IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 15, 2021, accepted April 22, 2021, date of publication April 28, 2021, date of current version May 10, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3076225

Adversarially Robust Hyperspectral Image Classification via Random Spectral Sampling and Spectral Shape Encoding

SUNGJUNE PARK, HONG JOO LEE, AND YONG MAN RO, (Senior Member, IEEE)
School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, South Korea

Corresponding author: Yong Man Ro (ymro@kaist.ac.kr)

This work was conducted by Center for Applied Research in Artificial Intelligence (CARAI) grant funded by DAPA and ADD (UD190031RD).

ABSTRACT Although the hyperspectral image (HSI) classification has adopted deep neural networks (DNNs) and shown remarkable performances, there is a lack of studies of the adversarial vulnerability for the HSI classifications. In this paper, we propose a novel HSI classification framework robust to adversarial attacks. To this end, we focus on the unique spectral characteristic of HSIs (*i.e.,* distinctive spectral patterns of materials). With the spectral characteristic, we present the random spectral sampling and spectral shape feature encoding for the robust HSI classification. For the random spectral sampling, spectral bands are randomly sampled from the entire spectrum for each pixel of the input HSI. Also, the overall spectral shape information, which is robust to adversarial attacks, is fed into the shape feature extractor to acquire the spectral shape feature. Then, the proposed framework can provide the adversarial robustness of HSI classifiers via randomization effects and spectral shape feature encoding. To the best of our knowledge, the proposed framework is the first work dealing with the adversarial robustness in the HSI classification. In experiments, we verify that our framework improves the adversarial robustness considerably under diverse adversarial attack scenarios, and outperforms the existing adversarial defense methods.

INDEX TERMS Adversarial robustness, hyperspectral image classification, random spectral sampling, spectral shape encoding.

I. INTRODUCTION

Hyperspectral images (HSIs) capture hundreds of abundant spectral information of materials with narrow wavelength band intervals (*e.g.,* 5-10 nm). Therefore, HSIs contain a discriminative spectral characteristic across the wavelength for each material [1]–[3]. Such an advantage of rich spectral information can help the HSI classification to identify every pixel of HSI (*i.e.,* ground objects in HSI), and it has been applied into various applications, such as environment management, medical diagnosis, and ground surveillance [4]–[8].

Due to the usefulness of HSIs, the HSI classification has been studied broadly, and deep neural networks (DNNs) have accelerated its improvement [2], [9]–[16]. For the first HSI classification framework adopting DNNs, Chen *et al.* [14] combined principal component analysis (PCA) with DNNs to extract deep features of HSIs. They merged both spatial and spectral features together to leverage each feature. Li *et al.* [2] utilized fully convolutional neural networks (CNNs) with deconvolutional and pooling layers to achieve a hyperspectral feature enhancement. Also, they proposed an optimization method to boost the classification performance. Furthermore, in the recent days, 2D and 3D CNNs have been adopted to exploit neighboring spatial and spectral features together, focusing on improving the performances [9], [12], [13], [17]. Moreover, Hong *et al.* [18] presented an augmented linear mixing model (ALMM) to deal with the HSI unmixing problem. Hong *et al.* modeled the main spectral variability, scaling factors, by the endmember dictionary, and other spectral variabilities which come from environmental effects. A prior knowledge for the spectral variability is also designed for effective data-driven learning. In [19], minibatch graph convolutional network (miniGCN) is developed by exploring the relations between each sample. Hong *et al.* [19] also

The associate editor coordinating the review of this manuscript and approving it for publication was Qingli Li.

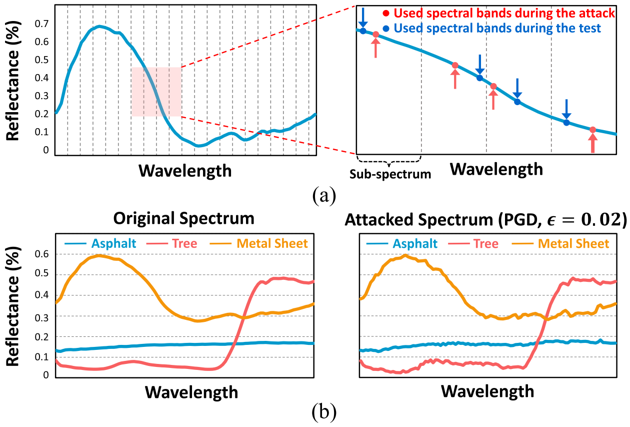
investigated CNN and GCN feature fusion with various strategies (*i.e.,* additive, elementwise multiplicative, and concatenation fusion). Yao *et al.* [20] explored the non-local smoothness property of HSIs and proposed non-local HSI total variation (NLHTV) regularizer, enabling the spatial distribution of different endmembers to be diverse.

Although the HSI classification has achieved remarkable improvements with DNNs, DNNs have a serious weakness. It has been widely known that DNNs are vulnerable to adversarial examples [21]–[24]. Adversarial examples are the data perturbed by imperceptible noise, while they are indistinguishable from clean data. Such examples induce DNNs to perform incorrect predictions. To deal with the problem, various defense methods have been studied, (*e.g.,* adversarial training [22], [25], randomization [26], [27], and ensemble model training [28], *etc.*). Adversarial training [22], [25] is one of the well-known defense strategies. Adversarial training uses adversarial examples for training DNNs and improves the adversarial robustness. Lee *et al.* [26] trained several ensemble networks, and randomly sampled each layer from different networks for randomization effects. Pang *et al.* [28] trained ensemble networks to promote diversity of predictions achieving the adversarial robustness.

Most of aforementioned adversarial defense methods focus on DNN classifiers of the general image domains (*e.g.,* RGB domain). Although the threat of adversarial examples on HSI has been raised [29], there is a lack of research to develop the robust HSI classifier against adversarial attacks. Since the HSI classification is used in safety and security-critical applications such as artwork authentication [30], [31], medical diagnosis [32]–[34], and homeland securities [35], [36], defending against adversarial attacks is necessary to adopt DNNs in the real-world. For example, in the case of homeland security application, attackers could conceal petards, panzers, or submarines.

Even though the existing defense methods can be applied into HSI classifiers [22], [25], [26], [28], they cannot fully exploit the advantage of HSIs (*i.e.,* unique spectral information). Moreover, when applying the existing adversarial defense methods to the HSI classification frameworks, they have the limitations: large increase in training time and network parameters. In order to alleviate such limitations and improve adversarial robustness at the same time, leveraging the spectral characteristic of HSIs could be one of the effective approaches.

In this paper, we introduce a novel HSI classification framework robust to adversarial attacks. In the proposed method, we exploit the spectral characteristic of HSIs, to tackle the aforementioned limitations (*i.e.,* more training time and parameters) at the same time. The key idea of the proposed framework is to make adversaries not being aware of which spectral bands to be used during test time. Also, we focus on exploiting overall shape of spectral bands to improve the adversarial robustness. To this end, we propose novel random spectral sampling and spectral shape feature encoding. The proposed framework consists of two feature encoding paths; patch feature and spectral shape feature encoding paths. The patch feature encoding path takes the input patch cube (*i.e.,* H×W×# of bands) including a target center pixel with its neighboring pixels. Here, we propose the random spectral sampling to sample random spectral bands, not encoding the entire input patch directly. It is designed to improve the adversarial robustness via randomization effects. For each pixel, we first decompose the entire spectrum (*e.g.,* 100 bands) into several sub-spectrum (*e.g.,* 5 bands for each sub-spectrum). Then, one spectral band is sampled categorically from each sub-spectrum, preserving the overall shape of entire spectrum. After that, random sampled HSI patch is fed into the patch feature extractor. With the proposed random spectral sampling, as shown in Figure 1(a), some sampled spectral bands (indicated by red dots) are perturbed during adversarial attack generation. However, other sampled spectral bands (indicated by blue dots) could secure the adversarial robustness of the HSI classifiers. Also, even though we sample spectral bands, we could preserve the overall pattern of spectral bands.



(a)

(b)

FIGURE 1. The examples of spectral bands from university of pavia HSI dataset. (a) Describes how the proposed random sampling works. (b) Shows that the overall shapes of spectrum are not changed largely under adversarial attack.

Furthermore, we propose the spectral shape feature encoding to leverage overall shape of spectral bands. Since the adversarial attacks perturb the input data by adding small noise, the perturbed noise could not change the overall (*increasing/decreasing*) shape of the entire spectrum. For example, as shown in Figure 1(b), the attacked spectra (right) are not deformed from each original spectrum (left). To leverage the overall shape of spectrum, we extract the overall shape information of the target pixel's spectral bands from the differences between the spectral bands. Then, we feed it into the spectral shape feature extractor. Since we employ the overall shape information of the spectral bands, which is not changed largely even with adversarial attacks, the proposed spectral shape feature encoding helps the HSI classifiers to be robust against adversarial attacks.

