutional layers in the discriminator. For AD-GAN, the 3-D HSI cubes of size  $27 \times 27 \times 3$  were considered as input, and an AdapDrop block was executed once at both the discriminator and the generator, each of which set  $k = 40$  and  $b\_size = 7$ .

As for the proposed SSAT-GAN, we set the spatial size of input HSI cubes to  $7 \times 7$  and trained 300 epochs. Both  $D$  and  $G$  were built with consecutive spectral-spatial attention feature spread blocks and spectral-spatial attention feature generation blocks, each of whose kernel number was 28. The minibatch was 16. To avoid the mode collapse, we set unlabeled samples with the same number of the labeled training samples that were used for training. Furthermore, all the comparison methods were

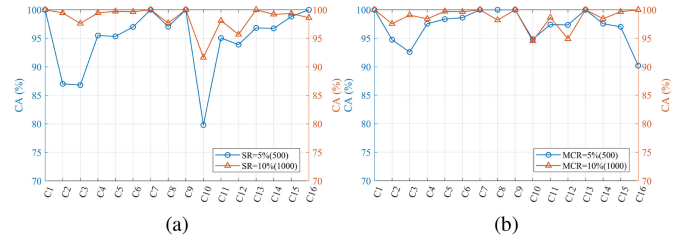


Fig. 12. Accuracy value of the IN dataset with SSAT-GAN model under different training data sampling. We report the average results of ten experiments. (a) Randomly selection strategy. (b) Monte Carlo sampling strategy.

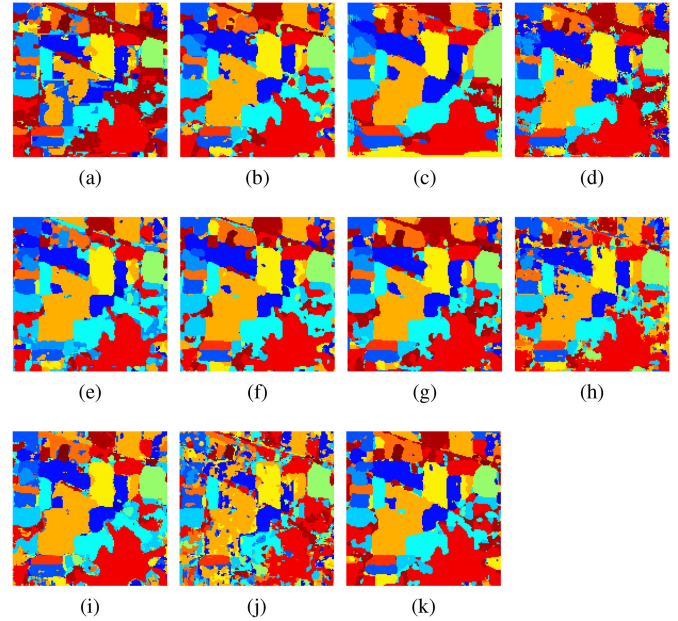


Fig. 13. Classification visualization of comparison models on IN dataset. (a) EPF-SVM. (b) SSRN. (c) 3D-Conv-Capsule. (d) HSI-BERT. (e) 3D-GAN. (f) GAN-CRF. (g) AD-GAN. (h) Spa-AT-GAN. (i) Spc-AT-GAN. (j) Spa-Spc-AT-GAN. (k) SSAT-GAN.

trained and evaluated using 10 randomly sampled experiments, and the average results and their standard deviations for the report were recorded.

1) *Experimental Results on IN Dataset:* For various methods, 500 labeled pixels were employed as training samples on the IN dataset. Table II lists the quantitative classification results of comparison methods, and the visualization maps are illustrated in Fig. 13. As shown in Table II, EPF-SVM yielded poor accuracies in the “Corn,” “Soybean-notil,” and “Buildings-Grass-Trees-Drivers” classes, which are 63.09%, 67.93%, and 52.74%, respectively. This is caused by their similarity of spectral curves, which makes them difficult to identify. In contrast, we observed that SSRN, 3D-Conv-Capsule, and HSI-BERT acquired better results than EPF-SVM in the three classes. However, in the HSI-BERT, it improved at least 12.21% in the “Corn” class. It can be analyzed that deep-learning methods have a certain positive effect on interpreting complex spectral characteristics. Different from the former, GAN-based methods showed superior prediction in the three classes, and 3D-GAN improved at least 26.66% in “Corn”. Besides, GAN-CRF achieved 93.05% in

“Soybean-notil,” and AD-GAN had classified the “Corn” completely accurate. As for SSAT-GANs, both Spa-AT-GAN and Spc-AT-GAN achieved advanced prediction in the three classes.

As SSAT can improve the intraclass aggregation, which effectively distinguishes the difference between spectra during hyperspectral interpretation. In the “Alfalfa,” “Grass-pasture-mowed,” and “Oats” categories, only 5, 3, and 5 pixels were used as training samples. SSAT-GAN gains superior classification, all with accuracies of 100%. This indicates that our SSAT-GAN can extract sensitive features under the classes of small samples. Among the competitive methods, SSAT-GAN also gathered the best accuracies in the “Soybean-mintil” and “Soybean-clean,” which contain redundant spectral signatures. Meanwhile, SSAT-GAN outperformed various comparison methods according to OA, AA, and kappa. In contrast to SSAT-GAN with its extensions, SSAT-GAN improved the OA with extensions by at least 1.01%, AA by 1.35%, and kappa by 1.25%. This illustrates that the SSAT module can extract discriminative spectral signatures and adaptive homogeneous areas to mitigate the impact of interfering pixels of HSIs. Besides, it should be noted that the experiment in Spa-Spc-AT-GAN showed inferior performance because the abundant spectral features are more difficult to learn than spatial features.

