Self-Attention Context Network: Addressing the Threat of Adversarial Attacks for Hyperspectral Image Classification

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

***Abstract*— Deep learning models have shown their great capability for the hyperspectral image (HSI) classification task in recent years. Nevertheless, their vulnerability towards adversarial attacks could not be neglected. In this study, we systematically analyze the influence of adversarial attacks on the HSI classification task for the first time. While existing research of adversarial attacks focuses on the generation of adversarial examples in the RGB domain, the experiments in this study show such adversarial examples could also exist in the hyperspectral domain. Although the difference between the generated adversarial image and the original hyperspectral data is imperceptible to the human visual system, most of the existing state-of-the-art deep learning models could be fooled by the adversarial image to make wrong predictions. To address this challenge, a novel self-attention context network (SACNet) is further proposed. We discover that the global context information contained in HSI can significantly improve the robustness of deep neural networks when confronted with adversarial attacks. Extensive experiments on three benchmark HSI datasets demonstrate that the proposed SACNet possesses stronger resistibility towards adversarial examples compared with the existing state-of-the-art deep learning models.**

***Index Terms*— Hyperspectral image (HSI) classification, adversarial example, adversarial attack, adversarial defense, convolutional neural network (CNN), deep learning.**

I. Introduction

DEEP neural networks, which can hierarchically extract features [1], have achieved state-of-the-art performance in the hyperspectral image (HSI) classification task [2]–[4]. Compared with traditional hand-crafted methods that rely on expert knowledge from designers [5], [6], deep learning models can learn discriminative features from the input image automatically without much human intervention [7], [8].

Chen *et al.* introduced the concept of deep learning into the hyperspectral community for the first time [9]. They proposed a spectral-spatial classification framework based on the stacked auto-encoders (SAEs) using a layer-wise training mechanism. In a like manner, the restricted Boltzmann machine (RBM) and the deep belief network (DBN) are adopted to simultaneously extract spectral and spatial features from hyperspectral data [10]. Considering the inherent sequential property contained in HSI, Mou *et al.* introduced the recurrent neural network (RNN) and long short-term memory (LSTM) model for deep spectral feature extraction [11]. Since there are too many wavebands in HSI, directly feeding the whole spectral vector into the RNN would make it difficult to optimize the network. To this end, Xu *et al.* further proposed two novel band grouping strategies for RNN [12].

Apart from the aforementioned methods, convolutional neural networks (CNNs) deserve our special attention because of their powerful feature representation capability [13], [14]. Zhao *et al.* proposed a spectral-spatial feature extraction method for HSI based on 2D-CNNs and balanced local discriminant embedding [15]. To directly extract spectral-spatial features from the hyperspectral cube, 3D-CNN-based methods were further proposed [16]–[18], which could be trained in an end-to-end manner. Paoletti *et al.* proposed a novel deep pyramidal residual networks for HSI classification [19]. The proposed pyramid architecture helps to gradually increase the feature map dimension at different convolutional layers, achieving state-of-the-art performance. Xu *et al.* further proposed a spectral-spatial fully convolutional network (SSFCN) for HSI classification [20]. Different from previous CNN-based methods that the input data of the network is a spatial patch with a size of $w \times w$, SSFCN can directly process the whole hyperspectral cube without patch extraction.

Although deep learning models have achieved great success in the remote sensing field [21], their vulnerability towards adversarial examples should not be neglected [22]. Szegedy *et al.* first discovered that deep neural networks are very fragile to adversarial examples [23]. Such adversarial examples can be simply generated by adding subtle adversarial perturbation to the original clean image. Although these adversarial examples may look the same as the original images for

Manuscript received December 6, 2020; revised June 19, 2021 and August 10, 2021; accepted September 16, 2021. Date of publication October 14, 2021; date of current version October 22, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61822113, Grant 41820104006, Grant 61871299, and Grant 41871243; and in part by the Science and Technology Major Project of Hubei Province (Next-Generation AI Technologies) under Grant 2019AEA170. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Yannick Berthoumieu. *(Corresponding authors: Liangpei Zhang; Bo Du.)*

Yonghao Xu and Liangpei Zhang are with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing and the Institute of Artificial Intelligence, Wuhan University, Wuhan 430079, China (e-mail: yonghaoxu@ieee.org; zlp62@whu.edu.cn).

Bo Du is with the National Engineering Research Center for Multimedia Software, School of Computer Science, Institute of Artificial Intelligence, Wuhan University, Wuhan 430079, China, and the Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan 430079, China (e-mail: dubo@whu.edu.cn).

Digital Object Identifier 10.1109/TIP.2021.3118977

a human observer, most of the state-of-the-art deep neural networks tend to misclassify these samples into a wrong category with very high confidence [24]. To make the adversarial attacks more efficient, Goodfellow *et al.* proposed the fast gradient sign method (FGSM) to generate adversarial perturbation [25]. Arnab *et al.* presented the first rigorous evaluation of adversarial attacks on modern semantic segmentation models [26]. He *et al.* explored the influence of adversarial attacks on biomedical image segmentation task with FGSM [27].

While the aforementioned research focuses on adversarial examples in the computer vision field, there are also some preliminary analyses about adversarial examples in geoscience and remote sensing. Czaja *et al.* first revealed that adversarial examples also exist in the satellite remote sensing image classification task [28]. They found that adversarial attacks on a small patch inside the remote sensing image could fool the deep learning models to make a wrong prediction. They also provided some practical considerations that an attacker would need to take into account when mounting physical adversarial attacks. Chen *et al.* showed that the sensitivity of deep neural networks to adversarial attacks is also related to the remote sensing dataset [29]. Generally, datasets with a higher diversity tend to be less vulnerable to adversarial attacks. Xu *et al.* systematically analyzed the influence of adversarial examples on deep neural networks for remote sensing scene classification task [22]. They found that adversarial examples generated from a specific deep learning model could also be detrimental to other deep neural networks or even traditional shallow methods like support vector machine (SVM) and k-nearest neighbors (KNN). They also investigated the adversarial training strategy to conduct a preliminary adversarial defense on the generated adversarial examples.

So far, researchers have put much attention on the adversarial examples in the RGB domain [30]. Whether adversarial examples also exist in the hyperspectral domain has not been explored yet. Compared to traditional image classification tasks in the computer vision field, the training data size in HSI classification is much smaller [31]. Under this circumstance, the trained deep learning models are more likely to be over-fitting and susceptible to noise. Thus, the threat of adversarial attacks may be even more serious. An illustration of adversarial attacks on deep neural networks for HSI classification is shown in Fig. 1. Here, we use the SSFCN [20] as the deep learning model for example. As shown in Fig. 1, although the difference between the original HSI and the generated adversarial example is hard to perceive for the human visual system, the state-of-the-art deep neural network like SSFCN could be fooled to make wrong predictions. While SSFCN can achieve an OA of 93.39% on the original clean image, its OA on the adversarial image decreases to only 19.12%. This phenomenon would undoubtedly pose a threat to the deployed deep learning models for HSI classification.

One possible way to defend from such attacks is adversarial training, where the generated adversarial examples are adopted to extend the original training set [22]. However, there are two limitations to this strategy. First, adversarial training requires additional training samples, which increases the computation burden. Second, since adversarial training

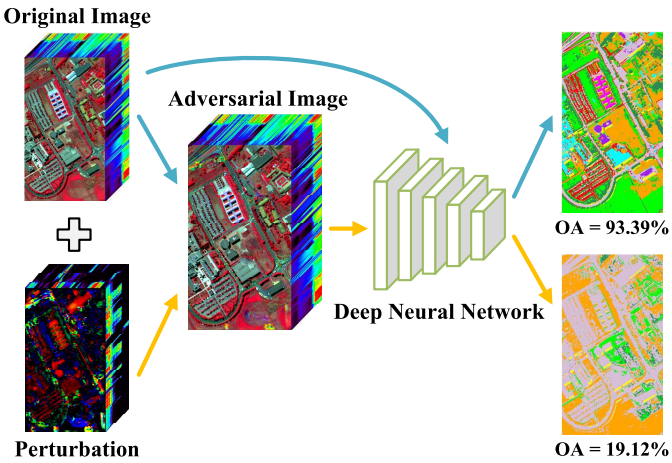

Fig. 1. An illustration of adversarial attacks on deep neural networks for hyperspectral image classification task. Although the difference between the generated adversarial image and the original hyperspectral image is imperceptible to the human visual system, the state-of-the-art deep learning model could be fooled by the adversarial image to make wrong predictions.

can hardly improve the inherent robustness of deep neural networks, the trained model may get attacked again with newly generated adversarial examples [30]. Instead of using the adversarial training scheme, we want to improve the network's intrinsic resistibility towards adversarial examples by well-designed network architecture. To achieve this goal, a novel self-attention context network (SACNet) for HSI classification is proposed. Compared with local feature extraction, the extraction of global context information requires building relationships with all related pixels in the whole image for a given pixel. In this way, the produced prediction by the network at this pixel would be affected by its related pixels. Under such a circumstance, if an incorrect label was assigned to this pixel by the adversarial attack, the incorrect loss at this pixel would also be passed to all other related pixels through back-propagation. Thus, the total loss at this pixel would be shared by all other related pixels, and attacking such networks may require a higher level of perturbation. The main contributions of this study are summarized as follows.

1) We introduce the concept of adversarial attack into the hyperspectral remote sensing community for the first time. Our research reveals the significance of the resistibility and robustness of deep learning models when addressing the safety-critical hyperspectral remote sensing tasks.
2) We systematically analyze the characteristics of adversarial examples in the hyperspectral domain. Experiments on three benchmark HSI classification datasets demonstrate that most of the existing state-of-the-art deep learning models could be seriously fooled by adversarial examples to make wrong classification maps.
3) To defense against adversarial attacks, a novel self-attention context network (SACNet) is further proposed. While most of the existing deep learning models focus on local feature extraction, the proposed SACNet can capture the global context information by self-attention learning and context encoding. We discover that such

global context information contained in HSI can significantly improve the resistibility of deep neural networks towards adversarial attacks for the HSI classification task. The experimental results show the proposed SACNet can achieve superior performance on the adversarial test set compared with existing methods.

The rest of this paper is organized as follows. Section II reviews the adversarial attack and defense algorithms. Section III describes the proposed SACNet in detail. Section IV presents the information on datasets used in this study and the experimental results. Conclusions and other discussions are summarized in Section V.

II. Related Works

In this section, we will make a brief review of the existing adversarial attack and defense methods.

A. Adversarial Attacks

1) Box-Constrained L-BFGS: Szegedy *et al.* first found an intriguing property of deep neural networks that applying specific perturbations to an image may fool the network to make wrong predictions [23]. Such perturbations can be generated by maximizing the network's prediction error.

Formally, let $f : x \in \mathbb{R}^n \rightarrow y \in \mathbb{L}$ be the mapping function of a deep neural network that maps an image with n pixels into a discrete label set. Given an image x and a target label $\hat{y}$, where $\hat{y}$ denotes the wrong label that we expect the network would predict, the adversarial perturbation ρ can be generated by solving the box-constrained optimization problem as below:

$$\min_{\rho} \|\rho\|_2, \quad subject\ to : \begin{cases} f(x+\rho) = \hat{y} \\ x + \rho \in [0,1]^n . \end{cases} \tag{1}$$

Note that (1) is nontrivial only when $\hat{y} \neq y$, where y denotes the true label of x. Otherwise, we will simply get $\rho = 0$. In general, directly solving (1) is a hard problem. Szegedy *et al.* proposed to approximate the solution of (1) using a box-constrained L-BFGS. More concretely, they performed line-search to find the minimum $c > 0$ for which the minimizer ρ of the following optimization problem satisfies $f(x+\rho) = \hat{y}$, and $\hat{y} \neq y$:

$$\begin{aligned} &\min_{\rho} c\|\rho\|_2 + J\left(\theta, x+\rho, \hat{y}\right), \\ &subject\ to : x + \rho \in [0,1]^n , \end{aligned} \tag{2}$$

where θ represents the parameters in the deep neural network, and $J(\cdot)$ denotes the loss function used for training the network (e.g., the cross-entropy loss).