In the experiments, we verify the effectiveness of the proposed approach with various HSI classification networks

(2D- and 3D-CNNs). The experiments demonstrate that the proposed framework improves the adversarial robustness significantly for the HSI classification, outperforming the existing adversarial defense methods on public datasets (University of Pavia and Salinas), under various adversarial attack scenarios (FGSM, PGD, CW, Adaptive attack, Expectation over transformation, and black-box attack). Moreover, the proposed approach can be applied into the existing HSI classification frameworks with small modifications.

Our contributions can be summarized as follows:

- To the best of our knowledge, it is the first work dealing with the adversarial robustness for HSI classification, exploiting the unique characteristic of spectral information in HSIs.
- We present 1) random spectral sampling and 2) spectral shape feature encoding to improve adversarial robustness of the HSI classifiers considerably without large increase of training time and parameters.
- We demonstrate the effectiveness of the proposed framework with the general HSI classification networks and public HSI datasets under various adversarial attack scenarios.

II. RELATED WORK

A. ADVERSARIAL ATTACKS

It has been widely known that DNNs are highly vulnerable to adversarial attacks making the networks conduct incorrect predictions. Adversaries generate adversarial examples by maximizing the loss values to fool the networks. Fast Gradient Sign Method (FGSM) [37] is a simple and fast attack method. It generates adversarial examples by perturbing the intensity of the pixel without any iteration. As an extension of FGSM, Projected Gradient Descent (PGD) [22] is proposed. It conducts FGSM iteratively with a small step size. Carlini & Wagner (C&W) [38] attack is optimization based adversarial attack method optimized with the attack objective function to generate adversarial perturbations. It generates adversarial perturbations that changes the logit values with a minimal perturbation. Burnel *et al.* [23] propose a natural hyperspectral adversarial example generation method with Wasserstein GAN [39]. It randomly generates adversarial examples by pre-trained generator for the specific class. In the experiment, we verify the effectiveness of the proposed method with widely used adversarial attack benchmark methods (FGSM, PGD, and C&W). Through the experiment, we verify that existing adversarial attack methods could sufficiently fool the existing HSI classification network while the proposed method effectively defends such attacks.

B. ADVERSARIAL ATTACK DEFENSE

To tackle the vulnerability against adversarial attacks, various research have been studied, such as adversarial training (AT) [22], [25], [37], ensemble model training [28], [40], and randomization [26], [41]. As the early adversarial training approach, [37] generates adversarial examples generated by FGSM attacks, and trains the networks with them to improve the adversarial robustness. However, it has the problem, not effective with iterative attacks (*e.g.,* PGD). Therefore, Madry *et al.* [22] proposed another adversarial training using the PGD attack. Such PGD-based adversarial training methods show great robustness performances, even if the network is exposed to other types of attack. Another way to improve the adversarial robustness is the ensemble model training methods [28]. While training several networks, adaptive diversity promoting (ADP) regularization [28] is proposed to make non-maximal values of each network prediction to be orthogonal. Also, Lee *et al.* [26] use the randomization method, and present a random layer sampling (RLS) which selects network layers from several trained networks during testing. By generating different networks with RLS, it makes the attacker could not effectively generate adversarial examples. However, when applying these defense methods into the HSI classification frameworks, they have the limitations. The adversarial training methods require very large training time, because it generates adversarial examples consistently during the network training. The other methods need to train more than one network, so that, it should cause much increase of network parameters. In this paper, we present a novel HSI classification framework which is robust to adversarial attacks. Moreover, it is so efficient that it alleviates the limitations of previous works by leveraging the spectral characteristic of HSIs effectively.

C. HYPERSPECTRAL IMAGE CLASSIFICATION

Hyperspectral image (HSI) classification is one of the most active fields in remote sensing community, and it has shown a great success with the advancement of DNNs. Chen *et al.* [42] proposed deep CNN architectures to extract spectral-spatial HSI features. By combining both spectral-spatial features, they achieved great performance in HSI classification. Haut *et al.* [17] presented a 2D-CNN residual-attention network to characterize spectral-spatial information effectively. Through the residual-attention network, they solved an overfitting problem and achieved high classification accuracy. Also, Wang *et al.* [13] adopted 3D-CNN to construct a fast and dense convolution network composed of spectral and spatial encoding modules. Furthermore, generative adversarial networks (GANs) have been adopted for HSI classification. Zhu *et al.* [43] explored the usefulness of GANs for HSI classification and designed 1D and 3D GANs as spectral and spectral-spatial classifiers. The synthetic HSI samples are used for data augmentation to improve classification accuracy. Zhong *et al.* [44] proposed GAN and conditional random field (GAN-CRF) based framework to integrate a deep learning and a probabilistic graphical model for HSI classification. In [44], dense CRFs are designed to give graph constraints to achieve the classification accuracy improvement. However, most of the aforementioned methods focused on developing high accuracy networks or learning methods. There is a lack of research to improve the adversarial robustness for the HSI classification. Since HSIs are widely used in

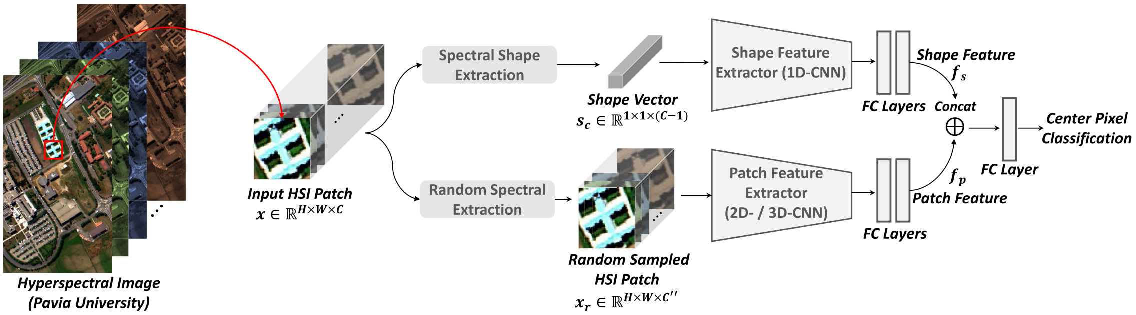

Hyperspectral Image
(Pavia University)
Input HSI Patch
$x \in \mathbb{R}^{H\times W\times C}$
Spectral Shape Extraction
Random Spectral Extraction
Shape Vector
$s_c \in \mathbb{R}^{1\times 1\times (C-1)}$
Random Sampled HSI Patch
$x_r \in \mathbb{R}^{H\times W\times C''}$
Shape Feature Extractor (1D-CNN)
Patch Feature Extractor (2D- / 3D-CNN)
FC Layers
Shape Feature
f_s
Concat
FC Layer
Center Pixel Classification
f_p
Patch Feature
FC Layers

FIGURE 2. Overall architecture of the proposed framework. It is composed of the two feature encoding paths; patch feature encoding path and spectral shape feature encoding path. The patch feature encoding path promotes randomization effects via random spectral extraction, and the spectral shape feature encoding path extracts the shape feature of target pixel's spectrum. Both paths are designed to improve the adversarial robustness of the HSI classifiers.

many important applications such as artwork authentication [30], [31], and homeland securities [35], [36], and medical diagnosis [32]–[34], the robust and reliable HSI classification framework needs to be considered. In this paper, we reveal the problem of adversarial vulnerability on the HSI classifiers, and propose a novel HSI classification framework to achieve the adversarial robustness using the spectral information.

III. PROPOSED METHOD

A. PRELIMINARIES

Before we introduce the proposed framework, we provide a brief description of adversarial attacks. Let $\boldsymbol{x}$ be the input of the network, $\boldsymbol{y}$ be the target label corresponding to the input $\boldsymbol{x}$, $\boldsymbol{\theta}$ the network parameters, and $L(\boldsymbol{\theta}, \boldsymbol{x}+\boldsymbol{r}, \boldsymbol{y})$ be the loss function used to train the network. The main goal of the adversarial attack is to generate perturbation $\boldsymbol{r}$ that maximizes the loss $L(\boldsymbol{\theta}, \boldsymbol{x}+\boldsymbol{r}, \boldsymbol{y})$. At the same time, we anticipate $\boldsymbol{x_{adv}}(= \boldsymbol{x}+\boldsymbol{r})$ to be indistinguishable from the clean input $\boldsymbol{x}$ by giving the constraint $\|\boldsymbol{r}\| \leq \epsilon$.

For fast gradient sign method (FGSM) [37], it generates adversarial examples using the gradient of the loss function with a single step. It generates adversarial examples that increase the gradient of loss function. The equation can be described as follows,

$$\boldsymbol{x_{adv}} = \boldsymbol{x} + \epsilon \; sign(\nabla_{\boldsymbol{x}} L(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y})), \tag{1}$$

where ϵ is the hyperparameter to manipulate the magnitude of perturbations. It increases the loss function linearly.