To verify the results of the SSAT-GAN under the Monte Carlo sampling, we also experiment with the randomly selection strategy under the training sampling ratio (SR) fixed as 5% (500 training samples) and 10% (1000 training samples). As shown in Fig. 12(a), the detailed class accuracy (CA) of each class shows qualitative comparisons with different circumstances, in which the “Soybean-notil” class obtained unsatisfactory results under the randomly selection strategy, with 79.79% of SR = 5% and 91.65% of SR = 10%. Besides, it is worth mentioning that the experiment under the SR = 10% achieves better realistic performance compared to that of SR = 5%. The reason is that the randomly chosen samples for any other classes contain more outstanding characteristics which confuse that of small-sample classes, when the data distribution is imbalanced. To confirm this stated opinion, we adopted the Monte Carlo sampling to redo the experiment with our SSAT-GAN with the SR MCR = 5% and

MCR = 10%. The Monte Carlo sampling considers the interclass sample distribution under preserving the total random sampling size (as shown in Table II). As can be seen from Fig. 12(b), the performance of each class in the style of Monte Carlo sampling has superior observations than the indication in Fig. 12(a). From Fig. 13(a)–(k), EPF-SVM and Spa-Spc-GAN has got more visual noise and had the most misclassified pixels; besides, visualizations of SSRN and HSI-BERT got rough boundaries in most classes. The reason is that the imbalanced sample distribution of the IN dataset, in which part of classes with a large number of samples may contain more discriminative characteristics to the identification. 3D-GAN, GAN-CRF, Spa-AT-GAN, and Spc-AT-GAN gained relatively little visual noisy scatter. In contrast, 3D-Conv-Capsule and AD-GAN significantly reduced the impact of noise and established homogeneous areas. Among them, SSAT-GAN had more uniform regions and set up an adaptive neighboring relationship, from which it can be noted that SSAT can effectively suppress information detrimental to classification.

2) *Experimental Results on UP Dataset.* The evaluation of the comparison methods on the UP dataset is listed in Table III using 500 labeled samples. We can see that OAs yielded with SSRN, 3D-GAN, and GAN-CRF are 95.31%, 93.89%, and 94.95%. Our proposed model can further increase the performance to 98.09% by incorporating the SSAT module. In similarity, the AA values are 94.33%, 94.25%, and 97.16% for the 3D-Conv-Capsule, HSI-BERT, and AD-GAN, respectively. As can be observed, the proposed SSAT-GAN has a relatively stable and balanced classification effect for each category under high-resolution neighboring relationships and acquired the maximum AA value (98.21%). It can be noted that the proposed SSAT module can capture discriminative interclass differences and is essential and beneficial for the proposed architecture.

The classification visualization on the UP dataset is described in Fig. 14. It can be seen that the comparison methods produced rough prediction maps, especially in 3D-GAN and Spa-Spc-AT-GAN, which was caused by atmospheric effects and instrument noises. SSAT-GAN aimed at neighboring correlation context as

TABLE II  
CLASSIFICATION ACCURACIES AND TRAINING AND TESTING TIMES OF VARIOUS COMPARISON METHODS USING 500 LABELED SAMPLES AND 500 UNLABELED SAMPLES FOR THE IN DATASET

Class	Train.(Test.)	EPF-SVM	SSRN	3D-Conv-Capsule	HSI-BERT	3D-GAN	GAN-CRF	AD-GAN	Spa-AT-GAN	Spc-AT-GAN	Spa-Spc-AT-GAN	SSAT-GAN
1	5 (41)	<b>100.00±0.00</b>	98.51±2.93	96.94±1.39	72.43±15.54	60.12±5.65	93.04±1.87	<b>100.00±0.00</b>	<b>100.00±0.00</b>	<b>100.00±0.00</b>	<b>100.00±0.00</b>	<b>100.00±0.00</b>
2	68 (1,360)	88.54±5.32	92.37±5.92	85.88±0.48	89.58±3.17	91.06±1.72	<b>96.10±1.20</b>	92.20±1.87	94.46±2.19	92.61±1.64	86.02±3.23	94.76±1.11
3	48 (782)	85.20±6.98	91.96±0.49	91.91±0.35	92.87±3.10	80.65±0.90	92.52±2.33	81.41±2.81	82.88±5.72	88.95±1.78	79.87±5.35	<b>92.60±2.24</b>
4	11 (226)	63.09±8.18	92.39±4.38	96.81±0.58	75.30±10.23	89.75±0.87	92.51±1.59	<b>100.00±0.00</b>	<b>100.00±0.00</b>	97.43±0.89	92.14±1.17	97.56±0.62
5	25 (458)	96.84±2.00	99.34±0.62	91.63±2.03	93.03±2.18	73.31±3.26	99.07±0.10	<b>99.76±0.19</b>	96.62±1.84	97.82±1.00	84.88±2.42	98.36±1.08
6	37 (693)	93.77±2.71	97.70±1.32	99.20±0.63	<b>99.26±0.44</b>	89.72±1.62	98.95±0.12	97.87±0.24	95.97±1.75	98.61±0.26	94.95±2.24	98.62±1.80
7	3 (25)	<b>100.00±0.00</b>	99.16±0.25	82.27±7.72	64.40±11.51	93.26±2.67	96.27±2.10	82.75±2.75	<b>100.00±0.00</b>	<b>100.00±0.00</b>	37.56±11.98	<b>100.00±0.00</b>
8	25 (453)	<b>100.00±0.00</b>	98.84±1.35	<b>100.00±0.00</b>	99.91±0.0	98.89±0.14	99.58±0.50	96.99±2.53	94.94±2.89	97.93±1.32	90.00±2.46	<b>100.00±0.00</b>
9	5 (15)	<b>100.00±0.00</b>	94.83±1.11	86.99±0.64	98.00±3.05	75.57±3.82	98.09±1.11	93.46±1.88	<b>100.00±0.00</b>	37.50±11.42	75.21±6.76	<b>100.00±0.00</b>
10	60 (912)	67.93±2.16	82.01±0.17	<b>98.55±1.08</b>	92.21±2.42	91.10±2.87	93.05±1.35	92.62±0.26	93.54±2.22	98.21±0.68	76.95±4.24	94.83±2.47
11	106 (2,349)	89.54±4.93	95.04±0.42	95.68±0.30	94.51±1.62	90.29±2.15	94.68±1.48	95.57±1.62	92.55±2.52	95.31±1.76	90.49±2.09	<b>97.41±1.08</b>
12	36 (557)	92.11±2.88	88.06±0.10	92.78±1.49	86.22±0.58	78.87±4.16	91.41±1.44	93.64±1.25	78.50±3.20	93.69±2.10	70.87±3.12	<b>97.35±1.27</b>
13	6 (199)	<b>100.00±0.00</b>	99.39±0.19	98.29±1.37	96.58±0.36	80.56±3.27	98.80±1.00	98.47±0.68	<b>100.00±0.00</b>	<b>100.00±0.00</b>	90.09±2.64	<b>100.00±0.00</b>
14	46 (1,219)	<b>99.49±1.91</b>	97.41±1.67	99.68±0.58	97.76±0.16	90.02±2.24	98.37±1.16	97.16±0.33	95.80±0.59	98.81±0.97	96.49±1.76	97.57±1.25
15	16 (370)	52.74±10.40	95.00±0.42	98.30±0.15	86.08±0.62	91.08±1.94	94.35±1.43	<b>98.61±0.24</b>	98.30±1.01	97.36±0.99	68.76±6.44	97.02±0.29
16	3 (90)	85.92±8.48	97.98±0.31	66.89±5.10	90.55±0.63	75.79±5.48	<b>96.96±1.16</b>	93.25±1.09	93.75±1.53	94.73±0.68	85.53±2.18	90.21±1.99
OA(%)		85.58±2.06	92.44±3.10	94.95±0.82	93.00±0.69	91.46±1.36	95.35±1.45	95.59±0.17	91.19±0.57	95.42±1.68	82.48±2.87	<b>96.60±1.63</b>
AA(%)		88.45±2.20	95.07±1.24	92.74±1.06	89.29±2.12	84.37±2.12	95.87±1.42	94.61±0.21	94.21±1.72	93.06±2.32	80.95±4.19	<b>97.22±1.31</b>
$\kappa \times 100$		83.57±2.36	91.42±3.45	94.25±0.93	92.00±0.80	89.98±1.78	94.87±1.54	93.85±0.20	89.91±1.26	94.77±0.96	81.97±1.79	<b>96.12±1.01</b>
Train. Time(mS)		45.07	1516.42	2409.04	1022.58	2038.33	2569.63	1226.13	608.46	1338.84	1361.29	1151.59
Test. Time(mS)		757.60	6509.89	5838.31	7352.00	5357.50	5095.77	6230.57	2490.72	4151.10	7783.87	3368.34