2) Fast Gradient Sign Method: The optimization of (2) is still very difficult in practice. To implement more efficient adversarial attacks, Goodfellow *et al.* proposed the fast gradient sign method (FGSM). Given an image x and a target label $\hat{y}$, the adversarial example x_{adv} can be calculated as:

$$\begin{cases} \rho = \epsilon\ \text{sign}\left(\nabla_x J\left(\theta, x, \hat{y}\right)\right) \\ x_{adv} = \text{clip}\left(x - \rho\right), \end{cases} \tag{3}$$

where $\nabla_x J\left(\theta, x, \hat{y}\right)$ calculates the gradients of the loss function $J(\cdot)$ with respect to the input sample x, sign $(\cdot)$ denotes the sign function, clip $(\cdot)$ clips the pixel values in the image, and ϵ is a small scalar value that controls the norm of the perturbation.

Miyato *et al.* [32] and Kurakin *et al.* [33] further modified FGSM by applying the ℓ_2 norm and ℓ_∞ norm to the generated perturbation:

$$\ell_2 :\ x_{adv} = \text{clip}\left(x - \epsilon \frac{\nabla_x J\left(\theta, x, \hat{y}\right)}{\|\nabla_x J\left(\theta, x, \hat{y}\right)\|_2}\right). \tag{4}$$

$$\ell_\infty :\ x_{adv} = \text{clip}\left(x - \epsilon \frac{\nabla_x J\left(\theta, x, \hat{y}\right)}{\|\nabla_x J\left(\theta, x, \hat{y}\right)\|_\infty}\right). \tag{5}$$

3) Basic Iterative Method/Projected Gradient Method: The basic iterative method (BIM) was first proposed by Kurakin *et al.*, which is an iterative version of FGSM [33]. At each iteration, the adversarial example can be updated as below:

$$x_{adv}^{t+1} = \text{clip}\left(x_{adv}^{t} - \epsilon\ \text{sign}\left(\nabla_{x_{adv}} J\left(\theta, x_{adv}^{t}, \hat{y}\right)\right)\right). \tag{6}$$

When $t = 0$, x_{adv}^{0} is initialized with the original clean image x. Madry *et al.* further pointed out that BIM is equivalent to the projected gradient descent (PGD) method if the algorithm is added by a random initialization on x [34].

4) Carlini and Wagner's Attack (C&W): In [35], Carlini and Wagner proposed to conduct adversarial attacks by encouraging x_{adv} to have a larger probability score for the target class $\hat{y}$ than other classes. Their experiments show that C&W can achieve successful attacks with imperceptible adversarial perturbations. A more comprehensive review of this method can be found in [36].

B. Adversarial Defenses

To defend against the aforementioned adversarial attacks, different strategies have been considered to conduct adversarial defense. A good review of this research topic can be found in [30]. Generally, existing adversarial defense methods can be categorized into three groups: improved training scheme, modified network architecture, and adversarial example detection.

1) Improved Training Scheme: One of the main reasons that adversarial examples could mislead deep neural networks to make wrong predictions is because of their absence in the training phase. A straightforward idea to conduct adversarial defense is thereby to train the deep neural network with both the original clean training samples and the generated adversarial examples. This training scheme is known as adversarial training [25].

Formally, the loss function of the adversarial training $\hat{J}(\theta, x, y)$ can be formulated as:

$$\hat{J}(\theta, x, y) = \lambda J(\theta, x, y) + (1-\lambda) J(\theta, x_{adv}, y), \tag{7}$$

where λ is a weighting factor to balance the optimization of the loss on clean samples and adversarial examples. Goodfellow *et al.* found adversarial training could not only help to resist adversarial attacks but also regularize the network [25].

Zantedeschi *et al.* further proposed to generalize the adversarial training with Gaussian noise perturbed samples [37]:

$$\hat{J}(\theta, x, y) = \mathbb{E}_{\Delta x \sim \mathcal{N}(0,\sigma^2)} J(\theta, x + \Delta x, y), \tag{8}$$

where σ denotes the acceptable non-perceivable perturbation. Despite its simplicity, (8) is proven to be useful in improving the robustness of neural networks to adversarial attacks [37].

Another possible training scheme is ensemble adversarial training, where adversarial examples crafted on other static pre-trained models are adopted to augment the training set of an undefended model [38].

2) Modified Network Architecture: Since adversarial training doesn't improve the inherent resistibility of the model, the trained model may get attacked again by newly generated adversarial examples [39]. To overcome this shortcoming, recent research attempts to directly modify the network architecture to get immunity towards adversarial attacks.

Gu *et al.* first explored the influence of network topology towards adversarial attacks [40]. Inspired by the contractive auto-encoder, they proposed Deep Contractive Networks with a smoothness penalty which increases the network resistibility to adversarial examples without a significant performance drop. Lyu *et al.* developed a family of gradient regularization modules that effectively penalize the gradient of the loss function for inputs [41]. Samangouei *et al.* proposed Defense-GAN to model the distribution of unperturbed images [42]. Given a perturbed image, Defense-GAN can find its close output which does not contain the adversarial changes. Hoffman *et al.* analyzed and developed a computationally efficient implementation of Jacobian regularization that increases the classification margins of neural networks [43].

3) Adversarial Example Detection: Another possible coping strategy to adversarial attacks is adversarial example detection, which directly detects whether a given sample is contaminated by adversarial perturbations. Grosse *et al.* proposed to augment the learning model with an additional outlier class C_{out}. In the test phase, the trained network is expected to classify the adversarial examples into the C_{out} class [44]. Similarly, Gong *et al.* [45] trained a binary classification model to discriminate all adversarial examples from the clean ones.

The consistency of the input sample's predictions is another possible clue for adversarial example detection. Feinman *et al.* proposed to generate multiple randomized classifiers with the dropout technic [46]. If all these classifiers make very different predictions on the input sample x, then x is very likely to be an adversarial example. A more comprehensive review of this research direction can be found in [47].

III. Self-Attention Context Network

Most of the existing deep learning models for the HSI classification task mainly focus on local feature extraction [21]. Although the receptive field of a CNN gets increased in the deep layer, it still focuses on a local region. Pixels locating in other areas are ignored in the feature extraction. This property makes existing models very vulnerable to adversarial attacks [23]. To improve the inherent resistibility of deep learning models, we propose a novel self-attention context network (SACNet) to extract global context information from HSI. Compared with local feature extraction, the extraction of global context information requires building relationships with all related pixels in the whole image for a given pixel. In this way, the produced prediction by the network at this pixel would be affected by its related pixels. Under such a circumstance, if an incorrect label was assigned to this pixel by the adversarial attack, the incorrect loss at this pixel would also be passed to all other related pixels through back-propagation. Thus, the total loss at this pixel would be shared by all other related pixels, and attacking such networks may require a higher level of perturbation.

The architecture of the proposed SACNet is presented in Fig. 2. We first adopt three dilated 2D convolutional layers and one average pooling layer as the backbone network to extract hierarchical features. Then, the learned features are fed into the self-attention module to build global spatial dependency. These attention features are further sent to the context-encoding module to extract global context information. In the classification phase, we concatenate the first two shallow convolutional features with the learned global context features to achieve multi-scale classification. In the following subsections, we will present the network architecture and the optimization of the framework in detail.

A. Backbone Network

The backbone network consists of three 3×3 2D convolutional layers and one 2×2 average pooling layer. Each convolutional layer contains m convolutional kernels (we set $m = 64$ in this study). Inspired by [20], we use dilated convolution with different dilation rates to expand the receptive field of the network, which can extract features from a larger neighbor region with the same convolutional kernel size.

Formally, let $X \in \mathbb{R}^{h \times w \times c}$ be the input hyperspectral data, where h, w, and c denote the height, width, and the number of bands in the image, respectively. Then, features in the first convolutional layer can be formulated as $\boldsymbol{C_1} = g(W_1 * X + b_1)$, where W_1 and b_1 are the weight matrix and bias vector in the first convolutional layer, respectively. $*$ denotes the 2D convolution operation. $g(x) = \max(0, x)$ is the rectified linear unit (ReLU) function. Similarly, $\boldsymbol{C_2} = g(W_2 * \boldsymbol{C_1} + b_2)$, and $\boldsymbol{C_3} = g(W_3 * \boldsymbol{P_1} + b_3)$ are features in the second and the third convolutional layers, where $\boldsymbol{P_1} = P_{avg}(\boldsymbol{C_2})$ denotes features in the first pooling layer, and $P_{avg}(\cdot)$ denotes the 2×2 average pooling operation.

B. Self-Attention Learning

Attention mechanism has been commonly used in machine translation [48] and computer vision tasks [49]. Inspired by the work in [50], [51], we propose a self-attention module to build global spatial dependency.

As shown in Fig. 2, the self-attention module receives the features of $\boldsymbol{C_3}$ generated by the backbone network as the input signal. To decrease the computation burden in attention learning, we use a pooling layer $\boldsymbol{P_2} = P_{avg}(\boldsymbol{C_3})$ to halve the spatial dimension of $\boldsymbol{C_3}$, where $\boldsymbol{P_2} \in \mathbb{R}^{\frac{h}{4} \times \frac{w}{4} \times m}$. Then, $\boldsymbol{P_2}$ is fed into three 1×1 convolutional layers with n filters to

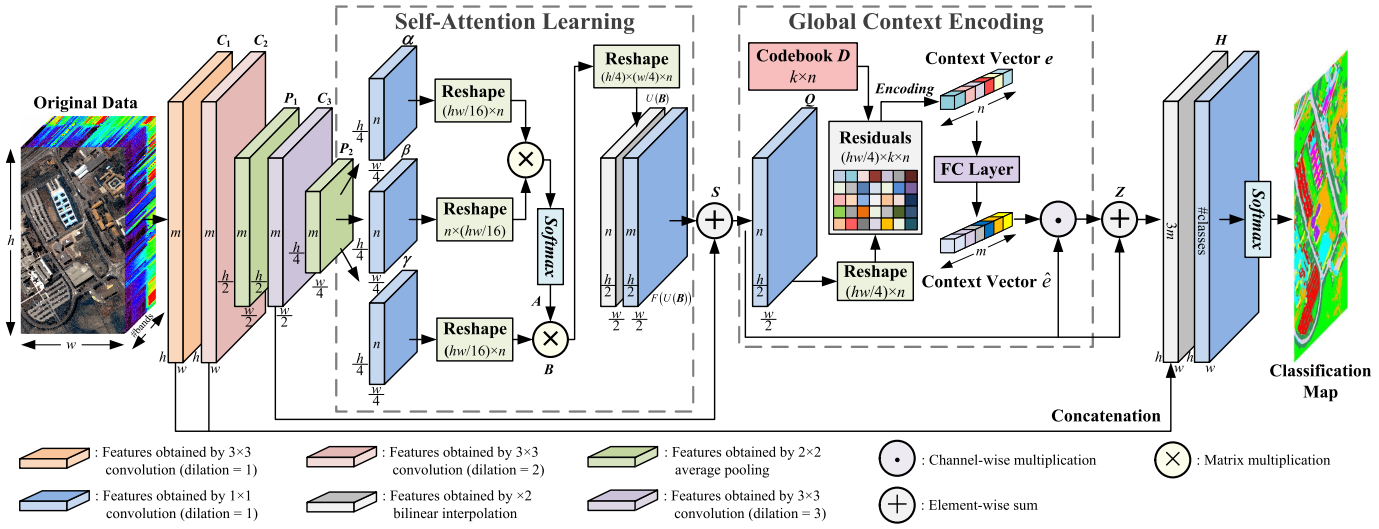

Fig. 2. An illustration of the proposed self-attention context network (SACNet). The dilated convolutional layers are first adopted to extract hierarchical features. Then, we use the self-attention module to build global spatial dependency. The obtained attention features are further strengthened by the context-encoding module. Finally, we concatenate the deep features with features in the first two convolutional layers to achieve multi-scale classification. Each feature map is shown with the size of its tensor (e.g., h, w, and m denote the height, width, and the number of channels, respectively).