Projected gradient descent (PGD) [22] is another attack algorithm similar to FGSM, but it is a more powerful attack. It generates adversarial perturbations with multi-steps (*iterative*) as follows,

$$\boldsymbol{x_{adv}}^{t+1} = \Pi_{\boldsymbol{x}+\boldsymbol{S}}[\boldsymbol{x}^t + \alpha \; sign(\nabla_{\boldsymbol{x}} L(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y}))], \tag{2}$$

where $\Pi_{\boldsymbol{x}+\boldsymbol{S}}$ is to project perturbations within bounded region $\boldsymbol{x} + \boldsymbol{S}$, t means iteration step, and α is the step size to control the magnitude of perturbations for each step. PGD also holds the constraint, $\|\boldsymbol{r}\| \leq \epsilon$, to limit the maximum size of perturbations, where $\boldsymbol{S} = [-\epsilon, \epsilon]^D$.

Carlini & Wagner (CW) [38] attack algorithm optimizes the below equations to generate adversarial perturbations,

$$minimize \; \|\boldsymbol{x_{adv}} - \boldsymbol{x}\| + c \cdot l(\boldsymbol{x_{adv}}), \tag{3}$$

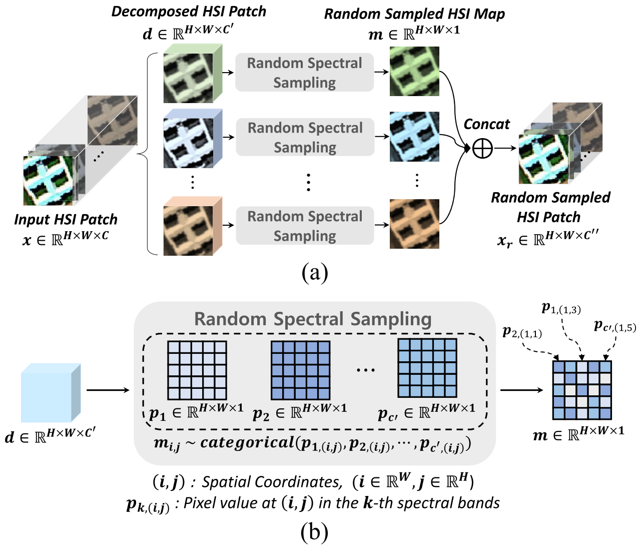
$$\begin{aligned} l(\boldsymbol{x_{adv}}) = max[&max\{Z(\boldsymbol{x_{adv}})_{\tilde{c}} : \tilde{c} \neq y_t\} \\ &- Z(\boldsymbol{x_{adv}})_{y_t}, -\kappa], \end{aligned} \tag{4}$$

where c denotes a constant chosen by a binary search, Z is a logit value, $\tilde{c}$ is one of the classes except target class, y_t, and $-\kappa$ is a fixed parameter to limit the maximum. It generates adversarial examples that changes the logit values with a minimal perturbation.

The above attacks can be categorized into white-box attack methods that the adversary has the whole information about the network, (*e.g.*, network architecture, network's weights, loss functions used in network training, *etc.*). Since white-box attack methods know the whole information about the network, it is crucial for DNNs. In the experiment, we verify the effectiveness of the proposed method with these white-box attack benchmarks.

B. OVERALL PROPOSED FRAMEWORK

The overall architecture of the proposed framework is shown in Figure 2. As the general HSI classification framework does [17], [42], [45], the proposed framework takes HSI patch, $\boldsymbol{x}$, for the input. The input patch is encoded through two paths (*i.e.*, patch feature encoding path and spectral shape feature encoding path). For the patch feature encoding, we generate random sampled HSI patch, $\boldsymbol{x_r}$, through the proposed random spectral sampling. $\boldsymbol{x_r}$ is fed into the patch feature extractor and fully connected (FC) layers to acquire the patch feature, $\boldsymbol{f_p}$. The patch feature extractor is composed of 2D or 3D convolution and pooling layers, which are usually used for the general HSI classifiers. Also, for the spectral shape feature encoding, we extract the spectral shape vector, $\boldsymbol{s}_c$, of the (target) center pixel from the input HSI patch $\boldsymbol{x}$. The spectral shape feature, $\boldsymbol{f_s}$, is obtained by feeding $\boldsymbol{s}_c$ into the shape feature extractor and FC layers. Then, we acquire two features, $\boldsymbol{f_p}$ and $\boldsymbol{f_s}$. The patch feature, $\boldsymbol{f_p}$, contains target and neighboring pixels' information for classification, and it is robust to adversarial attacks via randomization effects.


Decomposed HSI Patch
$d \in \mathbb{R}^{H\times W\times C'}$
Random Sampled HSI Map
$m \in \mathbb{R}^{H\times W\times 1}$
Random Spectral Sampling
Random Spectral Sampling
Random Spectral Sampling
Concat
Input HSI Patch
$x \in \mathbb{R}^{H\times W\times C}$
Random Sampled HSI Patch
$x_r \in \mathbb{R}^{H\times W\times C''}$
Random Spectral Sampling
$p_1 \in \mathbb{R}^{H\times W\times 1}$ $p_2 \in \mathbb{R}^{H\times W\times 1}$ $p_{c'} \in \mathbb{R}^{H\times W\times 1}$
$m_{i,j} \sim categorical(p_{1,(i,j)}, p_{2,(i,j)}, \cdots, p_{c',(i,j)})$
$d \in \mathbb{R}^{H\times W\times C'}$
$p_{1,(1,3)}$
$p_{2,(1,1)}$
$p_{c',(1,5)}$
$m \in \mathbb{R}^{H\times W\times 1}$
(i,j) : Spatial Coordinates, $(i \in \mathbb{R}^W, j \in \mathbb{R}^H)$
$p_{k,(i,j)}$: Pixel value at (i,j) in the k-th spectral bands

(a)

(b)

FIGURE 3. (a) Shows the overall flow of the random spectral extraction, how to extract random spectral bands. We first decompose the input HSI patch, x, and then, conduct random spectral sampling as shown in (b). A single spectral channel is selected from C' number of channels, per each pixel. After that, random sampled HSI patch, x_r, is acquired by combining random sampled HSI maps, m.

The spectral shape feature, $\boldsymbol{f_s}$, holds the overall shape information of target pixel's spectrum, which is not deformed largely under adversarial attacks, so that, $\boldsymbol{f_s}$ could have robust property against the attacks. Finally, both robust features, $\boldsymbol{f_p}$ and $\boldsymbol{f_s}$, are combined to predict the class of (target) center pixel. Note that, since the patch feature extractor is constructed as the general HSI classifiers, the proposed approach can be applied into the existing frameworks. The details of the proposed framework are described in the following sections.

C. RANDOM SPECTRAL SAMPLING

In the general HSI classification frameworks, the network takes HSI patch for its input to classify the center pixel. In this setting, since each pixel has a unique spectral pattern of material, we focus on two points: 1) preserving its spectral pattern to classify HSIs correctly, and 2) taking randomization effects to improve the adversarial robustness at the same time. To this end, we present random spectral sampling to extract random spectral bands. Figure 3(a) describes the overall flow of the random spectral extraction. As shown in Figure 3(a), we decompose the entire input HSI patch, $\boldsymbol{x} \in \mathbb{R}^{H\times W\times C}$, into the C'' number of small patches across the spectral direction (sub-spectrum). Each of them can be expressed as $\boldsymbol{d} \in \mathbb{R}^{H\times W\times C'}$, where $C' = \lfloor C/C'' \rfloor$. Then, for each decomposed HSI patch, $\boldsymbol{d}$, we conduct the proposed random spectral sampling, as described in Figure 3(b). Given each pixel, a single spectral channel is selected from C' number of channels by random spectral sampling. Each $\boldsymbol{d}$ is composed of the C' number of spatial maps, $\boldsymbol{p}$, that is, $\boldsymbol{d} = \boldsymbol{p}_1 \oplus \cdots \oplus \boldsymbol{p}_{c'}$, where $\oplus$ means a concatenation in the spectral direction, $\boldsymbol{p}_k \in \mathbb{R}^{H\times W\times 1}$, and $k \in \{1, 2, \cdots, C'\}$. From each decomposed HSI patch, $\boldsymbol{d}$, we create a random sampled HSI map, $\boldsymbol{m} \in \mathbb{R}^{H\times W\times 1}$. Regarding the spatial coordinate (i, j) for each random sampled HSI map $\boldsymbol{m}$, we randomly sample a single value from the C' number of spatial maps $\boldsymbol{p}$ at the corresponding pixel of (i, j). In other words, for each pixel, we sample a single spectral bands out of the C' number of spectral bands as follows,

$$\boldsymbol{m}_{i,j} \sim categorical(\boldsymbol{p}_{1,(i,j)}, \boldsymbol{p}_{2,(i,j)}, \cdots, \boldsymbol{p}_{c',(i,j)}; \boldsymbol{q}_{1,(i,j)}, \boldsymbol{q}_{2,(i,j)}, \cdots, \boldsymbol{q}_{c',(i,j)}), \qquad (5)$$

where $\boldsymbol{q}$ represents the probability of each spectral band to be sampled, and each $\boldsymbol{q}$ is fixed as $1/C'$ to have the uniform distribution. From Equation (5), $\boldsymbol{m}_{i,j}$ is sampled from the set of $\{\boldsymbol{p}_{1,(i,j)}, \boldsymbol{p}_{2,(i,j)}, \cdots, \boldsymbol{p}_{c',(i,j)}\}$ with the same probability. Note that, the random spectral sampling of each pixel value is conducted independently, so that, the created HSI map, $\boldsymbol{m}$, is expected to contain varied spectral information from different spectral bands. Then, the whole set of $\boldsymbol{m}$ is concatenated to acquire random sampled HSI patch, $\boldsymbol{x_r} \in \mathbb{R}^{H\times W\times C''}$, which is fed into the patch feature extractor.