Note: The best values are highlighted in bold font.

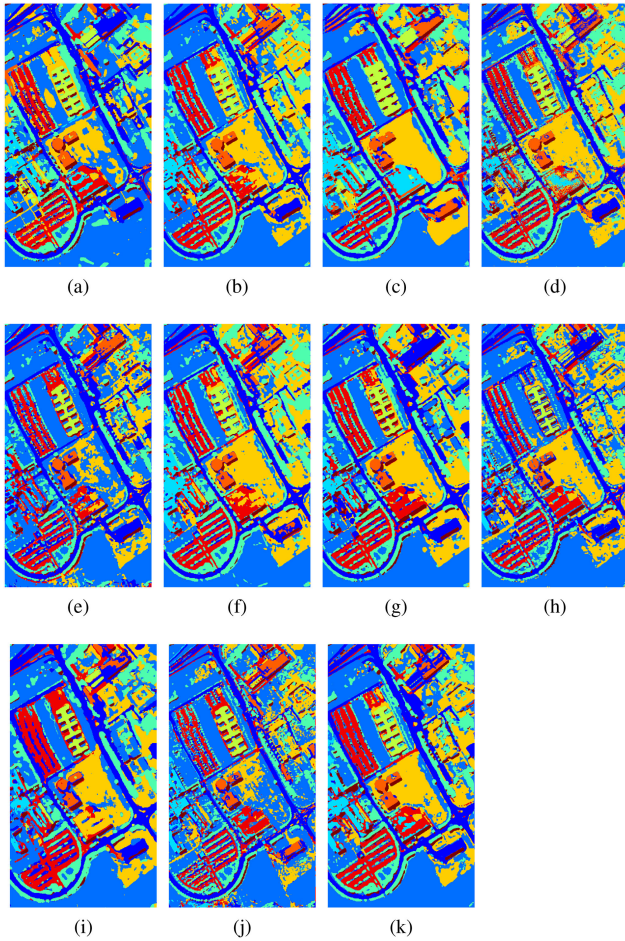


Fig. 14. Classification visualization of comparison models on UP dataset. (a) EPF-SVM. (b) SSRN. (c) 3D-Conv-Capsule. (d) HSI-BERT. (e) 3D-GAN. (f) GAN-CRF. (g) AD-GAN. (h) Spa-AT-GAN. (i) Spc-AT-GAN. (j) Spa-Spc-AT-GAN. (k) SSAT-GAN.

auxiliary information and had the smoothest results and clearest boundary.

3) *Experimental Results on KSC Dataset*: The last experiment is performed at the KSC dataset using 250 labeled pixels as training samples. As shown in Table IV, the SSAT-GAN achieved the best OA of 97.72% higher than GAN-CRF (95.38%) and AD-GAN (96.15%). In comparison, the SSRN

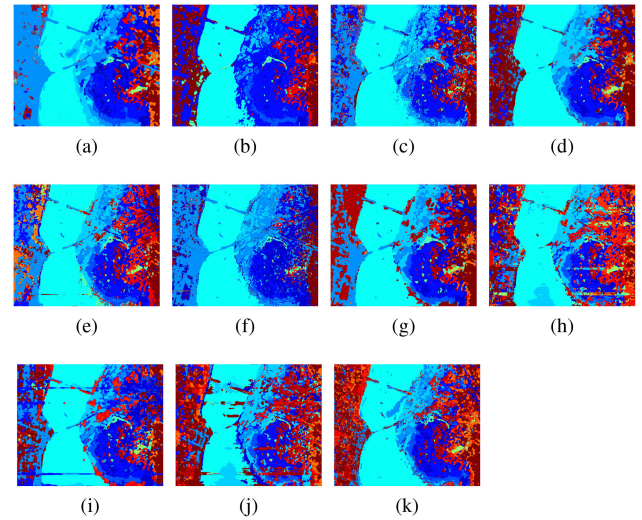


Fig. 15. Classification visualization of comparison models on KSC dataset. (a) EPF-SVM. (b) SSRN. (c) 3D-Conv-Capsule. (d) HSI-BERT. (e) 3D-GAN. (f) GAN-CRF. (g) AD-GAN. (h) Spa-AT-GAN. (i) Spc-AT-GAN. (j) Spa-Spc-AT-GAN. (k) SSAT-GAN.

yielded an OA of only 94.19%. The reasonable analysis is that the KSC dataset contains a relatively sparse characteristic so that the traditional network generally has more difficulties in interpreting spectral–spatial features. With the SSAT operation, the proposed model achieved superior performance in contrast to the other state-of-the-art methods. In addition, it needs to be noted that PCA-based 3D-GAN yielded the worst assessment with an OA of 93.38%, which illustrates that the representation of the primary components gains poor effect in the spectral–spatial feature extraction for HSIs with the characteristic of high sparsity. In contrast, our proposed architecture with the SSAT model acquires better robustness for the sparsity.