generate new feature maps $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$, where $\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}\} \in \mathbb{R}^{\frac{h}{4}\times\frac{w}{4}\times n}$ (we set $n = 32$ in this study). To get the global spatial attention map $\boldsymbol{A} \in \mathbb{R}^{\frac{hw}{16}\times\frac{hw}{16}}$, we reshape $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ into $\mathbb{R}^{\frac{hw}{16}\times n}$, and apply the matrix multiplication between $\boldsymbol{\alpha}$ and the transpose of $\boldsymbol{\beta}$ with the softmax function:

$$\boldsymbol{A}_{(i,j)} = \frac{\exp\left(\boldsymbol{\alpha}_i \times \boldsymbol{\beta}_j^T\right)}{\sum_{k=1}^{\frac{hw}{16}} \exp\left(\boldsymbol{\alpha}_k \times \boldsymbol{\beta}_j^T\right)}, \tag{9}$$

where $\boldsymbol{A}_{(i,j)}$ measures the impact of position i on position j. The generated attention map $\boldsymbol{A}$ is further multiplied by $\boldsymbol{\gamma}$ and the result $\boldsymbol{B} = \boldsymbol{A} \times \boldsymbol{\gamma}$ is reshaped to $\boldsymbol{B} \in \mathbb{R}^{\frac{h}{4}\times\frac{w}{4}\times n}$. The final attention enhanced features $\boldsymbol{S} \in \mathbb{R}^{\frac{h}{2}\times\frac{w}{2}\times m}$ is formulated as:

$$\boldsymbol{S} = F\left(U\left(\boldsymbol{B}\right)\right) + \boldsymbol{C_3}, \tag{10}$$

where $U(\cdot)$ denotes the bilinear interpolation with an upsampling rate of 2, and $F(\cdot)$ represents the nonlinear transformation implemented by an 1×1 convolutional layer with m kernels. It can be inferred from (10) that the attention features $\boldsymbol{S}$ is the sum of the original feature map and the global feature map containing relationships across all positions in the image. This property enables the network to build the spatial dependency for pixels belonging to the same category.

C. Global Context Encoding

Our initial inspiration for extracting global context information comes from the work in [52], where the authors used context information to selectively highlight the class-dependent feature maps for semantic segmentation. Since the extraction of global context information involves the participation of all pixels in the image, it may also help to improve the resistibility of the network towards adversarial attacks.

As shown in Fig. 2, the global context encoding module receives the attention features $\boldsymbol{S}$ as the input signal. We first use an 1×1 convolutional layer with n kernels to decrease the dimensionality of $\boldsymbol{S}$ and the result feature map $\boldsymbol{Q}$ is reshaped into $\mathbb{R}^{\frac{hw}{4}\times n}$. Let $q_i \in \mathbb{R}^n$ be the ith element in $\boldsymbol{Q}$, where $i = 1, 2, \ldots, N$, and $N = \frac{hw}{4}$. Then, we define a codebook $\boldsymbol{D} = \{d_j\}_{j=1}^k$ which aims to learn the visual centers based on the global statistic information in $\boldsymbol{Q}$, where $d_j \in \mathbb{R}^n$ denotes the jth codeword (we set $k = 48$ in this study). The normalized residuals between $\boldsymbol{Q}$ and $\boldsymbol{D}$ can be calculated as:

$$e_{ij} = \frac{\exp\left(-s_j \|r_{ij}\|^2\right)}{\sum_{l=1}^{k} \exp\left(-s_l \|r_{il}\|^2\right)} r_{ij}, \tag{11}$$

where $r_{ij} = q_i - d_j$ denotes the residuals between the ith element in $\boldsymbol{Q}$ and the jth codeword in $\boldsymbol{D}$, and s_j is the scaling factor for the jth codebook. The global context vector $e \in \mathbb{R}^n$ is then defined as $e = \sum_{j=1}^{k} \mathcal{N}\left(\sum_{i=1}^{N} e_{ij}\right)$, where $\mathcal{N}(\cdot)$ denotes the batch normalization with ReLU activation. A fully connected (FC) layer is further utilized to transform the dimension of e into $\mathbb{R}^m$. The transformed context vector is denoted as $\hat{e} = \sigma\left(W_{fc} e + b_{fc}\right)$, where $\sigma(x) = \frac{1}{1+\exp(-x)}$ denotes the sigmoid function, W_{fc} and b_{fc} are the weight matrix and bias vector in the FC layer. Finally, the context enhanced features $\boldsymbol{Z} \in \mathbb{R}^{\frac{h}{2}\times\frac{w}{2}\times m}$ are formulated as:

$$\boldsymbol{Z} = \boldsymbol{S} \odot \left(\hat{e} + 1\right), \tag{12}$$

where $\odot$ is the channel-wise multiplication.

D. Optimization

Since the context enhanced feature map $\boldsymbol{Z}$ contains global contextual information, directly conducting classification with $\boldsymbol{Z}$ may lose the detailed object boundaries in the image. To conduct multi-scale classification, we fuse the features in the first two convolutional layers $\boldsymbol{C_1}$ and $\boldsymbol{C_2}$ with $\boldsymbol{Z}$ by concatenation: $\boldsymbol{H} = [\boldsymbol{C_1}; \boldsymbol{C_2}; U(\boldsymbol{Z})] \in \mathbb{R}^{h\times w\times 3m}$, where $\boldsymbol{H}$ is the fused feature map, and $U(\cdot)$ denotes the bilinear interpolation with an upsampling rate of 2. The final classification is achieved by an 1×1 convolutional layer with softmax function.

Algorithm 1 Adversarial Attack on SACNet

Input:

(1) A HSI X and the corresponding training label Y.
(2) A SACNet f with parameters θ.
(3) A small scalar value ϵ that controls the norm of the adversarial perturbation, the number of training epochs τ, and the learning rate η.

Initialize θ with random Gaussian values.
for t in $range\,(0, \tau)$ **do**
  Compute the attention enhanced features $\boldsymbol{S}$ via (10).
  Compute the context enhanced features $\boldsymbol{Z}$ via (12).
  Compute the cross-entropy loss $\mathcal{L}_{cls}$ via (13).
  Update θ by descending its stochastic gradients via $\theta \leftarrow \theta - \eta \nabla_\theta \left(\mathcal{L}_{cls}\right)$.
end for
Generate the adversarial image X_{adv} via (5).
Feed the adversarial image X_{adv} to the trained SACNet f to accomplish the classification.

Output: The predictions on the adversarial image X_{adv}.

Let $\hat{Y}$ and Y be the predicted probability map and the label map, respectively. The cross-entropy loss $\mathcal{L}_{cls}$ can thereby be defined as:

$$\mathcal{L}_{cls} = -\frac{1}{hw} \sum_{i=1}^{h} \sum_{j=1}^{w} \sum_{k=1}^{v} Y_{(i,j,k)} \log \left(\hat{Y}_{(i,j,k)} \right), \qquad (13)$$

where v denotes the number of categories.

The whole framework is optimized with the mini-batch stochastic gradient descent algorithm. The detailed steps of conducting adversarial attacks on the proposed SACNet are given in Algorithm 1. Note that the goal of adversarial attacks for the HSI classification task is to perturb the HSI to maximize the number of incorrect class decisions on all test pixels in the image. However, adversarial attack algorithms like the FGSM are initially proposed for changing a single class decision produced by the network for a given image. Thus, to generate adversarial examples for a HSI, the calculation of $J\left(\theta, x, \hat{y}\right)$ in (5) should involve the sum of cross-entropy losses over all pixels in the input image.

IV. Experiments

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

A. Data Descriptions

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

The first dataset is Pavia University,[1] which consists of 103 spectral bands with 610 by 340 pixels. The second dataset is Houston [53], which consists of 144 spectral bands with 349 by 1905 pixels. The third dataset is Salinas[1], which consists of 204 spectral bands with 512 by 217 pixels. The false-color composition images of these three datasets and the corresponding ground-truth maps are shown in Fig. 3 to Fig. 5. The training and test sets are listed in Table I to Table III.

[1] http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes

TABLE I
Numbers of Training and Test Samples Used in the Pavia University Dataset

Class number	Class name	Training	Test
1	Asphalt	300	6331
2	Meadows	300	18349
3	Gravel	300	1799
4	Trees	300	2764
5	Metal sheets	300	1045
6	Bare soil	300	4729
7	Bitumen	300	1030
8	Bricks	300	3382
9	Shadows	300	647
	Total	2700	40076

TABLE II
Numbers of Training and Test Samples Used in the Houston Dataset

Class number	Class name	Training	Test
1	Grass healthy	300	951
2	Grass stressed	300	954
3	Grass synthetic	300	397
4	Trees	300	944
5	Soil	300	942
6	Water	300	25
7	Residential	300	968
8	Commercial	300	944
9	Road	300	952
10	Highway	300	927
11	Railway	300	935
12	Parking lot1	300	933
13	Parking lot2	300	169
14	Tennis court	300	128
15	Running track	300	360
	Total	4500	10529

TABLE III
Numbers of Training and Test Samples Used in the Salinas Dataset

Class number	Class name	Training	Test
1	Brocoli green weeds 1	300	1709
2	Brocoli green weeds 2	300	3426
3	Fallow	300	1676
4	Fallow rough plow	300	1094
5	Fallow smooth	300	2378
6	Stubble	300	3659
7	Celery	300	3279
8	Grapes untrained	300	10971
9	Soil vinyard develop	300	5903
10	Corn senesced green weeds	300	2978
11	Lettuce romaine 4wk	300	768
12	Lettuce romaine 5wk	300	1627
13	Lettuce romaine 6wk	300	616
14	Lettuce romaine 7wk	300	770
15	Vinyard untrained	300	6968
16	Vinyard vertical trellis	300	1507
	Total	4800	49329

B. Experimental Design and Implementation Details

The experiments are conducted in three parts. The first part reports the classification results of different deep learning methods on adversarial test sets. We adopt the FGSM with ℓ_∞

TABLE IV

QUANTITATIVE CLASSIFICATION RESULTS OF THE PAVIA UNIVERSITY DATASET ON THE ADVERSARIAL TEST SET WITH $\epsilon = 0.04$. THE TARGETED CATEGORY IN THE ATTACK IS SET AS CLASS 1. BEST RESULTS ARE SHOWN IN **BOLD**