The proposed random spectral sampling preserves the distinctive spectral patterns, and makes adversaries not being aware of which spectral bands to be used during the inference time. In other words, since the random sampled patch, $\boldsymbol{x_r}$, is acquired with the different set of $\boldsymbol{m}$ in every inference time, the adversarial examples, generated with some specific set of $\boldsymbol{m}$, would not be effective that they could not affect the network performance. As a result, the adversarial robustness of the HSI classification network can be achieved. Also, we use the patch feature extractor as the general HSI classifiers (2D or 3D-CNNs). Therefore, the proposed random spectral sampling can be applied into the existing HSI classifiers by considering the modification of the patch's spectral dimension only (because it reduces the spectral dimension by random spectral sampling). Note that the proposed random spectral sampling is conducted during both network training and inference times. Therefore, the network could take various channel combinations for the same spectral sample during the network training, which could avoid the overfitting issue.

D. SPECTRAL SHAPE FEATURE ENCODING

Adversarial attacks generate adversarial examples which are indistinguishable from the clean data. Therefore, adversaries apply the small size of attack perturbations within ϵ. In this point of view, we focus on the overall (*increasing/decreasing*) shape information of spectral bands, which is not changed largely by adversarial attacks. Note that, the shape represents the overall spectral shape as shown in Figure 1, given a single pixel from the input hyperspectral patch. To exploit the overall shape of (target) center pixel's spectrum, we extract the overall shape information by acquiring the differences between its spectral bands. Before extracting the overall shape information, we apply Gaussian kernel function $G(\cdot)$ to the spectral vector of the center pixel $\boldsymbol{x}_c \in \mathbb{R}^{1\times 1\times C}$ in

Algorithm 1 Training of the Proposed Framework

```
Input HSI patch : $x$     Predictions : $\hat{y}$
Total # of training iterations : $N$
The set of the decomposed HSI patches, $D$ : $D = \{d_1, \cdots, d_{c''}\}$
for $n = 1, 2, \cdots, N$ do
    Random Spectral Extraction
    $D \leftarrow Decompose\ x$
    for each $d \in D$ do
        for each pixel (i, j) do
            $m_{i,j} \sim categorical(p_{1,(i,j)}, \cdots, p_{c',(i,j)};$
            $q_{1,(i,j)}, q_{2,(i,j)}, \cdots, q_{c',(i,j)})$
        end for
    end for
    $x_r \leftarrow m_1 \oplus m_2 \oplus \cdots \oplus m_{c''}$

    Spectral Shape Extraction
    $x_c \leftarrow Extract\ center\ pixel\ vector\ from\ x$
    $x_{sc} \leftarrow G(x_c)$
    $s_c^l \leftarrow x_{sc}^{l+1} - x_{sc}^l$

    After the extraction process
    $f_p \leftarrow FC(PatchFeatureExtractor(x_r))$
    $f_s \leftarrow FC(ShapeFeatureExtractor(s_c))$
    $\hat{y} \leftarrow FC(f_p \oplus f_s)$
    Minimize CrossEntropyLoss
end for
```

order to mitigate noise effects (including attack perturbations) via smoothing. After applying Gaussian function, we acquire the smoothed center pixel $x_{sc} \in \mathbb{R}^{1\times1\times C}$. Then, we extract the overall shape information by calculating the differences between each spectral elements of the smoothed center pixel. In other words, the l-th element of shape vector is obtained by $s_c^l = x_{sc}^{l+1} - x_{sc}^l$, where $l \in \{1, \cdots, C-1\}$ is the index of spectral bands. Finally, we acquire the spectral shape vector $s_c \in \mathbb{R}^{1\times1\times(C-1)}$. The spectral shape vector s_c can capture and represent the overall shape of the spectral bands. Then, s_c is fed into the shape feature extractor to acquire the spectral shape feature, f_s. Since the proposed method is to employ the overall shape of the spectral bands, which is rarely changed even with adversarial perturbations, it can improve the adversarial robustness. Also, the proposed spectral shape feature encoding can be applied into the other existing other HSI classification frameworks easily.

The overall training procedure is described in Algorithm 1. The testing follows same procedure except for mini-batch iterations and minimization of the loss function.

IV. EXPERIMENTS

A. EXPERIMENT SETUP

1) DATASETS

We conduct experiments to verify the proposed framework with two public HSI datasets: *Salinas* and *University of Pavia* datasets. Salinas dataset consists of 16 classes, and it has 512 × 217 resolutions with 204 spectral bands ranging from 400 *nm* to 2,500 *nm*. It is composed of 54,086 pixels used for classification without backgrounds. Among them, we use 15% of pixels for training, and the residues for testing. University of Pavia dataset (we call it Pavia dataset) consists of 9 categories, and it has 610×340 resolutions with 103 spectral bands ranging from 430 *nm* to 860 *nm*. Pavia dataset is composed of total 42,776 pixels except for backgrounds. We also use 15% of pixels for training, and the rest of them for testing. Following the validation protocol in [46], we separate training and testing sets of both datasets without any overlap region.

TABLE 1. 2D-CNN patch feature extractor architecture and following fully connected layers. The bracket represents the size for salinas dataset.

Operation	# of Filters	Filter Size $h \times w \times c$	Stride $h \times w$	Padding $h \times w$	Output Size $h \times w \times c$
Input HSI Patch	-	-	-	-	$11 \times 11 \times 103\,(204)$
Random Sampled HSI Patch	-	-	-	-	$11 \times 11 \times 20\,(40)$
CONV(2D)	32	$3 \times 3 \times 20$	1×1	-	$9 \times 9 \times 32$
CONV(2D)	64	$3 \times 3 \times 32$	1×1	-	$7 \times 7 \times 64$
CONV(2D)	64	$3 \times 3 \times 64$	1×1	1×1	$7 \times 7 \times 64$
CONV(2D)	128	$3 \times 3 \times 64$	1×1	-	$5 \times 5 \times 128$
CONV(2D)	128	$3 \times 3 \times 128$	1×1	1×1	$5 \times 5 \times 128$
CONV(2D)	128	$3 \times 3 \times 128$	1×1	1×1	$5 \times 5 \times 128$
CONV(2D)	256	$3 \times 3 \times 256$	1×1	-	$3 \times 3 \times 256$
FC	-	-	-	-	256
FC	-	-	-	-	256

TABLE 2. 3D-CNN Patch feature extractor architecture and following fully connected layers. The bracket represents the size for salinas dataset.

Operation	# of Filters	Filter Size $h \times w \times c \times d$	Stride $h \times w \times c$	Padding $h \times w \times c$	Output Size $h \times w \times c \times d$
Input HSI Patch	-	-	-	-	$11 \times 11 \times 103\,(204) \times 1$
Random Sampled HSI Patch	-	-	-	-	$11 \times 11 \times 20\,(40) \times 1$
CONV(3D)	16	$3 \times 3 \times 3 \times 1$	$1 \times 1 \times 1$	$0 \times 0 \times 1$	$9 \times 9 \times 20\,(40) \times 16$
CONV(3D)	32	$3 \times 3 \times 3 \times 16$	$1 \times 1 \times 2$	$0 \times 0 \times 1$	$7 \times 7 \times 10\,(20) \times 32$
CONV(3D)	32	$3 \times 3 \times 3 \times 32$	$1 \times 1 \times 2$	$1 \times 1 \times 1$	$7 \times 7 \times 5\,(10) \times 32$
CONV(3D)	64	$3 \times 3 \times 3 \times 32$	$1 \times 1 \times 2$	$0 \times 0 \times 1$	$5 \times 5 \times 3\,(5) \times 64$
CONV(3D)	64	$3 \times 3 \times 3 \times 64$	$1 \times 1 \times 2$	$1 \times 1 \times 1$	$5 \times 5 \times 2\,(3) \times 64$
CONV(3D)	128	$3 \times 3 \times 3 \times 64$	$1 \times 1 \times 2$	$0 \times 0 \times 1$	$3 \times 3 \times 1\,(2) \times 128$
FC	-	-	-	-	256
FC	-	-	-	-	256

2) NETWORK DETAILS

The proposed framework consists of the spectral shape and patch feature encoding paths. The shape feature extractor consists of 1D convolution layers, and there are two kinds of the patch feature extractor. Each version of the patch feature extractor is composed of 2D or 3D convolution layers. These kinds of architectures are used for the general HSI classification networks [42], [45]. We evaluate two versions of the patch feature extractor for the generalization of the proposed framework. Also, to verify the effectiveness of the proposed approach, we construct baseline models for comparisons. The baseline model has the same architecture of the patch feature extractor (2D-CNN or 3D-CNN). In other words, the baseline model takes full input patch without random spectral sampling, and encoding them through patch feature extractor.