Classification maps are shown in Fig. 15. In contrast, it can be seen that SSAT-GAN achieved smoother and more adaptive visual results, which indicates that its SSAT module can both emphasize the intraclass consistency and increase interclass differences for HSI classification with high sparsity distribution. All the quantitative experiments conducted on the three datasets demonstrated that the SSAT-GAN framework reflects the excellence and robustness of HSI classification.

TABLE III  
CLASSIFICATION ACCURACIES AND TRAINING TIMES OF VARIOUS COMPARISON METHODS USING 500 LABELED SAMPLES AND 500 UNLABELED SAMPLES FOR THE UP DATASET

Class	Train.(Test.)	EPF-SVM	SSRN	3D-Conv-Capsule	HSI-BERT	3D-GAN	GAN-CRF	AD-GAN	Spa-AT-GAN	Spc-AT-GAN	Spa-Spc-AT-GAN	SSAT-GAN
1	75 (6,556)	97.31±0.44	<b>99.96±0.04</b>	92.93±0.36	97.32±1.10	90.33±0.83	92.01±2.00	98.65±0.01	96.37±0.85	97.34±1.41	96.99±0.79	97.87±1.02
2	220 (18,429)	98.29±0.92	97.87±0.64	99.32±0.12	98.64±0.14	97.12±1.26	98.63±0.39	97.86±0.37	96.20±0.78	99.54±0.21	97.49±0.87	<b>99.55±0.15</b>
3	24 (2,075)	92.22±5.07	96.88±2.91	83.73±0.69	80.25±5.21	92.19±0.23	80.83±3.61	96.80±0.35	92.36±1.62	83.29±2.12	83.76±2.87	<b>97.61±0.69</b>
4	35 (3,029)	92.14±7.67	97.85±0.34	97.44±1.42	94.13±1.34	95.27±0.75	92.68±1.87	<b>100.00±0.00</b>	<b>100.00±0.00</b>	98.66±0.66	98.27±0.88	99.93±0.10
5	18 (1,327)	97.09±1.82	99.56±0.36	99.82±0.15	99.98±0.03	98.71±0.51	<b>100.00±0.00</b>	<b>100.00±0.00</b>	<b>100.00±0.00</b>	99.62±0.24	99.62±0.37	<b>100.00±0.00</b>
6	59 (4,970)	88.55±1.63	<b>99.87±0.16</b>	99.04±0.59	96.27±1.53	95.53±0.83	95.14±0.25	98.32±0.64	94.98±0.37	98.61±1.02	97.12±0.94	92.01±2.14
7	16 (1,314)	72.40±8.06	69.67±10.07	82.09±5.63	86.23±5.22	92.33±1.02	97.65±0.97	97.68±1.43	95.57±1.47	98.84±1.02	94.23±0.97	<b>100.00±0.00</b>
8	43 (3,639)	90.09±1.52	81.08±4.57	96.27±0.27	95.41±2.23	80.57±3.46	89.18±2.67	85.17±3.78	88.47±4.32	93.29±1.92	92.13±1.16	<b>96.92±0.99</b>
9	10 (937)	95.58±2.01	99.92±0.12	97.72±0.54	98.99±0.75	99.10±0.51	99.26±0.42	<b>100.00±0.00</b>	<b>100.00±0.00</b>	97.89±0.72	98.93±0.37	<b>100.00±0.00</b>
	OA(%)	94.34±1.28	95.31±0.61	96.84±0.38	96.76±0.29	93.89±1.64	94.95±1.90	97.42±0.92	94.80±2.13	97.61±1.17	96.23±0.96	<b>98.09±0.42</b>
	AA(%)	91.85±1.45	93.74±1.11	94.33±0.86	94.25±0.84	93.46±1.26	94.42±1.43	97.16±0.77	95.16±0.85	96.34±0.40	95.39±1.40	<b>98.21±0.88</b>
	$\kappa \times 100$	92.52±1.63	93.77±0.81	95.81±0.50	95.69±0.39	91.87±1.27	96.87±1.35	96.53±0.23	93.07±1.79	96.83±0.97	94.98±1.66	<b>97.46±1.12</b>
	Train. Time(mS)	110.71	1281.99	2808.07	1157.66	2601.86	1768.93	1277.15	485.53	740.71	1349.75	1095.48
	Test. Time(mS)	458.08	20109.53	17830.94	16430.55	17890.22	17050.29	6950.39	9588.30	13544.43	13854.00	14868.67

Note: The best values are highlighted in bold font.

#### D. Investigation of the Impact of Attention Mechanism

To evaluate the effectiveness and the contribution of the attention mechanism, we compared various classical and representative attention modules which were executed over our GAN-baseline in Table V, including SE\_Block [47], CBAM [40], FA [48], and MAFN [24], and reported OAs of three datasets. It can be seen that both CBAM and our SSAT can obtain a considerable result on all three cases. This is caused by their forms of cascade connection that fit our architecture better. Besides, the FA module has a more promising result on both UP and KSC datasets in contrast to the IN dataset. The reason is that it requires a high spatial resolution to calculate the utility of covariance matrices over FA.

Moreover, we also investigated feature visualization with the guidance of the attention weights under the SSAT modules. In this experiment, only the  $7 \times 7$  neighboring HSI cubes were used to train the SSAT-GAN over the UP dataset. Each category of HSI cubes with the false color and their corresponding feature maps from the penultimate layer of the discriminator are shown in Fig. 16. As illustrated in Fig. 16, the more significant the features, the darker the gradient distribution of the attention. However, some target pixels to be classified in the fact do not exactly belong to the same category as their neighboring pixels in their corresponding HSI cubes, such as the central pixels with its surroundings in Fig. 16(c), (e), and (g). In contrast, it can be seen that there also consist of some bright areas in the corresponding mixed attention distribution. The reason is that our SSAT modules can effectively activate the spectral–spatial attribute and assign an independent attention weight for each pixel of the HSI cubes. In this case, our SSAT-GAN performs the guidance of the attention weights with the feature extraction simultaneously, which can improve the efficiency of hyperspectral characterization.

#### E. Execution Time Analysis on Different Datasets

The training and testing time on the three datasets are also illustrated in Tables II–IV. To assess the computational complexity, we reported the execution time (in *milliseconds*, *i.e.*, *ms*) at each epoch or iteration of various methods.