Class	1D-CNN	SpeFCN	SpaFCN	3D-CNN	3D-DL	SSFCN	PResNet	DilatedFCN	SACNet
1	91.47±2.80	92.93±5.37	97.89±0.69	89.74±30.69	**99.89±0.13**	98.54±0.73	97.72±1.91	91.51±1.13	91.96±3.18
2	41.33±14.15	18.50±8.72	12.05±17.28	17.13±21.22	20.07±13.34	13.58±11.91	27.09±18.26	86.83±5.19	**91.94±6.31**
3	6.63±7.92	0.41±1.16	5.26±7.89	12.65±12.44	4.27±8.61	2.00±1.73	25.79±15.56	85.35±6.23	**93.55±6.04**
4	89.33±5.88	92.35±9.66	98.65±0.53	75.36±27.42	85.10±11.42	98.91±0.55	87.21±5.84	88.15±1.69	**95.66±1.54**
5	99.65±0.18	99.55±0.31	99.90±0.14	89.91±30.75	98.48±2.50	99.89±0.19	99.78±0.27	96.22±2.26	**99.54±0.39**
6	53.45±11.98	53.73±34.02	60.53±21.53	31.26±21.39	54.91±23.26	65.12±15.53	48.88±16.04	45.08±29.36	**93.95±4.89**
7	1.23±1.92	6.15±12.31	37.46±14.17	5.00±22.36	0.39±0.52	32.22±14.65	4.50±6.49	33.05±16.97	**81.11±23.32**
8	36.71±12.37	38.50±11.61	88.91±7.69	19.58±22.28	1.29±2.51	90.45±3.63	23.07±13.52	88.86±3.13	**95.66±2.41**
9	99.90±0.29	**99.95±0.14**	99.64±0.33	89.53±30.62	99.90±0.14	99.61±0.28	98.55±1.05	95.83±1.26	98.88±0.76
OA (%)	53.48±7.10	43.49±4.71	47.84±7.72	37.04±10.41	41.81±7.12	49.06±5.89	47.04±9.19	81.84±5.08	**92.86±3.07**
κ (%)	42.37±7.62	33.02±5.50	40.87±7.61	24.03±10.90	30.30±7.56	41.84±5.53	36.34±9.34	75.69±6.65	**90.48±3.90**
AA (%)	57.74±3.37	55.78±4.41	66.70±3.04	47.79±13.65	51.59±3.62	66.70±3.00	56.95±3.42	78.99±4.72	**93.58±3.30**
Runtime (s)	34.30±1.42	83.06±2.73	35.44±2.26	75.70±2.98	99.31±4.30	56.59±2.38	67.20±4.07	**25.27±2.18**	71.17±2.01

norm to conduct targeted adversarial attacks using (5), where the ϵ is fixed to 0.04 and the targeted label $\hat{y}$ is set to class 1 in each dataset. The second part provides an ablation study about the proposed SACNet to evaluate how each module in SACNet influences the adversarial defense performance and explores the influence of different values of ϵ on the classification results. Finally, the third part further analyzes the generated adversarial hyperspectral examples.

All the experiments in this study are randomly repeated 20 times with random training and test data. In each repetition, we first randomly select 300 samples per category from the reference data to form the training set. Then, the remaining reference samples make up the test set. The overall accuracy (OA), kappa coefficient (κ), average accuracy (AA), and the producer accuracy for each class are utilized to quantitatively estimate different methods. Both the average value and the standard deviation are reported. We implement the experiments with the PyTorch platform in this study. The Adam optimizer is used to train the networks with a learning rate of $5e-4$ and a weight decay of $5e-5$. The training epochs are set as 1000. The experiments are carried out with two Intel Xeon Silver 4114 2.20-GHz CPUs with 128 GB of RAM, and one NVIDIA GeForce RTX 2080 Ti GPU.

C. Classification Results

In this subsection, we report the classification results of the proposed method along with other methods on the adversarial test set. We adopt the FGSM with ℓ_∞ norm to generate adversarial examples using (5), where the ϵ is fixed to 0.04. A brief introduction to each method included in the experiments is summarized below.

1) 1D-CNN: Spectral classification with a five-layer 1D-CNN [54].
2) SpeFCN: Spectral classification using a fully convolutional network with 1×1 convolution [20].
3) SpaFCN: Spatial classification using a multi-scale fully convolutional network [20].
4) 3D-CNN: Spectral-spatial classification with 3D-CNN. This method receives the hyperspectral cube as the input to extract spectral-spatial features [17].
5) 3D-DL: Spectral-Spatial classification with a novel 3D deep learning framework [18].
6) SSFCN: Spectral-spatial classification using a fully convolutional network which can adaptively learn spectral and spatial features with a two-branch architecture [20].
7) PResNet: Spectral-Spatial classification with a deep pyramidal residual network [19]. The proposed pyramid architecture helps to gradually increase the feature map dimension at different convolutional layers.
8) DilatedFCN: The proposed backbone network.
9) SACNet: The proposed self-attention context network.

The quantitative results are reported in Table IV to Table VI. Although fusing both spectral and spatial features has become a classical paradigm for the HSI classification task [21], we find the superiority of spectral-spatial classification methods is not always prominent on the adversarial test set. It can be observed from Table IV to Table VI that 1D-CNN outperforms 3D-CNN for all three datasets with more than 5% in the OA metric. This phenomenon indicates that simply combining spectral and spatial information in the classification may not be enough to address the threat of adversarial attacks in HSI classification. Besides, we find most of the existing state-of-the-art deep learning-based approaches are very sensitive to adversarial attacks. Take the results of PResNet for example. It only achieves an OA of around 47% on the Pavia University dataset. For the Houston and Salinas datasets, PResNet can only yield an OA of about 36%. These results are significantly worse than the ones on the original clean test set, as reported in [19]. We also find the proposed DilatedFCN can achieve competitive performance on the adversarial test set despite its simple architecture. In the Pavia University dataset, DilatedFCN even gets an OA of around 81%, which is much higher than the ones of SSFCN and PResNet. The reason for this phenomenon may lie in the adopted dilated convolution, which provides a larger receptive field compared to traditional convolutional operation. With the help of the proposed self-attention learning and context encoding, SACNet can significantly improve the resistibility of the framework towards adversarial examples. For all three datasets, SACNet can achieve an OA of more than 90%, which dramatically outperforms the existing state-of-the-art deep learning-based methods.

To visually evaluate the influence of adversarial attacks on the classification results of the aforementioned methods,

TABLE V

QUANTITATIVE CLASSIFICATION RESULTS OF THE HOUSTON DATASET ON THE ADVERSARIAL TEST SET WITH $\epsilon = 0.04$. THE TARGETED CATEGORY IN THE ATTACK IS SET AS CLASS 1. BEST RESULTS ARE SHOWN IN **BOLD**

Class	1D-CNN	SpeFCN	SpaFCN	3D-CNN	3D-DL	SSFCN	PResNet	DilatedFCN	SACNet
1	99.02±0.57	99.61±0.66	98.54±2.86	99.48±2.25	**100±0.00**	99.29±0.89	99.41±0.95	89.60±3.99	93.57±4.34
2	66.48±9.58	55.23±17.32	71.32±7.64	45.55±23.08	54.78±16.65	72.80±5.55	43.40±14.79	58.09±10.91	**91.17±4.82**
3	21.68±30.17	0.00±0.00	13.07±24.89	15.74±21.15	36.32±27.84	3.59±7.93	7.85±10.78	52.93±19.38	**99.45±0.77**
4	73.17±13.83	61.02±28.76	91.51±2.30	49.69±32.75	71.11±18.52	91.03±3.95	54.61±22.18	71.72±7.69	**93.42±3.22**
5	72.17±12.41	34.26±34.16	32.96±22.43	56.17±27.47	75.27±11.58	51.69±21.58	59.41±12.46	41.20±14.68	**97.29±1.65**
6	89.00±7.66	61.80±23.23	68.00±19.68	78.40±21.10	90.20±9.58	74.20±13.26	69.80±18.47	87.60±9.35	**97.00±4.08**
7	26.95±12.25	30.48±28.07	86.61±5.61	15.36±10.30	45.37±9.89	87.55±5.69	19.12±7.06	74.37±7.41	**92.46±2.73**
8	49.57±7.10	32.30±10.03	35.25±6.47	38.87±8.85	57.55±6.85	42.01±8.30	34.79±8.63	37.10±6.56	**80.58±6.31**
9	17.44±5.63	7.48±12.31	45.37±13.80	51.64±27.14	65.92±14.15	55.18±16.82	13.66±5.78	75.34±4.63	**90.72±4.24**
10	12.32±8.31	0.01±0.05	2.32±4.30	4.95±8.28	6.20±7.98	3.82±5.97	2.84±2.47	25.67±14.03	**93.12±7.26**
11	1.81±1.49	0.01±0.05	30.53±18.31	4.25±4.59	3.35±3.96	42.05±18.98	2.79±2.18	25.63±9.23	**91.94±4.25**
12	16.82±7.06	10.01±7.80	12.62±7.99	9.49±16.34	15.86±10.74	15.56±8.10	5.49±4.28	34.58±12.58	**87.81±9.41**
13	32.51±6.30	18.20±18.67	31.92±17.04	25.65±26.96	77.37±6.65	41.12±22.54	56.33±9.60	51.57±13.52	**92.69±4.39**
14	62.97±12.07	61.99±31.35	55.63±26.43	75.47±19.99	80.23±16.09	58.67±22.56	45.27±18.84	43.83±30.03	**99.84±0.41**
15	94.56±8.84	67.86±31.76	82.32±12.94	89.89±9.50	97.07±4.33	84.17±7.80	84.15±13.03	72.10±13.00	**98.83±1.54**
OA (%)	44.77±3.91	33.31±5.47	50.43±6.01	39.00±10.90	51.77±6.03	55.17±4.91	35.01±3.50	54.08±4.89	**91.92±2.80**
κ (%)	40.60±4.10	29.17±5.52	46.14±6.50	33.58±11.83	47.42±6.60	51.18±5.30	30.15±3.67	49.93±5.36	**91.20±3.04**
AA (%)	49.10±3.98	36.02±5.88	50.53±7.55	44.04±10.71	58.44±6.29	54.85±5.84	39.93±4.23	56.09±6.16	**93.33±2.19**
Runtime (s)	**67.66±2.38**	157.55±1.81	97.49±1.94	223.60±9.61	220.37±9.57	135.10±2.63	206.42±8.63	89.28±2.00	180.55±1.63