For the network details, TABLE 1, 2, and 3 illustrate the network architectures of the patch feature extractor and shape feature extractor. The first two tables, TABLE 1 and 2, show

TABLE 3. The Shape feature extractor architecture and following fully connected layers. The bracket represents the size for salinas dataset.

Operation	# of Filters	Filter Size	Stride	Padding	Output Size
Input HSI Patch	-	-	-	-	$11 \times 11 \times 103\,(204)$
Shape Vector	-	-	-	-	102 (203)
CONV(1D)	16	11	2	5	$51\,(102) \times 15$
CONV(1D)	32	5	2	2	$26\,(51) \times 32$
CONV(1D)	64	2	2	1	$13\,(26) \times 64$
FC	-	-	-	-	256
FC	-	-	-	-	256

the 2D and 3D-CNN patch feature extractors, respectively. The 2D-CNN patch extractor is composed of 7 2D convolutional layers and, 3D version consists of 6 3D convolutional layers. In the tables, the output size is described mainly with respect to Pavia dataset, and the bracket represents the size for Salinas dataset. Also, TABLE 3 illustrates the network architecture of the shape feature extractor and following fully connected layers. It includes 3 convolutional layers, where each of them contains 1D convolution, batch normalization, and hyperbolic tangent non-linear functions. Following the tables, the sizes of shape feature, f_s, and patch feature, f_p, are the same as 256 dimension. Since both features are concatenated in the channel direction, the last fully connected layer for the final classification takes the feature of 512 dimension as an input.

3) IMPLEMENTATION DETAILS

We adopt Pytorch 1.2 [47] and CUDA 9.2 with a single GEFORCE GTX 1080Ti GPU for every implementation in this paper. We set spatial resolution of the input HSI patch as $H = 11$ and $W = 11$. The spectral dimensions, C, of Pavia and Salinas datasets, are 103 and 204, respectively. We set the size of gaussian kernel as 7 empirically [48]. For the random spectral sampling, we set $C' = 5$, where C'' becomes 20 and 40, while the unused spectral dimensions (remaining 3 and 4 spectral bands) for each dataset are not considered.

For both datasets, we train 2D-CNN frameworks with Adam optimizer with 0.001 learning rate, 0.00005 learning rate decay, and 300 epochs. Also, 3D-CNN frameworks are trained with SGD optimizer with 0.1 learning rate, 0.00005 learning rate decay, and 300 epochs.

4) ADVERSARIAL ROBUSTNESS EVALUATION PRINCIPLES

Calini *et al.* [49] suggested guidelines regarding how to evaluate adversarial robustness. They established four adversarial robustness evaluation principles: 1) applying a diverse set of adversarial attacks, 2) comparing with previous works, 3) performing black-box attacks using similar networks, and 4) attacking the randomness adaptively for those utilizing randomness effects. According to robustness evaluation principles, we demonstrate the effectiveness of the proposed approach as follows:

- Regarding adversarial attacks, we apply diverse adversarial attacks (FGSM [37], PGD [22], and CW [38]) to evaluate the adversarial robustness on different datasets. To verify the effectiveness of the proposed framework, we conduct the experiments with the various size of ϵ. (Evaluation Principle (1))
- We compare the adversarial robustness of our approach with other defense methods (*i.e.,* adversarial training (AT) [22], ADP [28], RLS [26]). For AT, we train the baseline model (2D-CNN and 3D-CNN) with PGD ($\epsilon = 0.01$) attacked examples. For the ADP, we construct three ensemble models. Each model has architecture of baseline models (2D-CNN and 3D-CNN). Then, they are optimized by the ADP loss function proposed in [28]. For the RLS, we construct ensemble model set with two ensemble models. Each model has architecture of baseline models (2D-CNN and 3D-CNN). Then, the ensemble model set is optimized by random layer sampling method proposed in [26]. (Evaluation Principle (2))
- To verify the robustness of the proposed method, we conduct experiment under various adversarial attack scenarios such black-box attack and adaptive attack scenario. Detailed attack settings are described in Section IV-C and IV-D (Evaluation Principle 3 and 4)

B. ADVERSARIAL ROBUSTNESS EVALUATION

TABLE 4 shows the classification accuracy under adversarial attack settings along with the two different datasets and two kinds of baseline networks (2D-CNN and 3D-CNN). In the case of Pavia dataset, as shown in the table, the baseline model tends to be vulnerable to adversarial attacks. With the clean image ('No Attack'), the accuracy of baseline is 92.38% and 90.92% on 2D-CNN and 3D-CNN, respectively. However, when the adversarial perturbation is added, the accuracy drops to 8.04%, 4.92%, and 7.84% under the FGSM, PGD, and CW attacks on 2D-CNN model. Also, in the case of 3D-CNN, the accuracy drop to 12.72%, 6.29%, and 28.79% under the FGSM, PGD, and CW attacks respectively.

In the case of our proposed method, the accuracy on clean image shows 87.82% and 87.96% on 2D-CNN and 3D-CNN respectively. When the adversarial perturbation is added, our proposed method shows better performance than the baseline method. Furthermore, we compare our approach with the other defense methods. Even though previous defense method shows better robustness than the baseline model, our proposed method shows superior performances against most of the attack scenarios. Especially, for the case of CW attack, the proposed framework secures the adversarial robustness with large margins from the other methods. Also, in the case of Salinas dataset, our results show superior robustness than the baseline model and other defense methods.

As mentioned before, the existing adversarial defense methods have the limitations when applying them into the HSI classification frameworks; large increase of training time and network parameters. For the proposed framework, we consider the spectral characteristic of HSIs to improve the adversarial robustness, alleviating the limitations as well. As described in TABLE 4, ADP [28] and RLS [26] require 2-3 times more network parameters, compared with the

TABLE 4. Classification accuracy (%) on adversarial examples of two datasets (Pavia and Salinas). We compare the adversarial robustness with the Baseline, Adversarial Training (AT), Adaptive Diversity Promoting (ADP), and Random Layer Sampling (RLS) methods. 2D- and 3D-CNN indicate the type convolution operation of the baseline and the patch feature extractor. Also, we compare the number of network parameters, consuming time for each iteration during the training, and the prediction time for the inference.