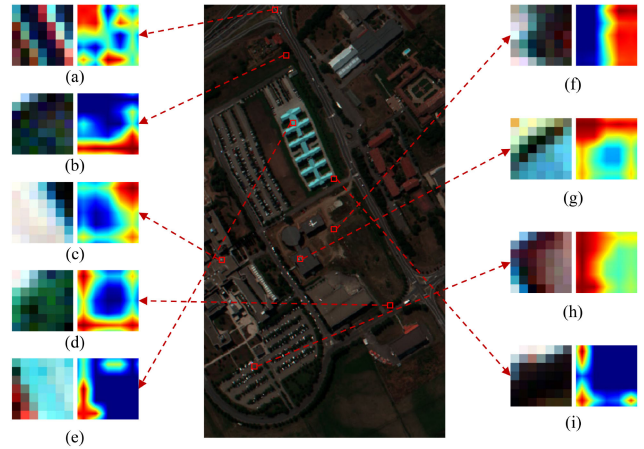


Fig. 16. Feature visualization with the guidance of the attention weights over SSAT on the UP dataset. Each land cover category is randomly selected from the labeled training set and described with the false color. The corresponding feature visualization is obtained by applying Grad-CAM [49].

In general, EPF-SVM obviously consumed the shortest time for training in all three cases. 3D-Conv-Capsule took the longest time as the reason it needs to construct a dynamic route for optimal vector search during training. GAN-based methods need to optimize the discriminator and the generator alternately and, thus, gathered a relatively long time for training. In addition, Spa-AT-GAN took the shortest time to train on all three datasets, among the deep learning methods, which took about 4–6 times faster than GAN-CRF. In contrast, we can find that the time cost is relatively similar between SSAT-GAN and HSI-Bert, while our SSAT-GANs contains better accuracy as illustrated in Tables II–IV.

For testing, 3D-GAN took more time to test because of its large candidate neighboring areas and the deep network architecture. In contrast, SSAT-GAN consumes relatively less time due to the high efficiency that existed in the feature representation of the spectral–spatial attention spread and generation blocks. In summary, it can be concluded that the proposed framework is the most efficient method with advanced performance under fair comparison.

TABLE IV  
CLASSIFICATION ACCURACIES AND TRAINING AND TESTING TIMES OF VARIOUS COMPARISON METHODS USING 250 LABELED SAMPLES AND 250 UNLABELED SAMPLES FOR THE KSC DATASET

Class	Train.(Test.)	EPF-SVM	SSRN	3D-Conv-Capsule	HSI-BERT	3D-GAN	GAN-CRF	AD-GAN	Spa-AT-GAN	Spc-AT-GAN	Spa-Spc-AT-GAN	SSAT-GAN
1	36 (725)	<b>100.00±0.00</b>	96.90±2.55	99.92±0.01	99.84±0.16	98.19±0.72	96.10±1.32	98.29±0.79	98.05±0.88	<b>100.00±0.00</b>	96.81±0.18	98.06±1.11
2	11 (232)	<b>100.00±0.00</b>	98.69±0.28	77.60±3.52	90.34±5.94	95.19±0.52	<b>100.00±0.00</b>	96.08±1.68	99.45±0.34	98.95±0.45	87.13±2.04	99.50±0.16
3	12 (244)	<b>100.00±0.00</b>	88.87±1.11	91.50±3.75	80.81±6.84	76.44±7.79	<b>99.72±1.58</b>	85.08±2.13	93.64±1.42	94.59±0.98	52.27±8.89	90.94±3.42
4	12 (240)	77.78±5.83	74.93±2.11	54.42±3.11	50.54±8.69	62.54±6.73	<b>99.12±0.08</b>	69.16±4.28	80.90±2.36	71.78±3.02	60.14±4.44	91.34±2.10
5	10 (151)	53.99±10.57	66.02±2.54	95.24±1.98	61.78±11.59	82.85±1.55	58.10±10.45	85.09±2.77	83.54±1.68	87.85±1.25	60.00±3.66	<b>90.97±1.49</b>
6	11 (218)	97.77±1.39	96.35±0.61	99.13±0.77	82.06±11.19	93.88±0.22	96.80±1.24	83.62±0.88	87.06±2.48	98.55±0.25	82.67±0.76	<b>100.00±0.00</b>
7	6 (99)	93.33±2.10	88.52±1.50	97.24±1.36	96.16±4.66	97.43±1.67	84.21±0.86	97.43±1.27	69.46±3.79	78.44±4.26	85.98±1.92	<b>97.78±0.68</b>
8	20 (411)	74.72±2.09	95.77±0.53	<b>98.77±1.78</b>	98.73±1.37	88.88±0.28	96.61±2.36	98.39±0.44	96.72±1.22	98.46±0.66	97.69±0.79	92.94±2.14
9	24 (496)	82.27±7.38	99.54±0.97	95.40±2.23	99.41±0.27	96.12±0.69	<b>100.00±0.00</b>	99.35±0.25	99.78±0.02	98.28±0.62	<b>100.00±0.00</b>	<b>100.00±0.00</b>
10	19 (385)	90.33±1.94	99.80±0.38	92.21±5.70	<b>100.00±0.00</b>	98.83±0.28	97.60±0.76	<b>100.00±0.00</b>	98.90±0.10	94.45±0.55	<b>100.00±0.00</b>	<b>100.00±0.00</b>
11	21 (398)	95.36±3.76	99.40±0.90	<b>100.00±0.00</b>	99.84±0.25	97.48±1.18	94.18±2.24	<b>100.00±0.00</b>	98.94±1.49	99.47±0.14	96.89±1.08	96.79±1.31
12	24 (479)	<b>100.00±0.00</b>	99.00±0.01	97.93±1.73	98.95±1.15	98.09±0.39	<b>100.00±0.00</b>	99.87±0.07	97.48±1.17	99.52±0.20	98.39±1.06	<b>100.00±0.00</b>
13	44 (883)	<b>100.00±0.00</b>	99.95±0.14	<b>100.00±0.00</b>	<b>100.00±0.00</b>	98.86±0.98	<b>100.00±0.00</b>	<b>100.00±0.00</b>	99.52±0.05	99.87±0.10	<b>100.00±0.00</b>	99.50±0.14
OA(%)		91.31±3.14	94.19±2.90	94.75±0.45	93.88±0.59	93.38±1.32	95.38±1.86	96.15±1.44	95.67±1.55	96.42±0.99	94.63±2.10	<b>97.72±1.11</b>
AA(%)		89.66±3.01	92.60±2.89	92.41±0.46	99.11±0.99	91.14±2.32	93.62±1.28	93.50±1.50	92.54±0.64	93.98±1.47	93.55±1.25	<b>96.75±0.95</b>
$\kappa \times 100$		90.33±3.49	93.53±3.23	94.15±0.51	93.18±0.66	92.63±2.23	94.86±2.16	95.72±1.28	95.18±1.19	96.02±0.99	93.90±1.42	<b>97.08±1.82</b>
Train. Time(ms)		218.91	684.01	3770.44	1194.08	1946.93	2768.10	1691.16	355.98	462.28	663.21	990.91
Test. Time(ms)		1014.41	2602.88	2916.12	4529.00	10809.45	7039.77	7512.25	1178.84	1496.23	1596.17	2094.20