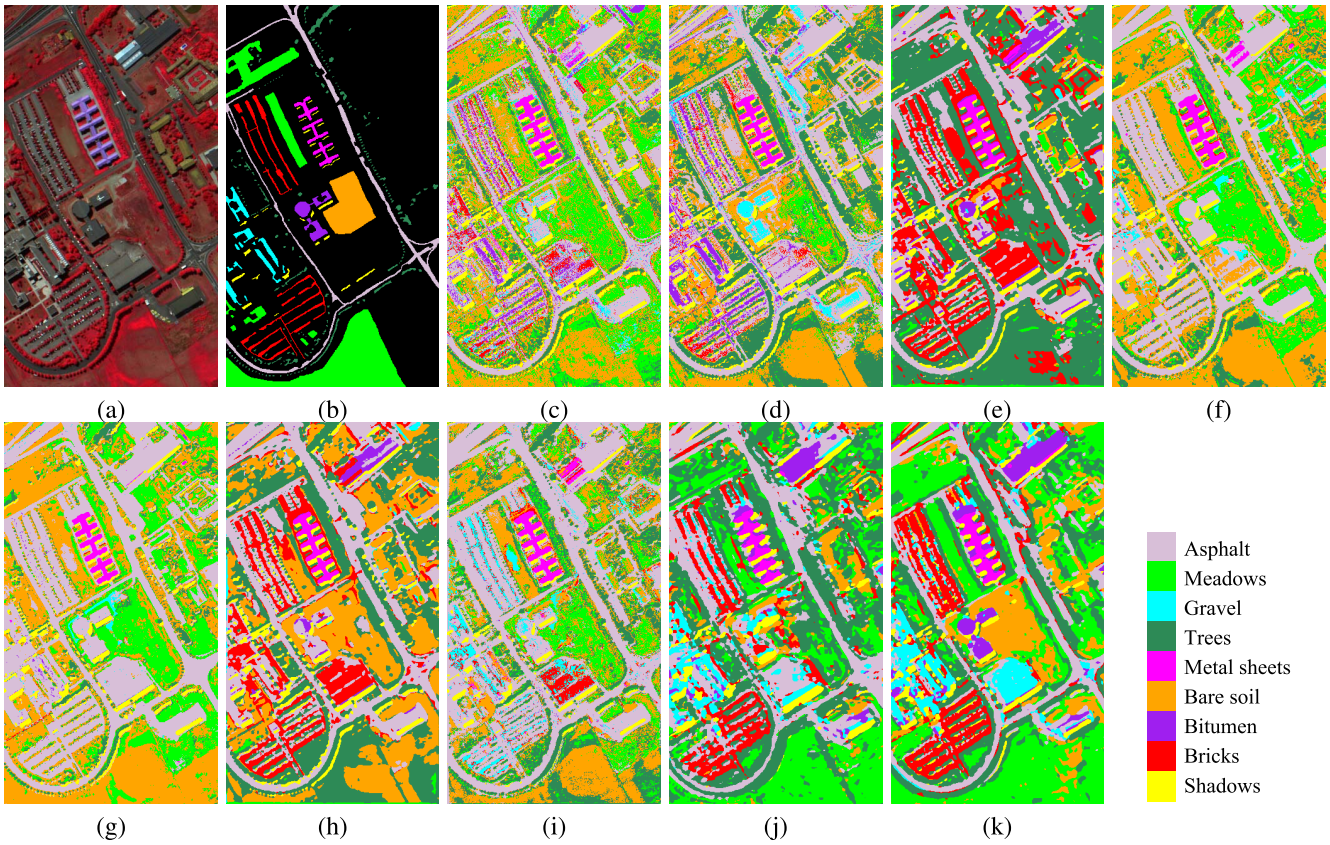

Fig. 3. Classification maps for the Pavia University dataset on the adversarial test set with $\epsilon = 0.04$. The targeted category in the attack is set as class 1. (a) The false color image. (b) Ground-truth map. (c) 1D-CNN. (d) SpeFCN. (e) SpaFCN. (f) 3D-CNN. (g) 3D-DL. (h) SSFCN. (i) PResNet. (j) DilatedFCN. (k) SACNet.

we further present the classification maps in Fig. 3 to Fig. 5. We find that the generated adversarial test set can seriously mislead the state-of-the-art methods to make wrong predictions. Take the classification maps of the Pavia University dataset for example. Although the region in the bottom part of the image belongs to the "meadows" category, most of the existing methods misclassify this region into either "trees" or "bare soil". By contrast, the proposed SACNet is more robust towards adversarial attacks, and the classification map of SACNet is much closer to the ground truth annotation. Similar phenomenons can be observed in the Houston and Salinas datasets.

The time cost of each method is also reported in Table IV to Table VI. It can be observed that the proposed DilatedFCN is very efficient owing to its simple architecture. It costs the least time in both the Pavia University and

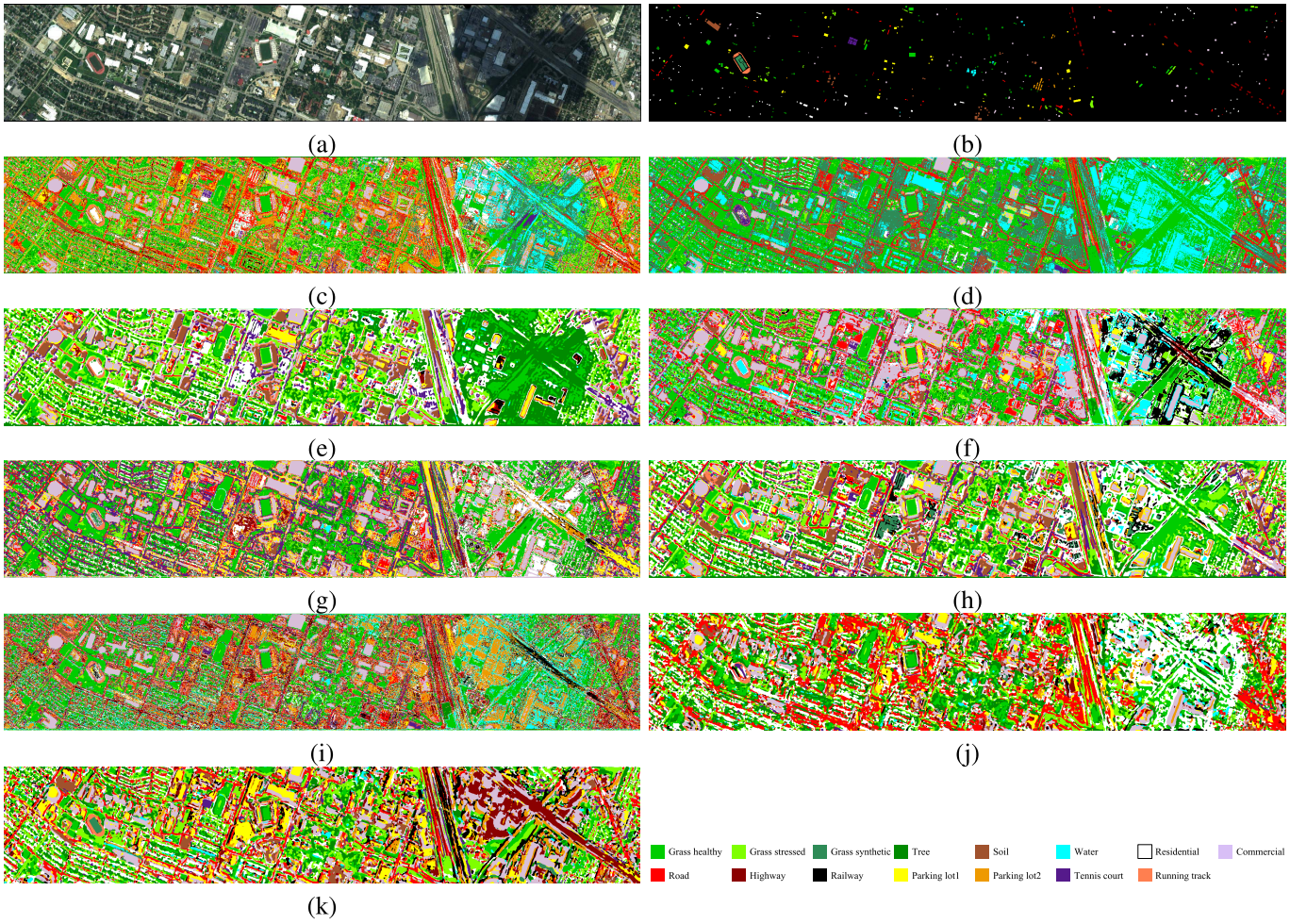

Fig. 4. Classification maps for the Houston dataset on the adversarial test set with $\epsilon = 0.04$. The targeted category in the attack is set as class 1. (a) The false color image. (b) Ground-truth map. (c) 1D-CNN. (d) SpeFCN. (e) SpaFCN. (f) 3D-CNN. (g) 3D-DL. (h) SSFCN. (i) PResNet. (j) DilatedFCN. (k) SACNet.

TABLE VI

QUANTITATIVE CLASSIFICATION RESULTS OF THE SALINAS DATASET ON THE ADVERSARIAL TEST SET WITH $\epsilon = 0.04$. THE TARGETED CATEGORY IN THE ATTACK IS SET AS CLASS 1. BEST RESULTS ARE SHOWN IN **BOLD**

Class	1D-CNN	SpeFCN	SpaFCN	3D-CNN	3D-DL	SSFCN	PResNet	DilatedFCN	SACNet
1	99.92±0.10	99.80±0.08	99.85±0.14	94.96±22.35	**100±0.00**	99.90±0.13	**100±0.00**	99.89±0.22	99.39±1.53
2	78.33±8.67	53.41±15.41	75.34±19.60	64.29±25.99	23.72±25.73	70.93±15.27	20.59±27.11	60.72±20.66	**98.75±1.13**
3	15.22±13.08	52.63±33.72	7.10±12.71	57.96±34.88	18.46±28.29	10.91±19.28	12.32±16.83	39.15±32.14	**97.44±2.98**
4	67.48±28.68	33.54±32.40	39.77±42.71	91.70±23.89	**95.72±5.27**	55.59±41.72	91.56±8.29	90.19±17.77	95.11±8.05
5	59.75±31.08	0.94±3.72	28.56±30.35	61.87±30.81	42.29±33.39	34.35±37.49	31.45±19.61	77.89±21.07	**94.29±10.68**
6	**99.77±0.12**	99.35±1.18	95.46±14.65	89.99±30.78	93.96±19.35	99.46±0.86	99.23±0.80	93.40±13.21	99.73±0.31
7	81.95±18.77	31.90±37.22	40.87±32.12	58.09±40.14	44.27±38.45	26.57±32.33	46.00±33.64	87.66±11.09	**96.70±2.54**
8	79.55±10.08	59.72±25.20	81.65±30.99	62.28±30.88	17.86±4.29	71.66±36.71	26.67±20.74	**85.63±11.04**	80.72±6.46
9	62.79±29.82	27.10±30.68	7.27±12.68	61.08±33.11	22.14±23.45	5.76±12.76	6.10±11.77	0.24±0.47	**96.05±4.82**
10	73.88±9.01	80.03±14.22	87.14±10.63	65.11±29.27	14.62±22.20	81.75±10.79	34.77±16.25	92.05±8.81	**98.16±1.09**
11	47.68±23.77	78.65±23.72	81.28±22.91	49.07±33.15	33.30±35.76	86.41±17.43	19.60±25.01	86.63±14.27	**99.83±0.36**
12	53.88±11.67	26.62±25.18	30.66±34.40	34.21±25.37	22.74±27.66	26.43±33.96	17.42±23.02	60.77±25.36	**99.29±2.28**
13	41.14±18.09	0.27±0.72	97.76±6.93	72.36±37.08	21.72±25.27	98.25±2.32	58.60±31.13	97.92±4.49	**99.89±0.21**
14	97.42±1.71	96.79±2.58	91.57±12.74	92.88±21.97	92.55±10.30	85.05±24.98	83.56±10.15	95.58±3.98	**99.44±0.53**
15	43.70±14.30	29.48±28.53	22.20±30.88	42.91±27.29	18.00±6.24	30.65±36.01	20.77±15.38	39.44±18.07	**81.07±7.51**
16	96.98±1.31	82.42±13.76	44.84±29.83	77.13±28.39	45.32±30.15	51.09±28.53	75.82±17.49	96.17±5.01	**99.02±1.89**
OA (%)	70.00±5.37	50.89±10.44	54.71±5.40	63.03±20.98	34.23±9.89	53.01±5.36	36.20±5.76	66.73±4.87	**91.57±2.13**
κ (%)	66.40±5.92	45.99±10.64	49.41±5.63	59.32±21.51	27.30±10.67	47.78±5.57	29.71±5.98	62.88±5.29	**90.56±2.38**
AA (%)	68.72±6.07	53.29±7.77	58.21±5.74	67.24±20.78	44.17±11.78	58.42±6.44	46.53±5.20	75.21±3.81	**95.93±1.43**
Runtime (s)	55.45±2.07	65.88±2.48	23.98±2.30	152.62±4.26	192.40±8.69	39.61±2.32	93.45±4.19	**19.33±2.47**	41.43±2.42

Salinas datasets. By contrast, the time cost of the proposed SACNet is relatively larger since the self-attention learning and the context encoding increase the computation burden of the whole framework. However, compared with other existing deep learning-based methods, the time cost of SACNet is still very competitive. Take the Pavia University data set for example. While 3D-DL costs about 99 seconds, the time cost of SACNet is only about 71 seconds.