Dataset	Attack	ϵ	2D-CNN					3D-CNN				
			Baseline	AT	ADP	RLS	Ours	Baseline	AT	ADP	RLS	Ours
Pavia	No attack	-	92.38	87.38	91.41	**92.45**	87.82	90.92	86.15	91.56	**92.13**	87.96
	FGSM	0.01	38.70	79.28	41.19	83.09	**85.96**	50.62	81.00	28.80	85.14	**86.28**
		0.02	19.37	59.42	18.71	77.60	**82.71**	25.28	67.56	14.85	81.60	**82.57**
		0.04	9.31	34.84	9.00	60.70	**67.68**	15.87	44.96	6.30	66.79	**70.03**
		0.06	8.04	24.67	6.95	41.01	**49.95**	12.72	31.03	4.99	51.31	**58.79**
	PGD	0.01	22.15	73.33	21.13	71.58	**78.20**	34.62	77.05	18.72	74.64	**78.19**
		0.02	8.62	34.16	7.98	43.09	**49.03**	14.79	41.54	8.27	40.04	**56.75**
		0.04	6.64	13.12	5.61	**27.03**	20.22	7.39	13.67	4.22	34.31	**43.78**
		0.06	4.92	9.34	4.15	**22.27**	14.85	6.29	11.24	4.02	20.41	**39.07**
	CW	-	7.84	39.25	17.83	23.78	**50.22**	28.79	43.07	17.24	22.81	**62.24**
	# of parameters (M)		**1.34**	**1.34**	4.02	2.68	1.61	**0.90**	**0.90**	2.72	1.81	1.02
	Iteration time (s)		**0.020**	0.313	0.047	0.024	0.045	**0.014**	0.281	0.045	0.028	0.038
	Prediction time (s)		**0.0016**	**0.0016**	0.0028	0.0128	0.0018	**0.0018**	**0.0018**	0.0030	0.0142	0.0020
Salinas	No attack	-	90.28	82.47	90.98	**91.27**	85.80	92.00	87.42	92.11	**92.89**	86.80
	FGSM	0.005	40.17	71.98	45.44	72.10	**83.43**	51.16	79.77	39.08	81.48	**85.62**
		0.010	24.31	50.97	23.72	60.42	**78.73**	35.17	62.23	22.61	72.38	**83.13**
		0.015	19.05	37.95	13.83	49.97	**71.08**	28.81	42.89	13.01	62.33	**80.05**
		0.020	18.74	27.88	9.97	41.81	**61.56**	22.89	35.38	8.04	51.76	**75.73**
	PGD	0.005	23.38	66.73	28.11	68.10	**74.26**	44.44	78.71	31.94	70.48	**81.22**
		0.010	7.79	36.12	6.10	41.51	**47.96**	20.29	51.66	10.47	59.66	**66.32**
		0.015	4.44	24.02	5.40	28.86	**32.56**	10.35	31.82	8.04	44.54	**46.86**
		0.020	4.07	16.29	4.92	24.00	**24.20**	6.88	18.24	7.29	34.84	**35.01**
	CW	-	7.76	27.36	9.12	27.17	**68.38**	7.06	17.62	19.69	37.14	**75.26**
	# of parameters (M)		**1.37**	**1.37**	4.12	2.75	1.85	**0.98**	**0.98**	2.93	1.95	1.54
	Iteration time (s)		**0.025**	0.575	0.054	**0.025**	0.060	**0.022**	0.568	0.049	0.030	0.048
	Prediction time (s)		**0.0018**	**0.0018**	0.0029	0.0130	0.0020	**0.0019**	**0.0019**	0.0031	0.0149	0.0021

baseline model. On the contrary, the proposed approach needs only 1.2-1.5 times increase of network parameters. Also, the table shows the consuming time per iteration with the same batch size during training. Especially, AT [22] demands 15-25 times more network training time, while our framework requires only 2-3 times longer training time. RLS [26] seemingly increases a little training time, but it needs more training iterations following training procedure in [26], because they optimize more than two models. Also, in TABLE 4, the prediction time is the time consumed to estimate a single sample during the inference phase. As shown in the table, the proposed frameworks requires a small amount of time increase, while achieving the adversarial robustness via the random spectral sampling and spectral shape feature encoding. Note that the process of generating adversarial examples are not considered in measuring the prediction time of AT [28]. The experiment corroborates that the proposed approach improves the adversarial robustness, and reduces the increase in training time and network parameters, by employing the spectral characteristic of HSIs.

C. ROBUSTNESS EVALUATION UNDER BLACK-BOX ATTACK SETTING

The black-box attacks mean that adversaries do not know the whole information of target networks. Therefore, the black-box attacks usually generate the adversarial examples from the other similar networks. Such black-box attack scenarios are necessary to prove the adversarial robustness against general and more realistic attacks, since getting full information of real-world application is hardly possible. TABLE 5 and 6 describe the experiment results to show adversarial robustness under black-box attack settings. In TABLE 5, we generate adversarial examples from 3D-CNN baseline model, and they are fed into the 2D-CNN baseline model and proposed framework for comparison. As illustrated in TABLE 5, our framework shows

TABLE 5. Classification accuracy (%) under black-box attack scenarios. The adversarial examples generated from 3D-CNN baseline model. Then, we evaluate on 2D-CNN baseline model and Ours (2D-CNN).

Attack	Pavia			Salinas		
	ϵ	Baseline	Ours	ϵ	Baseline	Ours
FGSM	0.01	70.69	**83.67**	0.005	72.163	**77.21**
	0.02	52.85	**73.38**	0.010	50.494	**56.64**
	0.04	31.92	**50.20**	0.015	41.12	**44.21**
	0.06	23.51	**34.33**	0.020	30.18	**39.86**
PGD	0.01	68.02	**83.10**	0.005	67.00	**77.45**
	0.02	47.05	**72.16**	0.010	42.64	**55.64**
	0.04	27.65	**41.97**	0.015	37.79	**43.87**
	0.06	21.39	**29.66**	0.020	32.55	**39.53**

TABLE 6. Classification accuracy (%) under black-box attack scenario. The adversarial examples generated from 2D-CNN baseline model. Then, we evaluate on 3D-CNN baseline model and ours (3D-CNN).

Attack	Pavia			Salinas		
	ϵ	Baseline	Ours	ϵ	Baseline	Ours
FGSM	0.01	76.43	**79.19**	0.005	74.84	**79.63**
	0.02	48.44	**62.64**	0.010	53.49	**71.06**
	0.04	24.60	**38.08**	0.015	46.11	**59.66**
	0.06	17.57	**25.23**	0.020	41.12	**48.79**
PGD	0.01	75.88	**78.66**	0.005	76.33	**80.30**
	0.02	51.61	**62.82**	0.010	57.80	**72.39**
	0.04	28.88	**40.72**	0.015	48.83	**60.84**
	0.06	20.99	**27.82**	0.020	42.54	**50.06**

better adversarial robustness than the baseline model under black-box attack scenarios. In TABLE 6, we generate adversarial examples from 2D-CNN baseline model, and they are fed into the 2D-CNN baseline model and proposed framework for comparison. As shown in the table, our framework also shows better robustness than the baseline model. From the experiment, the proposed approach is also robust to more general and realistic black-box attacks.

D. ADVERSARIAL ROBUSTNESS UNDER ADAPTIVE ATTACK

1) EXPECTATION OVER TRANSFORMATION (EOT)

Athalye *et al.* [50] proposed Expectation Over Transformation (EOT) to consider expectation of many possible transformations of input data for generating adversarial examples. Since our framework involves the spectral randomness which transforms the input of patch feature extractor, it is required to evaluate the adversarial robustness using EOT attack. To this end, we conduct 10 forward iterations (different random sampled HSI patch for each iteration), and average them to acquire final classification results during adversarial example generation, considering many possible input transformations. TABLE 7 illustrates the robustness performance with EOT attack. In the experiment, the proposed framework is still robust under EOT attack scenarios. Since our method exploits the overall shape of spectral robust to adversarial attack, we still maintain the robustness under EOT attack. Figure 4 shows more detail experiment results according to the number of iteration for EOT attack. As shown in the figure, although the performance is degraded with increasing number of iterations, the extent of the performance drop is saturated. Therefore, the figure corroborates that the proposed framework still shows the adversarial robustness under the EOT attack scenarios.

2) RANDOMNESS EXPOSURE SCENARIO

The adversarial defense methods which take randomness effects would be vulnerable, when the randomness is fully exposed to the adversaries. For example, if the adversaries are aware of which spectral bands are sampled during adversarial example generation, it could be crucial to our framework. To verify the adversarial robustness with this scenario, we predetermine which spectral bands are sampled for testing, and generate adversarial examples with those predetermined spectral bands. TABLE 8 shows the classification accuracy evaluated on the adversarial examples generated by using the predetermined spectral band sampling. Although it is exposed to one of the worst cases, our approach still shows much better adversarial robustness than the baseline. It can be interpreted that encoding the spectral shape feature can be helpful to improve adversarial robustness. To verify this, in the following section, we verify the effects of the spectral shape feature encoding.

TABLE 7. Classification accuracy (%) evaluated with adversarial examples from Expectation Over Transform (EOT) attack scenarios.

Dataset	Attack	ϵ	2D-CNN		3D-CNN	
			Baseline	Ours	Baseline	Ours
Pavia	FGSM	0.01	38.70	**70.45**	50.61	**70.53**
		0.02	19.36	**41.48**	25.28	**55.37**
	PGD	0.01	22.15	**57.27**	34.62	**61.64**
		0.02	8.62	**21.14**	14.78	**44.42**
	CW	-	7.83	**27.31**	28.79	**43.20**
Salinas	FGSM	0.005	40.16	**64.30**	51.16	**78.09**
		0.010	24.31	**38.09**	35.16	**57.81**
	PGD	0.005	27.37	**52.38**	44.44	**73.91**
		0.010	7.78	**26.44**	20.28	**44.11**
	CW	-	7.76	**39.66**	7.06	**63.22**

TABLE 8. Classification accuracy (%) when sampled spectral bands are exposed to adversaries.