Note: The best values are highlighted in bold font.

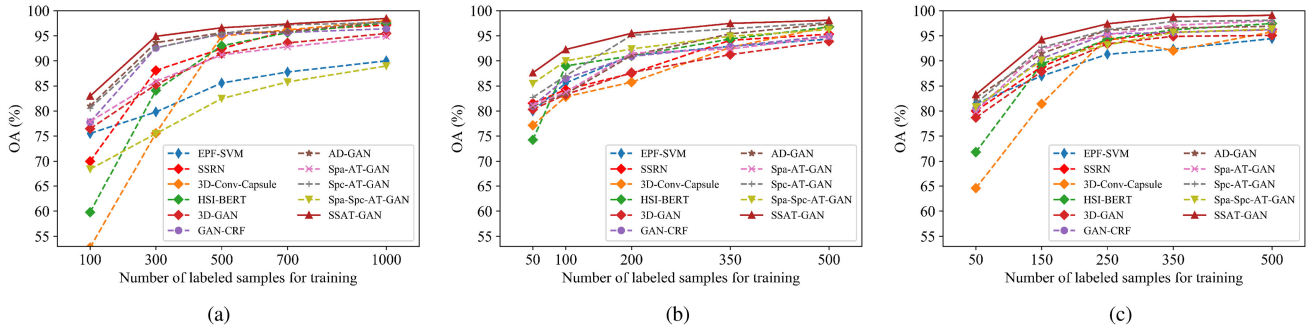


Fig. 17. Impact of different number of labeled samples on OA results for training. OA results were obtained by all algorithms on (a) IN dataset, (b) UP dataset, and (c) KSC dataset.

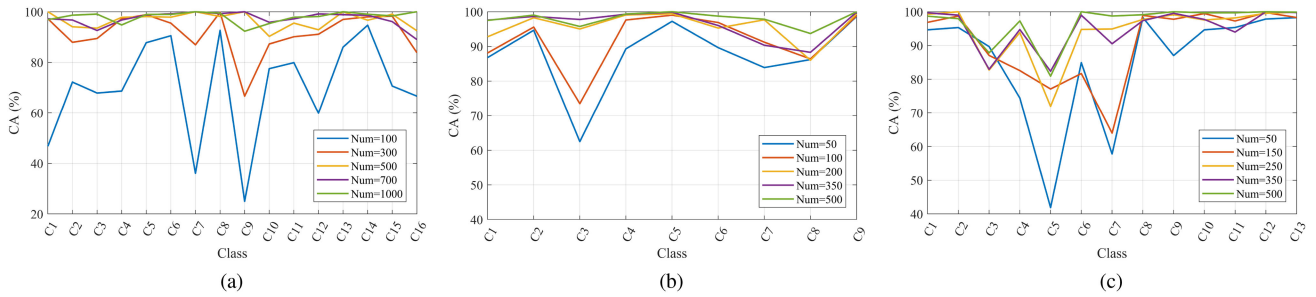


Fig. 18. Class accuracy results for each class with different number of total labeled samples for training over the SSAT-GAN on (a) IN dataset, (b) UP dataset, and (c) KSC dataset.

#### F. Sensitivity Analysis on Different Number of Labeled Sample for Training

To observe the effect of different number of labeled samples on OAs, we randomly selected labeled pixels in the range of  $\{100, 300, 500, 700, 1000\}$  on the IN,  $\{50, 100, 200, 350, 500\}$  on the UP, and  $\{50, 150, 250, 350, 500\}$  on the KSC with the Monte Carlo sampling strategy. Fig. 17(a)–(c) reports the OAs of all competitors on three datasets, respectively. It should be noted that the OAs gradually increase and then stabilize under a different number of labeled samples on the IN, UP, and KSC datasets. The reason is that the Monte Carlo sampling strategy can provide sufficient labeled samples and, thus, construct a complete dictionary for training. In addition, SSAT-GAN has an obvious advantage in classification performance in contrast to other methods.

To verify the contribution of different categories to both AA and kappa with modification of labeled samples for training, a new experiment was performed on the three datasets over the proposed SSAT-GAN. Fig. 18 illustrated the CA of each class on the three datasets. It is observed that the SSAT-GAN acquires stable CAs for “Grass-trees,” “Hay-windowed,” and “Woods”

class, no matter how many amounts of total labeled samples are considered, owing to the discriminative spectral characteristics of three ground materials in the IN dataset. Therefore, it still achieves a satisfactory classification performance, even in a relatively small labeled samples. Furthermore, the contribution to the CA of the remaining classes in the IN dataset is improved and then stabilizes since the number of labeled samples increases.

For the UP dataset illustrated in Fig. 18(b), for the “Meadows,” “Painted metal sheets,” and “Shadows” class, the CAs detail a negligible variation as the labeled samples increase. As for other classes, the accuracy values of the proposed SSAT-GAN tend to be stabilized as the number of samples increases. Similar achievements can be found in Fig. 18(b) and (c). Overall, not all the classes contribute to both AA and kappa to the same degree with modification of labeled training samples. The reason may be that the spectral signatures suffer from the challenge of spectral variability which stems from the illumination and atmospheric conditions. However, our SSAT modules can alleviate such limitations of spectral characteristics, which can be illustrated at those advanced accuracy values in the three datasets.