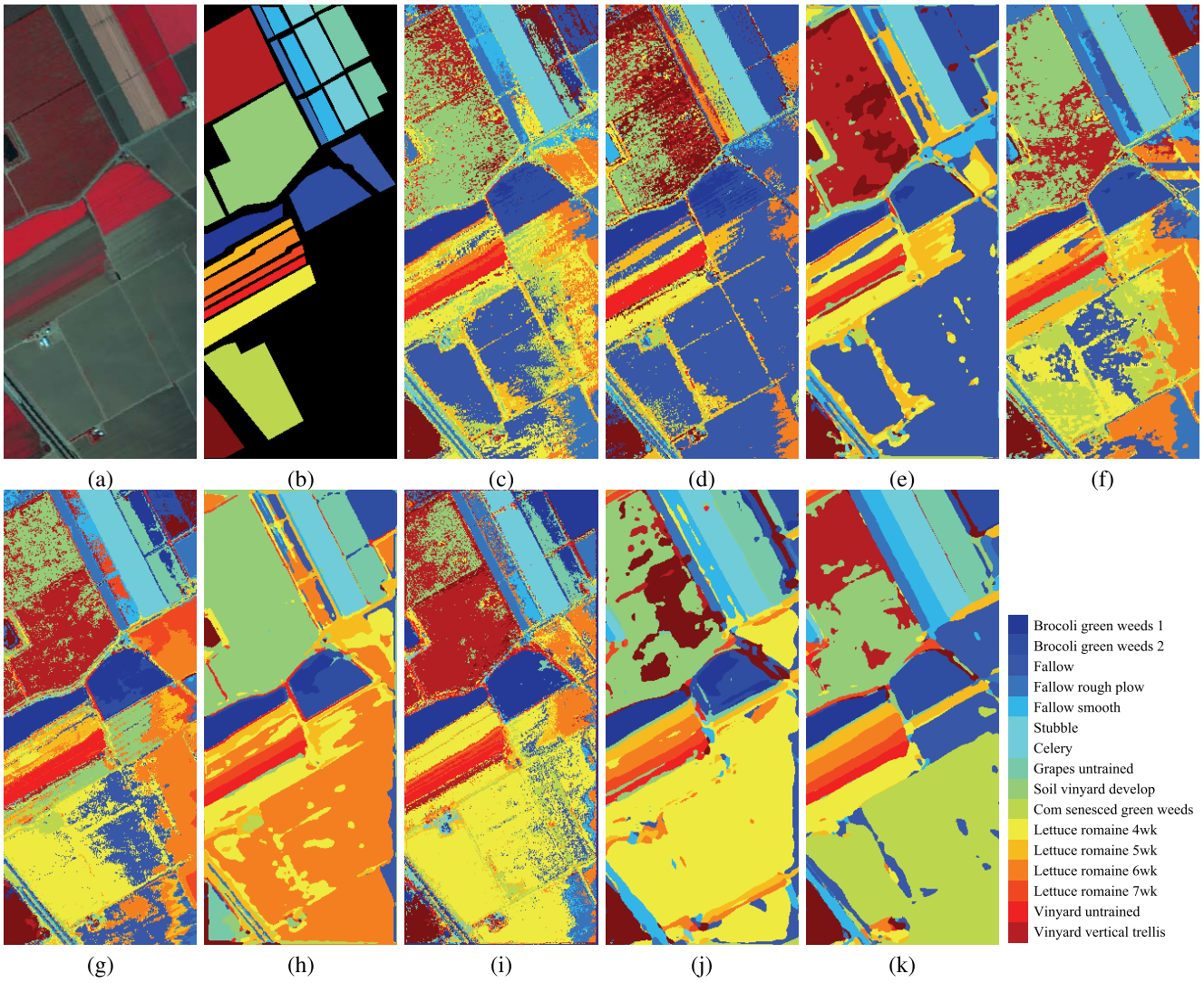

Fig. 5. Classification maps for the Salinas dataset on the adversarial test set with $\epsilon = 0.04$. The targeted category in the attack is set as class 1. (a) The false color image. (b) Ground-truth map. (c) 1D-CNN. (d) SpeFCN. (e) SpaFCN. (f) 3D-CNN. (g) 3D-DL. (h) SSFCN. (i) PResNet. (j) DilatedFCN. (k) SACNet.

TABLE VII

PERFORMANCE CONTRIBUTION OF EACH MODULE IN SACNET (REPORTED IN OA). BEST RESULTS ARE HIGHLIGHTED IN **BOLD**

Method	DilatedFCN	+SA	+CE	+SA+CE
DilatedFCN	✓	✓	✓	✓
Self-Attention (SA)		✓		✓
Context Encoding (CE)			✓	✓
Pavia University	81.84	89.25	91.36	**92.86**
Houston	54.08	84.74	90.87	**91.92**
Salinas	66.73	85.99	88.58	**91.57**

D. Ablation Study

In this subsection, we evaluate how each module in the proposed SACNet influences the classification performance on the adversarial test set. The FGSM with ℓ_∞ norm is adopted to generate adversarial examples using (5), where the ϵ is fixed to 0.04 in the experiments. The detailed classification results with different modules are presented in Table VII. Here, SA denotes self-attention learning, and CE represents the context encoding mechanism. It can be inferred from Table VII that both self-attention learning and context encoding can significantly improve the resistibility of DilatedFCN towards adversarial attacks. Besides, compared to self-attention learning, context encoding is more beneficial to the adversarial defense. Take the results in the Houston dataset for example. While self-attention learning enables the network to yield an OA of 84.74%, the adoption of context encoding can increase the OA to 90.87%. For all three datasets, combining both self-attention learning and context encoding mechanism can achieve the best OAs.

Another interesting issue is the influence of different levels of adversarial perturbations on classification performance. To this end, we generate adversarial examples with different adversarial intensity values using (5), where the ϵ takes the value from $\{0.01, 0.02, 0.04, 0.06, 0.08, 0.1, 0.2, 0.4, 0.6, 0.8, 1\}$ in the experiments. The classification results are reported in Fig. 6. It can be observed that as ϵ increases, the OA values of all methods tend to decrease, which indicates that adversarial examples generated with a larger perturbation

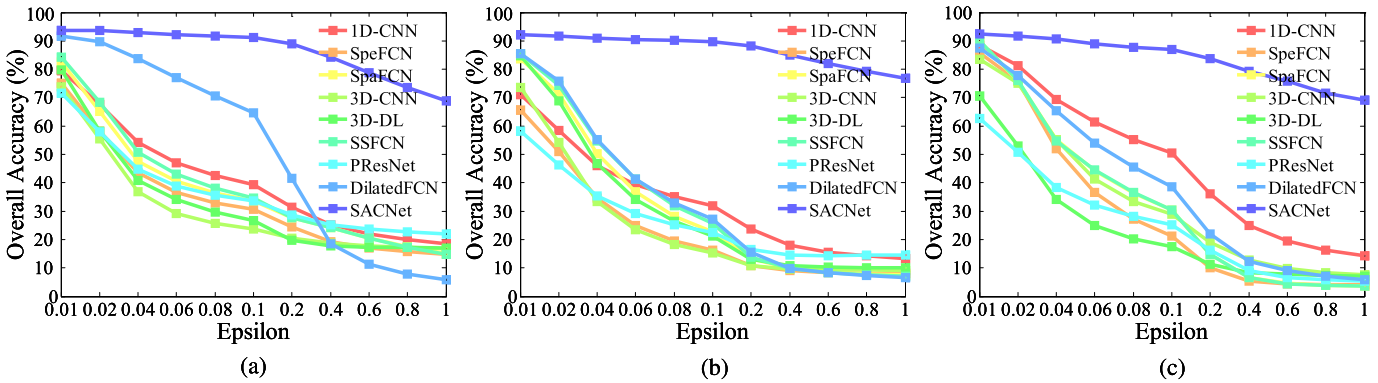

Fig. 6. The overall accuracies obtained by different methods on the adversarial test set with different values of ϵ. (a) Pavia University dataset. (b) Houston dataset. (c) Salinas dataset.

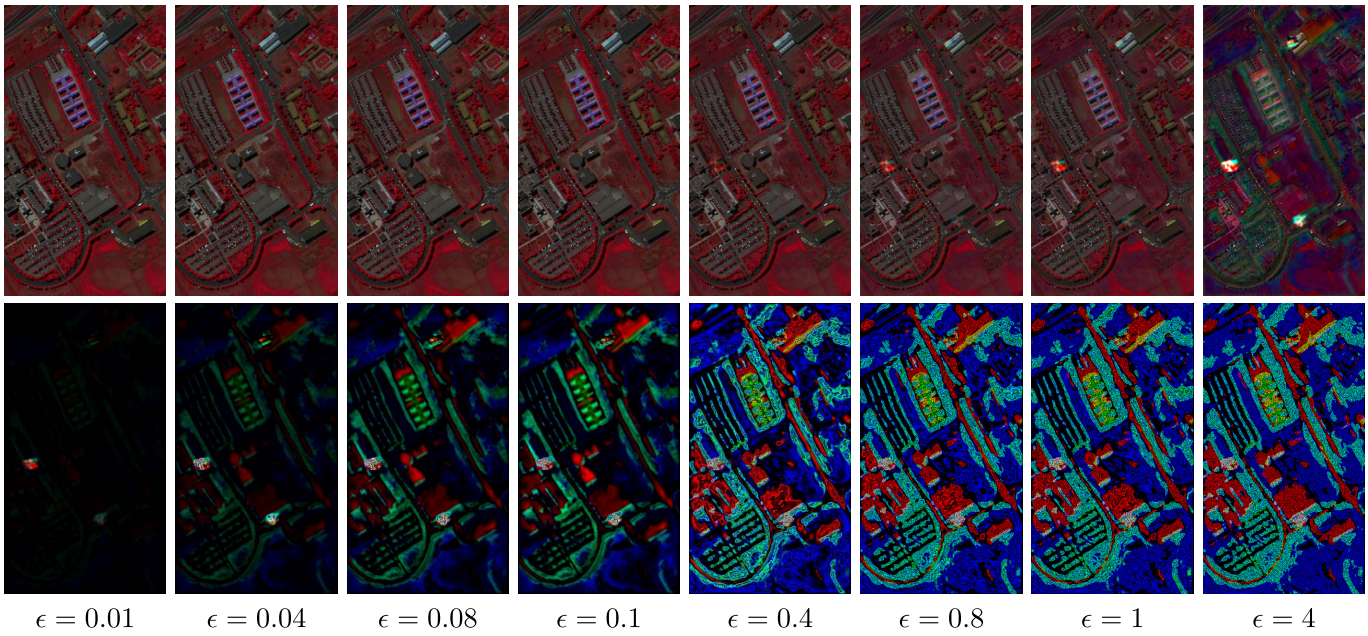

Fig. 7. The generated adversarial examples (top row) and the corresponding adversarial perturbations (bottom row) of the Pavia University dataset with different values of ϵ. The false-color image is adopted for illustration (R:102, G:56, B:31).

intensity generally are more detrimental to deep neural networks. Besides, compared to existing state-of-the-art methods like SSFCN and PResNet, the proposed SACNet shows the strongest resistibility towards adversarial examples for all three datasets under different levels of adversarial attacks. When ϵ reaches 1, most of the existing methods can only obtain an OA of less than 30%, while the proposed SACNet can still achieve an OA of more than 70%, which demonstrates the effectiveness of our method.