Dataset	Attack	ϵ	2D-CNN		3D-CNN	
			Baseline	Ours	Baseline	Ours
Pavia	FGSM	0.01	38.70	**66.53**	50.61	**65.97**
		0.02	19.36	**37.16**	25.28	**52.94**
	PGD	0.01	22.15	**55.32**	34.62	**60.30**
		0.02	8.62	**20.60**	14.78	**42.88**
Salinas	FGSM	0.005	40.16	**58.06**	51.16	**73.24**
		0.010	24.31	**35.30**	35.16	**47.77**
	PGD	0.005	27.37	**48.40**	44.44	**70.53**
		0.010	7.78	**26.03**	20.28	**39.51**

E. ABLATION STUDIES

In the ablation study section, we verify the effects of the spectral shape feature encoding and random spectral extraction. To verify that, we conduct three experiment settings. 1) Only using spectral shape feature encoding without random spectral extraction, 2) encoding the spectral value itself instead of shape of spectral, and 3) only using random spectral extraction. Followings are the experiment results.

1) ONLY USING SPECTRAL SHAPE FEATURE ENCODER

We are motivated from that the adversarial attacks using small perturbations do not deform the overall shape of spectrum largely. Therefore, we extract the overall shape information of the (target) center pixel's spectrum, and acquire the spectral shape feature following the section III-D. To corroborate the effects of the spectral shape feature encoding, we construct

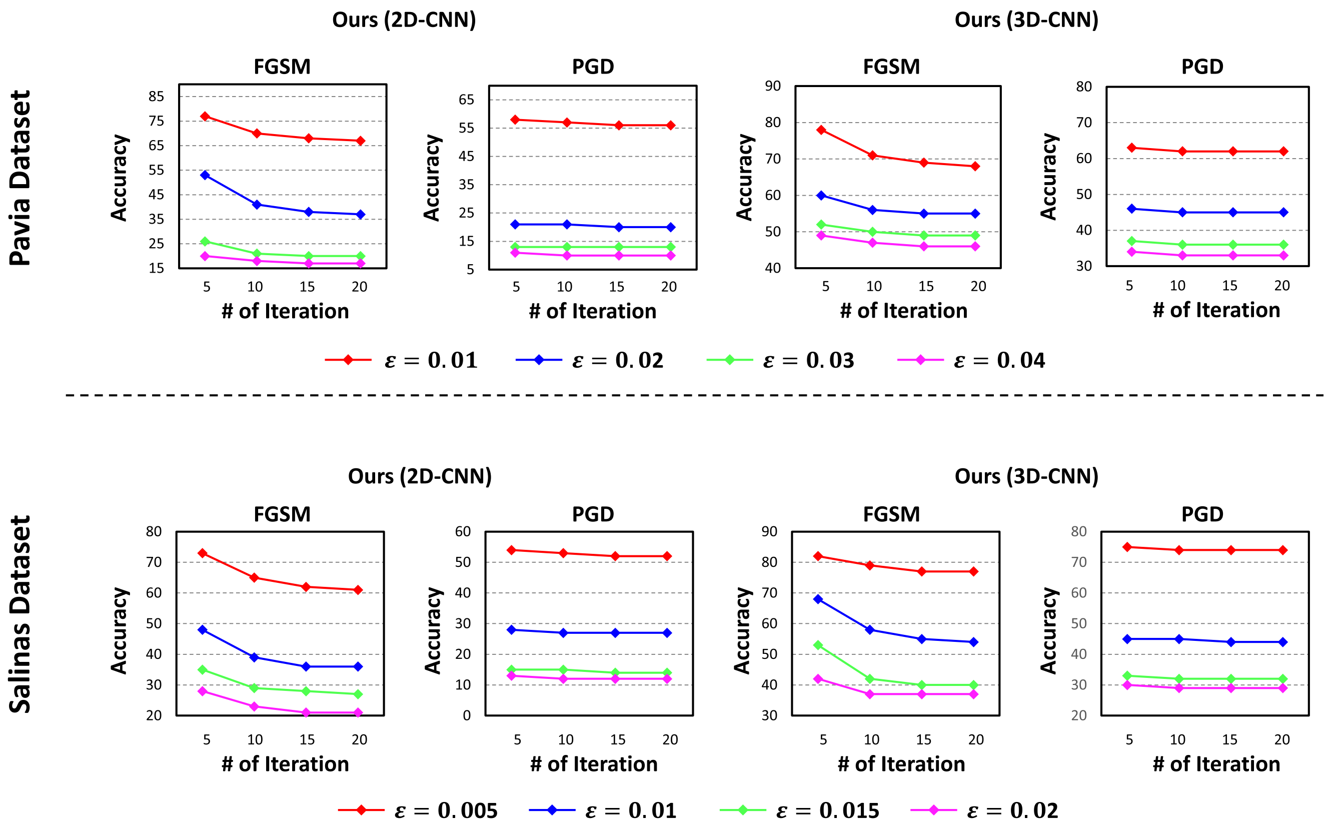


FIGURE 4. Classification accuracy (%) under EOT FGSM and PGD attacks. Although the performance is degraded with increasing number of iterations, the extent of the performance drop is saturated.

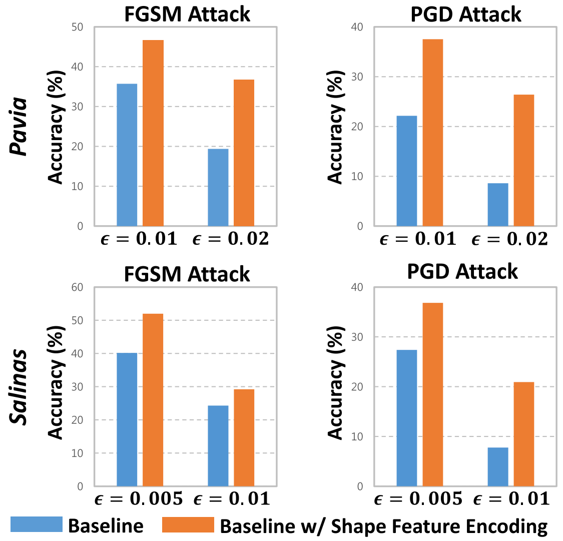


FIGURE 5. Effectiveness of the proposed spectral shape feature encoding. Even though the random spectral sampling is not applied, with the spectral shape feature encoding improves the adversarial robustness of the 2D-CNN baseline model.

TABLE 9. Robustness comparison along with the input types: the spectral value (x_{sc}) and the shape of spectral (s_c).

Dataset	Attack	ϵ	2D-CNN		3D-CNN	
			x_{sc}	s_c	x_{sc}	s_c
Pavia	FGSM	0.01	85.49	**85.96**	83.83	**86.28**
		0.02	79.73	**82.71**	74.78	**82.57**
		0.04	58.90	**67.68**	51.85	**70.03**
		0.06	43.86	**49.95**	34.95	**58.79**
	PGD	0.01	73.86	**78.20**	68.13	**78.19**
		0.02	40.90	**49.03**	38.52	**56.75**
		0.04	20.04	**20.22**	21.02	**43.78**
		0.06	14.71	**14.85**	18.22	**39.07**
Salinas	FGSM	0.005	82.45	**83.43**	85.38	**85.62**
		0.010	75.44	**78.73**	81.36	**83.13**
		0.015	68.39	**71.08**	75.80	**80.05**
		0.020	60.69	**61.56**	67.49	**75.73**
	PGD	0.005	72.33	**74.26**	78.45	**81.22**
		0.010	**52.25**	47.96	56.55	**66.32**
		0.015	**34.71**	32.56	38.02	**46.86**
		0.020	**27.32**	24.20	31.07	**35.01**

the networks without the random spectral extraction (sampling). In other words, the patch feature extractor takes the original input HSI patch that is not processed by the random spectral sampling. In Figure 5, the blue bars represent the classification accuracy (%) acquired from the 2D-CNN baseline model. The orange bars represent the accuracy from the network trained by applying the spectral shape feature encoding on the baseline model. As shown in the figure, it shows better adversarial robustness against FGSM and PGD attacks. From the experiment, we verify that leveraging the overall shape of the spectrum could be effective to mitigate the adversarial vulnerability.

TABLE 10. Ablation studies for verifying the effectiveness of random spectral extraction and spectral shape encoding.