#### V. DISCUSSION

There are three differences between the proposed SSAT-GAN and the GAN-based methods for HSI classification [29], [32], [33]. First, SSAT-GAN takes the attention information of HSIs into account for both the discriminator and the generator. Second, the discriminator in the adversarial framework adds unlabeled samples for semisupervised learning and alleviates

TABLE V  
OVERALL ACCURACIES (%) OF SEMISUPERVISED GAN METHODS WITH VARYING ATTENTION MODULES USING 300 UNLABELED SAMPLES AND LABELED SAMPLES ON THE IN, UP, AND KSC DATASETS

Datasets	SE_Block	CBAM	FA	MAFN	SSAT
IN	90.82	91.39	86.65	87.99	<b>94.04</b>
UP	96.49	96.17	97.17	97.11	<b>97.85</b>
KSC	96.27	96.53	97.75	96.51	<b>98.08</b>

the impact of small samples. Third, a mean minimization loss is employed for the unsupervised learning of SSAT-GAN to reduce the complex calculation parameters of high-dimensional features so as to achieve steady-state performance of GAN.

The SSAT-GAN models incorporate the SSAT as the feature perception enhancement step in the feature extraction stage, which builds a strong SNR spectral domain and a physical denoising contextual area upon both spectral and spatial dimensions, respectively. Compared with those attention mechanisms used in the vision community, the SSAT considers the long-range correlations between neighboring HSI cubes. This property helps the SSAT-GAN framework to better filter noises in the areas with different spectral purity and texture information.

We gain three major insights from the semisupervised HSI classification outcomes of GANs in all three datasets. First, by taking the spectral–spatial discriminative features of training data into account, the discriminators of SSAT-GANs extract efficient and significant HSI characteristics and achieve better classification accuracies. Second, the unlabeled samples and generated HSI samples of unsupervised learning make discriminators more robust among adversarial framework and learning complex real data distribution of HSIs to predict. This alternate training mode enables semisupervised GANs to promote superior classification outcomes than that of supervised deep learning derived frameworks. Third, the mean minimization loss takes the constrained optimization of the high-dimensional feature maps generated by the discriminators as the smooth filtering by calculating the efficiency values, which imposes the correlation in homogeneous regions including high texture areas or purity spectral domain.

## VI. CONCLUSION

In this article, an SSAT-GAN approach for HSI classification is proposed by using a cascade feature representation of spectral–spatial attributes with the SSAT. The proposed model improves the transmission of the characteristics with extended spectral–spatial attention feature spread and generation blocks to represent the feature. It effectively applied the attention weights to emphasize both spectral bands and spatial correlations to improve the characterization during feature extraction. Besides, SSAT-GAN constructs a semisupervised architecture by adding unlabeled samples for training to alleviate the scarcity of training samples. Furthermore, we employ the mean minimization loss for unsupervised learning of the discriminator to avoid the mode collapse. In terms of the accuracy and computation of the experiments, an analysis on the three HSI datasets indicates that our model achieves an excellent performance.

## ACKNOWLEDGMENT

The authors would like to thank the Associate Editor and the three anonymous reviewers for their outstanding comments and suggestions, which greatly helped the authors to improve the technical quality and presentation of this article.

## REFERENCES

- [1] H. Li, G. Xiao, T. Xia, Y. Y. Tang, and L. Li, "Hyperspectral image classification using functional data analysis," *IEEE Trans. Cybern.*, vol. 44, no. 9, pp. 1544–1555, Sep. 2014.
- [2] S. Le Mouelic *et al.*, "An iterative least squares approach to decorrelate minerals and ices contributions in hyperspectral images: Application to cuprite (Earth) and Mars," in *Proc. 1st Workshop Hyperspectral Image Signal Process., Evol. Remote Sens.*, 2009, pp. 1–4.
- [3] M. Teke, H. S. Deveci, O. Haliloğlu, S. Z. Gürbüz, and U. Sakarya, "A short survey of hyperspectral remote sensing applications in agriculture," in *Proc. 6th Int. Conf. Recent Adv. Space Technol.*, 2013, pp. 171–176.
- [4] J. Zhou, C. Kwan, B. Ayhan, and M. T. Eismann, "A novel cluster kernel RX algorithm for anomaly and change detection using hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 11, pp. 6497–6504, Nov. 2016.
- [5] X. Huang and L. Zhang, "An adaptive mean-shift analysis approach for object extraction and classification from urban hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 12, pp. 4173–4185, Dec. 2008.
- [6] Y. Yuan, J. Lin, and Q. Wang, "Hyperspectral image classification via multitask joint sparse representation and stepwise MRF optimization," *IEEE Trans. Systems, Man, Cybern.*, vol. 46, no. 12, pp. 2966–2977, Dec. 2016.
- [7] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Spectral-spatial classification of hyperspectral data using loopy belief propagation and active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 844–856, Feb. 2013.
- [8] J. Yang, D. Zhang, A. F. Frangi, and J.-Y. Yang, "Two-dimensional PCA: A new approach to appearance-based face representation and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 1, pp. 131–137, Jan. 2004.
- [9] C.-I. Chang and S. Wang, "Constrained band selection for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 6, pp. 1575–1585, Jun. 2006.
- [10] L. Zhang, L. Zhang, D. Tao, and X. Huang, "On combining multiple features for hyperspectral remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 3, pp. 879–893, Mar. 2012.
- [11] J. Xia, P. Ghamisi, N. Yokoya, and A. Iwasaki, "Random forest ensembles and extended multixinction profiles for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 202–216, Jan. 2018.
- [12] X. Kang, S. Li, and J. A. Benediktsson, "Spectral-spatial hyperspectral image classification with edge-preserving filtering," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2666–2677, May 2014.
- [13] J. Jiang, J. Ma, C. Chen, Z. Wang, Z. Cai, and L. Wang, "SuperPCA: A superpixelwise PCA approach for unsupervised feature extraction of hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4581–4593, Aug. 2018.
- [14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
- [15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [16] 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.
- [17] 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.
- [18] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectral-spatial residual network for hyperspectral image classification: A 3-D deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 847–858, Feb. 2018.
- [19] W. Wang, S. Dou, Z. Jiang, and L. Sun, "A fast dense spectral-spatial convolution network framework for hyperspectral images classification," *Remote Sens.*, vol. 10, no. 7, 2018, Art. no. 1068.
- [20] K. Zhu, Y. Chen, P. Ghamisi, X. Jia, and J. A. Benediktsson, "Deep convolutional capsule network for hyperspectral image spectral and spectral-spatial classification," *Remote Sens.*, vol. 11, no. 3, pp. 223–250, 2019.
- [21] A. Sellami and I. Farah, "Spectra-spatial graph-based deep restricted Boltzmann networks for hyperspectral image classification," in *Proc. Photon. Electromagn. Res. Symp.-Spring*, 2019, pp. 1055–1062.