We further conduct experiments with more adversarial attack methods including PGD and C&W. For all datasets, we run 10 iterations of PGD and C&W attacks with the ℓ_∞ norm using a step size of 0.04 to generate the adversarial examples. The experimental results are shown in Table VIII. It can be observed from Table VIII that PGD and C&W can achieve more powerful attacks compared to FGSM. Take the Houston dataset for example. While the OA of SSFCN on the FGSM adversarial test set is about 55.17%, it dramatically decreases to 10.20% and 9.85% on the PGD and C&W adversarial test sets, respectively. Similar phenomenons can be observed in other comparing methods. By contrast, the proposed SACNet can still obtain an OA of more than 70% in all three datasets when confronted with PGD and C&W attacks, which dramatically outperforms the existing state-of-the-art deep learning-based methods.

Table VIII also reports the classification performance of different methods on the original clean test sets. We can find that the proposed SACNet can yield an OA of more than 96% in all three datasets. It achieves the highest OAs on both the Houston and Salinas datasets and the second-highest OA on the Pavia University dataset (slightly lower than 3D-DL). These results demonstrate that the proposed method can also obtain competitive results on the clean images compared to the existing state-of-the-art methods.

E. Analysis About the Adversarial Examples

In this subsection, we first visualize the generated hyperspectral adversarial examples and perturbations with different values of ϵ. The Pavia University dataset is adopted as an example. As presented in Fig. 7, the adversarial perturbation is imperceptible when ϵ is set with small values (e.g., $\epsilon = 0.01$).

TABLE VIII

QUANTITATIVE CLASSIFICATION RESULTS ON THE ORIGINAL CLEAN TEST SETS AND THE ADVERSARIAL TEST SETS GENERATED WITH DIFFERENT METHODS (REPORTED IN OA). BEST RESULTS ARE SHOWN IN **BOLD**

Method	1D-CNN	SpeFCN	SpaFCN	3D-CNN	3D-DL	SSFCN	PResNet	DilatedFCN	SACNet
Pavia University									
Clean Test Set	88.86	89.52	94.58	91.76	**96.79**	95.61	91.61	92.49	96.01
FGSM	53.48	43.49	47.84	37.04	41.81	49.06	47.04	81.84	**92.86**
PGD	38.73	23.65	24.39	23.34	17.03	24.49	19.15	25.11	**78.29**
C&W	37.37	24.37	24.30	21.66	18.46	24.22	20.80	25.25	**72.55**
Houston									
Clean Test Set	94.55	89.14	96.01	96.03	96.29	96.62	95.09	90.78	**96.95**
FGSM	44.77	33.31	50.43	39.00	51.77	55.17	35.01	54.08	**91.92**
PGD	29.04	10.05	9.97	12.35	17.44	10.20	9.69	12.24	**82.86**
C&W	31.97	9.67	10.43	11.90	17.07	9.85	10.70	11.39	**83.62**
Salinas									
Clean Test Set	92.31	90.55	94.37	90.84	95.29	95.27	95.04	92.50	**97.09**
FGSM	70.00	50.89	54.71	63.03	34.23	53.01	36.20	66.73	**91.57**
PGD	38.13	12.88	7.12	19.96	6.87	6.13	5.94	14.49	**71.90**
C&W	39.44	11.25	8.97	28.06	7.61	11.55	6.58	14.01	**77.23**

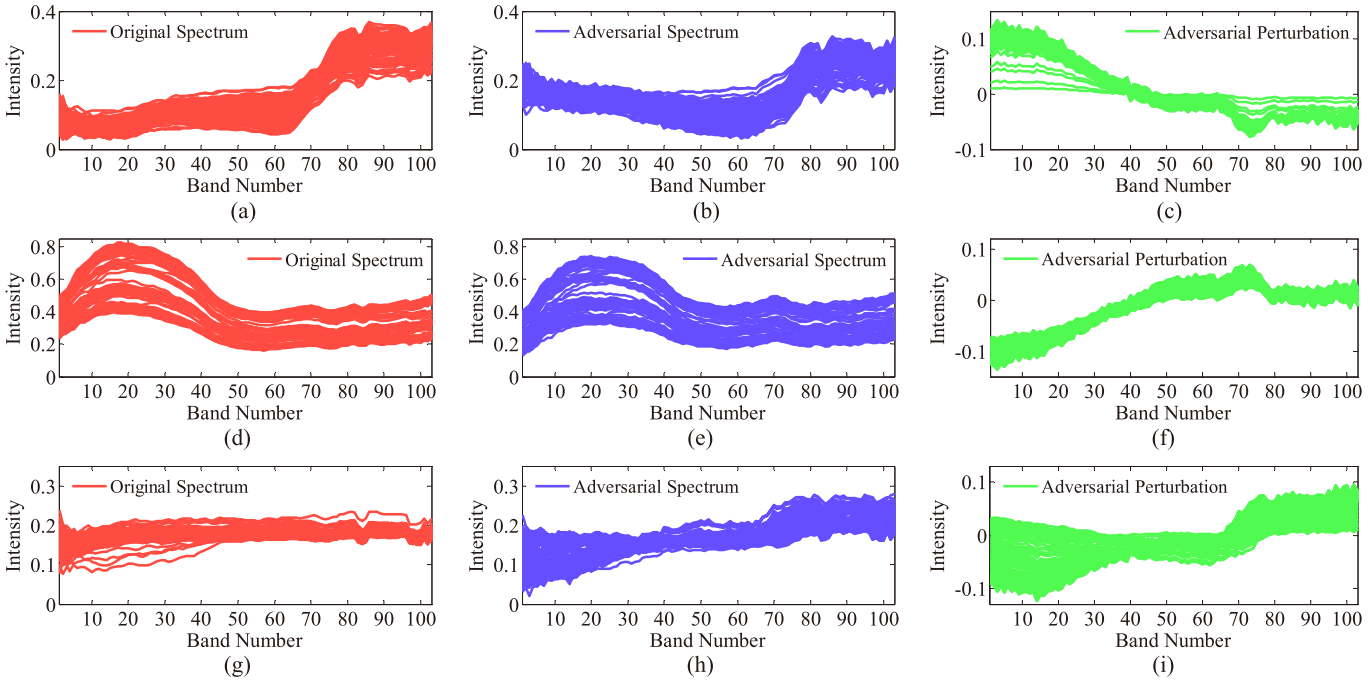

Fig. 8. The spectral curves of the original data, adversarial examples, and adversarial perturbations in different categories of the Pavia University dataset (with $\epsilon = 1$). (a)–(c): The 2nd class (meadows). (d)–(f): The 5th class (metal sheets). (g)–(i): The 7th class (bitumen).

As ϵ gradually increases, the intensity of the adversarial perturbation also gets larger. Besides, it can be observed that the difference between the generated adversarial examples and the original hyperspectral image is very hard to be perceived for the human visual system when $\epsilon \leq 0.1$. This phenomenon also indicates that adversarial attacks may bring about a serious threat for the HSI classification task since such imperceptible difference could mislead most of the existing state-of-the-art deep learning-based methods to make wrong predictions according to the results in Fig. 6 (a).

To better analyze how adversarial attacks influence the spectral reflectance characteristic and cheat deep neural networks, we plot the spectral curves of the original data, adversarial examples, and adversarial perturbations from different categories in Fig. 8. Compared to the false-color visualization in Fig. 7, spectral curves make the difference between the original data and adversarial examples more perceptible for human observers. Take the meadows class in the first row of Fig. 8 for example. As typical vegetation, the reflectance of meadows in the blue region of the spectrum (band 1–18) is low, due to absorption by chlorophyll for photosynthesis [55]. In the near-infrared region (band 77–103), the reflectance is much higher than that in the visible bands due to the cellular structure in the vegetation [55]. This property is in accordance with the spectral curves of the original data as shown in Fig. 8 (a). By contrast, adversarial attacks successfully change the spectral reflectance property of meadows in the adversarial example. The reflectance in the blue region gets strengthened while the reflectance in the near-infrared region is weakened as shown in Fig. 8 (b) and (c). These changes

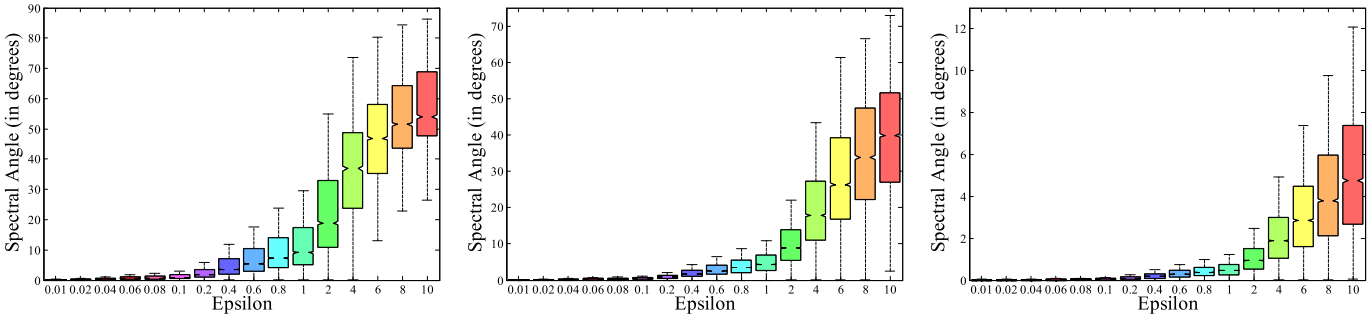

Fig. 9. The boxplot of the spectral angles between the adversarial example and the original hyperspectral data with different values of ϵ. (a) Pavia University dataset. (b) Houston dataset. (c) Salinas dataset.

TABLE IX
MEAN SPECTRAL ANGLES BETWEEN THE ADVERSARIAL EXAMPLES AND THE ORIGINAL HYPERSPECTRAL DATA WITH DIFFERENT VALUES OF ϵ. ANGLES ARE REPORTED IN DEGREES

ϵ	0.01	0.02	0.04	0.06	0.08	0.1	0.2	0.4	0.6	0.8	1	2	4	6	8	10
Pavia University	0.166	0.330	0.650	0.963	1.268	1.568	3.005	5.657	8.108	10.411	12.599	22.283	36.757	46.032	51.923	55.467
Houston	0.116	0.221	0.407	0.572	0.723	0.866	1.512	2.686	3.815	4.927	6.027	11.303	20.394	27.740	33.462	37.710
Salinas	0.009	0.014	0.026	0.041	0.054	0.069	0.140	0.280	0.418	0.555	0.691	1.353	2.602	3.775	4.887	5.943

bring about a challenge for deep neural networks to reveal the real categories of adversarial examples.

We further take the spectral angle as the metric [56] to quantitatively measure the distance between the generated adversarial examples and the original hyperspectral image. Let $x^i \in \mathbb{R}^c$ and $x^i_{adv} \in \mathbb{R}^c$ denote the spectral vectors of the ith pixel in the original hyperspectral image and the adversarial image, respectively, where $i = 1, 2, \ldots, hw$. Then, the spectral angle θ^i between x^i and x^i_{adv} can be formulated as:

$$\theta^i = \cos^{-1} \frac{\sum_{j=1}^{c} x^i_{(j)} \cdot x^i_{adv(j)}}{\sqrt{\sum_{j=1}^{c} \left(x^i_{(j)}\right)^2} \sqrt{\sum_{j=1}^{c} \left(x^i_{adv(j)}\right)^2}}, \quad (14)$$

where $\cos^{-1}$ denotes the arc-cosine function. For each dataset used in this study, we use the boxplot to analyze the statistical distribution of spectral angles with different values of ϵ as shown in Fig. 9. Each boxplot in Fig. 9 shows the minimum (the bottom bar), the maximum (the top bar), the median (the notch), and the first and third quartiles of the spectral angles. It can be observed that as the ϵ increases, the spectral angles also get larger. Besides, a lager ϵ also leads to a greater magnitude of fluctuation on spectral angles, especially when $\epsilon > 1$.

We further report the mean spectral angle $\bar{\theta} = \frac{\sum_{i=1}^{hw} \theta^i}{hw}$ of each dataset with different values of ϵ in Table IX. An intriguing discovery is that the magnitude of the mean spectral angle is highly related to the number of bands c in the hyperspectral data. A larger c generally corresponds to a smaller mean spectral angle. As described in Section IV-A, there are 103, 144, 204 bands in the Pavia University, Houston, and Salinas datasets, respectively. Accordingly, the mean spectral angles of these three datasets are in descending order under the same adversarial perturbation. Take $\epsilon = 1$ for example. While the mean spectral angle of the Pavia University is about 12.59 degrees, the Houston and Salinas datasets only have a mean spectral angle of about 6.02, and 0.69 degrees, respectively. One possible explanation for this phenomenon may lie in the fact that adversarial perturbations are shared by all channels in the hyperspectral data. With a larger number of bands c, each channel in the image may get a smaller adversarial perturbation under the same level of adversarial attacks. Thus, the mean spectral angle may also be smaller in this case.

V. CONCLUSION AND DISCUSSIONS

Deep learning-based methods have achieved great success in the HSI classification task. Nevertheless, their vulnerability towards adversarial attacks could not be neglected. In this study, we systematically analyze the influence of adversarial attacks on the HSI classification task for the first time. While existing research of adversarial attacks focuses on the generation of adversarial examples in the RGB domain, this study reveals that such adversarial examples could also exist in the hyperspectral domain. Although the difference between the adversarial examples and the original hyperspectral data could be imperceptible for the human visual system, most of the state-of-the-art deep learning-based methods may get seriously cheated to make wrong predictions. To address this challenge, a novel self-attention context network (SACNet) is proposed in this study. Extensive experiments on three benchmark HSI datasets demonstrate that the proposed SACNet possesses stronger resistibility towards adversarial examples compared with the existing state-of-the-art deep learning models. We further analyze the generated hyperspectral adversarial examples visually and statistically. The experiments reveal that adversarial attacks can successfully change the spectral reflectance characteristics of different categories in the adversarial examples.

While directly improving the inherent resistibility of deep neural networks is one way to defend against adversarial attacks, another possible coping strategy may be adversarial

example detection, which attempts to detect whether a given remote sensing image is attacked by adversarial perturbations. Although the visual difference between the original HSI and the adversarial example could be imperceptible, our study shows that the spectral angle may be a good clue for hyperspectral adversarial example detection. We will investigate this issue in our future work. Besides, whether the proposed SACNet would also help to defend from adversarial attacks for RGB images with adequate adaptation is another interesting research question. We will try to tackle this issue in our future work.

REFERENCES

[1] Z. Han *et al.*, "Deep spatiality: Unsupervised learning of spatially-enhanced global and local 3D features by deep neural network with coupled softmax," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 3049–3063, Jun. 2018.

[2] H. Lee and H. Kwon, "Going deeper with contextual CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4843–4855, Oct. 2017.

[3] M. Zhang, W. Li, and Q. Du, "Diverse region-based CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2623–2634, Jun. 2018.

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

[5] B. Guo, S. R. Gunn, R. I. Demper, and J. D. B. Nelson, "Customizing kernel functions for SVM-based hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 17, no. 4, pp. 622–629, Apr. 2008.

[6] K. Bernard, Y. Tarabaika, J. Angulo, J. Chanussot, and J. A. Benediktsson, "Spectral–spatial classification of hyperspectral data based on a stochastic minimum spanning forest approach," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2008–2021, Apr. 2012.

[7] H. Wu and S. Prasad, "Semi-supervised deep learning using pseudo labels for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1259–1270, Mar. 2018.

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

[9] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2014.

[10] Y. Chen, X. Zhao, and X. Jia, "Spectral–spatial classification of hyperspectral data based on deep belief network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2381–2392, Jun. 2015.

[11] L. Mou, P. Ghamisi, and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3639–3655, Jul. 2017.

[12] Y. Xu, L. Zhang, B. Du, and F. Zhang, "Spectral–spatial unified networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 5893–5909, Oct. 2018.

[13] X. X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[14] Y. Xu, B. Du, F. Zhang, and L. Zhang, "Hyperspectral image classification via a random patches network," *ISPRS J. Photogramm. Remote Sens.*, vol. 142, pp. 344–357, Aug. 2018.

[15] W. Shao 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, Oct. 2016.

[16] 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.

[17] Y. Li, H. Zhang, and Q. Shen, "Spectral–spatial classification of hyperspectral imagery with 3D convolutional neural network," *Remote Sens.*, vol. 9, no. 1, p. 67, 2017.

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

[19] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. J. Plaza, and F. Pla, "Deep pyramidal residual networks for spectral–spatial hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 740–754, Feb. 2019.

[20] Y. Xu, B. Du, and L. Zhang, "Beyond the patchwise classification: Spectral-spatial fully convolutional networks for hyperspectral image classification," *IEEE Trans. Big Data*, vol. 6, no. 3, pp. 492–506, Sep. 2020.

[21] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[22] Y. Xu, B. Du, and L. Zhang, "Assessing the threat of adversarial examples on deep neural networks for remote sensing scene classification: Attacks and defenses," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1604–1617, Feb. 2021.

[23] C. Szegedy *et al.*, "Intriguing properties of neural networks," 2013, *arXiv:1312.6199*. [Online]. Available: http://arxiv.org/abs/1312.6199

[24] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "Adversarial attacks and defences: A survey," 2018, *arXiv:1810.00069*. [Online]. Available: http://arxiv.org/abs/1810.00069

[25] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*. [Online]. Available: http://arxiv.org/abs/1412.6572

[26] A. Arnab, O. Miksik, and P. H. S. Torr, "On the robustness of semantic segmentation models to adversarial attacks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 888–897.

[27] X. He, S. Yang, G. Li, H. Li, H. Chang, and Y. Yu, "Non-local context encoder: Robust biomedical image segmentation against adversarial attacks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8417–8424.

[28] W. Czaja, N. Fendley, M. Pekala, C. Ratto, and I.-J. Wang, "Adversarial examples in remote sensing," in *Proc. 26th ACM SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, 2018, pp. 408–411.

[29] L. Chen, G. Zhu, Q. Li, and H. Li, "Adversarial example in remote sensing image recognition," 2019, *arXiv:1910.13222*. [Online]. Available: http://arxiv.org/abs/1910.13222

[30] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14410–14430, 2018.

[31] P. Ghamisi *et al.*, "Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 37–78, Dec. 2017.

[32] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1979–1993, Aug. 2019.

[33] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," 2016, *arXiv:1611.01236*. [Online]. Available: http://arxiv.org/abs/1611.01236

[34] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2017, *arXiv:1706.06083*. [Online]. Available: http://arxiv.org/abs/1706.06083

[35] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 39–57.

[36] H. Xu *et al.*, "Adversarial attacks and defenses in images, graphs and text: A review," *Int. J. Autom. Comput.*, vol. 17, no. 2, pp. 151–178, Apr. 2020.

[37] V. Zantedeschi, M.-I. Nicolae, and A. Rawat, "Efficient defenses against adversarial attacks," in *Proc. 10th ACM Workshop Artif. Intell. Secur.*, 2017, pp. 39–49.

[38] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," 2017, *arXiv:1705.07204*. [Online]. Available: http://arxiv.org/abs/1705.07204

[39] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1765–1773.

[40] S. Gu and L. Rigazio, "Towards deep neural network architectures robust to adversarial examples," 2014, *arXiv:1412.5068*. [Online]. Available: http://arxiv.org/abs/1412.5068

[41] C. Lyu, K. Huang, and H.-N. Liang, "A unified gradient regularization family for adversarial examples," in *Proc. IEEE Int. Conf. Data Mining*, Nov. 2015, pp. 301–309.

[42] P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-GAN: Protecting classifiers against adversarial attacks using generative models," 2018, *arXiv:1805.06605*. [Online]. Available: http://arxiv.org/abs/1805.06605

[43] J. Hoffman, D. A. Roberts, and S. Yaida, "Robust learning with Jacobian regularization," 2019, *arXiv:1908.02729*. [Online]. Available: http://arxiv.org/abs/1908.02729
[44] K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. McDaniel, "On the (Statistical) detection of adversarial examples," 2017, *arXiv:1702.06280*. [Online]. Available: http://arxiv.org/abs/1702.06280
[45] Z. Gong, W. Wang, and W.-S. Ku, "Adversarial and clean data are not twins," 2017, *arXiv:1704.04960*. [Online]. Available: http://arxiv.org/abs/1704.04960
[46] R. Feinman, R. R. Curtin, S. Shintre, and A. B. Gardner, "Detecting adversarial samples from artifacts," 2017, *arXiv:1703.00410*. [Online]. Available: http://arxiv.org/abs/1703.00410
[47] D. J. Miller, Z. Xiang, and G. Kesidis, "Adversarial learning targeting deep neural network classification: A comprehensive review of defenses against attacks," *Proc. IEEE*, vol. 108, no. 3, pp. 402–433, Mar. 2020.
[48] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," 2018, *arXiv:1803.02155*. [Online]. Available: http://arxiv.org/abs/1803.02155
[49] Y. Xu, B. Du, L. Zhang, Q. Zhang, G. Wang, and L. Zhang, "Self-ensembling attention networks: Addressing domain shift for semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 5581–5588.
[50] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2018, pp. 7794–7803.
[51] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.
[52] H. Zhang *et al.*, "Context encoding for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7151–7160.
[53] C. Debes *et al.*, "Hyperspectral and lidar data fusion: Outcome of the 2013 grss data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2405–2418, 2014.
[54] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sensors*, vol. 2015, pp. 1–12, 2015.
[55] J. E. Colwell, "Vegetation canopy reflectance," *Remote Sens. Environ.*, vol. 3, no. 3, pp. 175–183, Jan. 1974.
[56] P. E. Dennison, K. Q. Halligan, and D. A. Roberts, "A comparison of error metrics and constraints for multiple endmember spectral mixture analysis and spectral angle mapper," *Remote Sens. Environ.*, vol. 93, no. 3, pp. 359–367, 2004.

Yonghao Xu (Member, IEEE) received the B.S. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2016, where he is currently pursuing the Ph.D. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing.

His research interests include remote sensing, computer vision, and machine learning.

Bo Du (Senior Member, IEEE) received the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2010.

He is currently a Professor with the School of Computer Science and the Institute of Artificial Intelligence, Wuhan University. He is also the Director of the National Engineering Research Center for Multimedia Software, Wuhan University. He has more than 80 research articles published in the IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), IEEE TRANSACTIONS ON CYBERNETICS (TCYB), IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS), IEEE JOURNAL OF SELECTED TOPICS IN EARTH OBSERVATIONS AND APPLIED REMOTE SENSING (JSTARS), and IEEE GEOSCIENCE AND REMOTE SENSING LETTERS (GRSL), and 13 of them are ESI hot papers or highly cited papers. His major research interests include pattern recognition, hyperspectral image processing, and signal processing.

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

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

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

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