- [22] R. Xu, Y. Tao, Z. Lu, and Y. Zhong, "Attention-mechanism-containing neural networks for high-resolution remote sensing image classification," *Remote Sens.*, vol. 10, no. 10, 2018, Art. no. 1602.
- [23] H. Sun, X. Zheng, X. Lu, and S. Wu, "Spectral-spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3232–3245, May 2020.
- [24] Z. Li *et al.*, "Hyperspectral image classification with multiattention fusion network," *IEEE Geosci. Remote Sens. Lett.*, to be published, doi: [10.1109/LGRS.2021.3052346](https://doi.org/10.1109/LGRS.2021.3052346).
- [25] C. Yu, R. Han, M. Song, C. Liu, and C.-I. Chang, "Feedback attention-based dense CNN for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: [10.1109/TGRS.2021.3058549](https://doi.org/10.1109/TGRS.2021.3058549).
- [26] J. He, L. Zhao, H. Yang, M. Zhang, and W. Li, "HSI-BERT: Hyperspectral image classification using the bidirectional encoder representation from transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 165–178, Jan. 2020.
- [27] S. Fang, D. Quan, S. Wang, L. Zhang, and L. Zhou, "A two-branch network with semi-supervised learning for hyperspectral classification," in *Proc. IGARSS IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 3860–3863.
- [28] Z. Lei, Y. Zeng, P. Liu, and X. Su, "Active deep learning for hyperspectral image classification with uncertainty learning," *IEEE Geosci. Remote Sens. Lett.*, to be published, doi: [10.1109/LGRS.2020.3045437](https://doi.org/10.1109/LGRS.2020.3045437).
- [29] Y. Zhan, D. Hu, Y. Wang, and X. Yu, "Semisupervised hyperspectral image classification based on generative adversarial networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 212–216, Feb. 2018.
- [30] Y. Zhan *et al.*, "Semi-supervised classification of hyperspectral data based on generative adversarial networks and neighborhood majority voting," in *Proc. IGARSS IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 5756–5759.
- [31] Z. He, H. Liu, Y. Wang, and J. Hu, "Generative adversarial networks-based semi-supervised learning for hyperspectral image classification," *Remote Sens.*, vol. 9, no. 10, 2017, Art. no. 1042.
- [32] 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.
- [33] Z. Zhong, J. Li, D. A. Clausi, and A. Wong, "Generative adversarial networks and conditional random fields for hyperspectral image classification," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3318–3329, Jul. 2020.
- [34] J. Wang, F. Gao, J. Dong, and Q. Du, "Adaptive dropblock-enhanced generative adversarial networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5040–5053, Jun. 2021.
- [35] B. Sui, T. Jiang, Z. Zhang, and X. Pan, "ECGAN: An improved conditional generative adversarial network with edge detection to augment limited training data for the classification of remote sensing images with high spatial resolution," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1311–1325, Oct. 2021.
- [36] H. Liang, W. Bao, and X. Shen, "Adaptive weighting feature fusion approach based on generative adversarial network for hyperspectral image classification," *Remote Sens.*, vol. 13, no. 2, pp. 198–222, 2021.
- [37] J. Feng *et al.*, "Generative adversarial networks based on collaborative learning and attention mechanism for hyperspectral image classification," *Remote Sens.*, vol. 12, no. 7, 2020, Art. no. 1149.
- [38] Y. Cai, Z. Dong, Z. Cai, X. Liu, and G. Wang, "Discriminative spectral-spatial attention-aware residual network for hyperspectral image classification," in *Proc. 10th Workshop Hyperspectral Imag. Signal Process.: Evol. Remote Sens.*, 2019, pp. 1–5.
- [39] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
- [40] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
- [41] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1251–1258.
- [42] X. Cui, K. Zheng, L. Gao, B. Zhang, D. Yang, and J. Ren, "Multiscale spatial-spectral convolutional network with image-based framework for hyperspectral imagery classification," *Remote Sens.*, vol. 11, no. 19, 2019, Art. no. 2220.
- [43] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.
- [44] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*.
- [45] Y. Saatchi and A. G. Wilson, "Bayesian GAN," *Adv. Neural Inf. Process. Syst.*, vol. 2017, pp. 3623–3632, 2017.
- [46] G. Hinton, "Neural networks for machine learning online course lecture 6a," Coursera. [Online]. Available: [http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture\\_slides\\_lec6.pdf](http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf)
- [47] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
- [48] J. Bai, R. Chen, and M. Liu, "Feature-attention module for context-aware image-to-image translation," *Vis. Comput.*, vol. 36, no. 10, pp. 2145–2159, 2020.
- [49] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.



**Hongbo Liang** received the B.S. degree in computer science and technology from North Minzu University, Yinchuan, China, 2018, and the M.S. degree in computer technology from the School of Computer Science and Engineering, North Minzu University, in 2021.

His research interests include hyperspectral image processing, remote sensing image classification, and deep learning.



**Wenxing Bao** received the B.Eng. degree in industrial automation from Xidian University, Xi'an, China, in 1993, and the M.Sc. degree in electrical engineering and the Ph.D. degree in electronic science and technology from Xi'an Jiaotong University, Xi'an, China, in 2001 and 2006, respectively.

He is currently a Professor and Vice President of North Minzu University, Yinchuan, China. His research interests include digital image processing, remote sensing image classification, and fusing.



**Xiangfei Shen** received the B.S. degree in computer science and technology from North Minzu University, Yinchuan, China, in 2017, and the M.S. degree in computer science and technology from North Minzu University, Yinchuan, China, in 2020. He is currently working toward the Ph.D. degree in information and communication engineering at the school of microelectronics and communication engineering, Chongqing University, Chongqing, China.

His research interests include hyperspectral image processing, pattern recognition, and machine learning.



**Xiaowu Zhang** received the B.Eng. degree in computer science and technology from Yulin University, Yulin, China, in 2009, and the M.Sc. degree in computer application technology from North Minzu University, Yinchuan, China, in 2012.

He is currently a Lecturer with North Minzu University. His research interests include hyperspectral image classification, computer vision, machine learning, and deep learning.