Backbone	Random Spectral Extraction	Spectral Shape Encoding	Attacks					
			FGSM		PGD		Adaptive Attacks	
			$\epsilon = 0.01$	$\epsilon = 0.02$	$\epsilon = 0.01$	$\epsilon = 0.02$	$\epsilon = 0.01$	$\epsilon = 0.02$
2D-CNN	✘	✘	38.70	19.37	22.15	8.62	-	-
	✔	✘	**86.65**	**83.59**	77.3	47.89	42.73	16.92
	✔	✔	85.96	82.71	**78.2**	**49.03**	**55.32**	**20.67**
3D-CNN	✘	✘	50.62	25.28	34.62	14.79	-	-
	✔	✘	85.04	80.53	74.57	51.78	55.11	34.42
	✔	✔	**86.28**	**82.57**	**78.19**	**56.75**	**60.38**	**42.88**

2) ENCODING SPECTRAL VALUE ITSELF

Since adversarial attacks craft adversarial examples with small and indistinguishable noise, the overall (*increasing/decreasing*) shape information of spectral bands is not deformed largely even with adversarial attacks. Therefore, to leverage the overall shape of the target pixel's spectrum, we extract the shape vector, $\boldsymbol{s}_c$, rather than encoding the value itself of center pixel's spectral vector, $\boldsymbol{x}_c$ or smoothed one, $\boldsymbol{x}_{sc}$. Accordingly, we conduct the comparison experiments, taking $\boldsymbol{x}_{sc}$ or $\boldsymbol{s}_c$ as an input of the shape feature extractor. TABLE 9 describes the experimental results. As illustrated in the table, taking $\boldsymbol{s}_c$ as an input shows better classification accuracy performances under almost attack scenarios. The table demonstrates that taking the shape guided input would be more helpful to encode the overall shape information.

3) ONLY USING RANDOM SPECTRAL EXTRACTION

One of the main reason that our proposed method is robust against adversarial attack is the random spectral extraction. Through the random spectral extraction, we could hide the spectral information used for inference. In other words, it makes adversaries not being aware of which spectral bands to be used during inference time. Therefore, only using the random spectral extraction could improve the adversarial robustness. TABLE 10 shows the ablation study results on Pavia dataset. As shown in the table, without random spectral extraction and spectral shape encoding (baseline in TABLE 4), the accuracy drops significantly. However, when applying the random spectral extraction, the accuracy increases dramatically. When applying spectral shape encoding, the robustness further improves under FGSM and PGD attacks. Especially, under adaptive attack scenarios (same as Section IV-D2, PGD attack), applying spectral shape encoding makes model more robust. It could be interpreted that the spectral shape encoding plays a key role when randomness is exposed.

F. QUALITATIVE RESULTS

In this section, we visualize how adversarial attacks affect the classification performances qualitatively, to corroborate the effectiveness of the proposed method. The experiment is conducted on Pavia dataset, and Figure 6 shows the HSI map estimated by the 2D-CNN baseline and our framework. In the figure, the upper row represents the estimated HSI map of 2D-CNN baseline, and the bottom row contains the estimated HSI map, when the proposed method is applied. The ground truth HSI map is shown in the rightmost. As described in the estimated HSI maps of the baseline network (*top row*), adversarial attacks could lead the network to conduct misclassification. However, when the proposed method is applied to the network (*bottom row*), the less changes observed in the HSI maps, compared to corresponding HSI maps of baseline model. It means that the prediction results are not easy to be changed by the adversarial attacks. From the qualitative visualization results, it could demonstrate the effectiveness of the proposed method, achieving the robustness against adversarial attacks.

V. DISCUSSION

A. MEAN AND STANDARD DEVIATION OF CLASSIFICATION ACCURACY

Since the proposed random spectral sampling selects the different (spectral) channel combinations at every prediction, the classification accuracy could be fluctuated for each prediction time. However, the fluctuation would be insignificant, because the proposed framework is also trained with various channel combination during the training time. To verify it, we conduct the experiment repeating 10 evaluation times and obtain the mean and standard deviation of classification accuracy. TABLE 11 shows the experimental results (*i.e.,* mean and standard deviation) under various attacks including no attack scenario. As shown in TABLE 11, while achieving the adversarial robustness against various attacks, the standard deviation of classification accuracy is small enough to be ignored, ranging from 0.008 to 0.080. Although the proposed random spectral sampling could lead to the accuracy fluctuation by different channel selection, its impact is insignificant to the overall classification accuracy.

B. GENERALIZABILITY

In this section, we discuss the generalizability of the proposed method to the more recent hyperspectral classification frameworks [51]–[54]. Since SSRN and SSAN [51], [52] are based on 3D-CNN with residual connections and attention modules, we expect that the proposed method could be applied to SSRN and SSAN. Furthermore, RNN-based models have been introduced recently [53], [54]. In general, RNN-based models encode each spectral value of HSI pixel vector as an

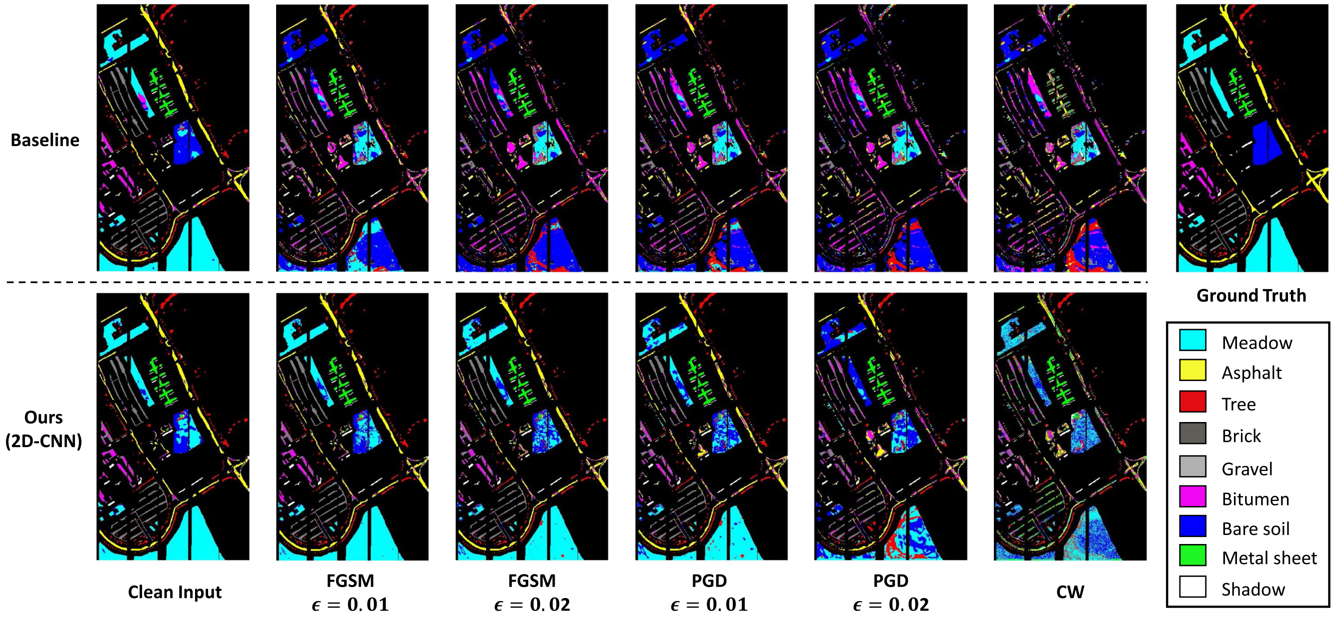


FIGURE 6. The visualization results showing the estimated HSI map from the 2D-CNN baseline (*top row*) and proposed framework (*bottom row*). With the clean input, both models could obtain the estimated HSI maps, which are similar with the ground truth HSI map. However, with the input perturbed by adversarial attacks, the estimated HSI maps of the baseline model are deteriorated, while the proposed framework shows the adversarial robustness (that is, less changes are observed in estimated HSI maps).

TABLE 11. The mean and standard deviation (std) of classification accuracy (%) obtained by 10 evaluation tests. The standard deviation ranges from 0.008 to 0.080, which is insignificant, compared to the overall classification accuracy.

Dataset	Attack	ϵ	2D-CNN		3D-CNN	
			Mean	Std	Mean	Std
Pavia	No Attack	-	87.83	0.019	87.96	0.027
	FGSM	0.01	86.03	0.046	86.29	0.028
		0.02	82.68	0.044	82.55	0.033
		0.04	67.67	0.046	69.99	0.041
		0.06	49.91	0.075	58.85	0.055
	PGD	0.01	78.21	0.039	78.15	0.040
		0.02	49.07	0.044	56.77	0.027
		0.04	20.22	0.008	43.78	0.014
		0.06	14.74	0.012	39.07	0.015
	CW	-	50.16	0.074	62.17	0.079
Salinas	No Attack	-	85.85	0.060	86.77	0.036
	FGSM	0.005	83.44	0.056	85.53	0.060
		0.010	78.75	0.062	83.14	0.046
		0.015	71.06	0.054	80.00	0.065
		0.020	61.50	0.080	75.69	0.056
	PGD	0.005	74.31	0.067	81.22	0.043
		0.010	47.96	0.054	66.28	0.041
		0.015	32.61	0.066	46.90	0.048
		0.020	24.21	0.037	35.00	0.030
	CW	-	68.43	0.051	75.31	0.070

TABLE 12. The classification accuracy (%) under various adversarial attacks with RNN-based network using GRU. Pavia dataset is used for this experiment. As shown in the table, the RNN baseline becomes adversarially robust with the proposed method.

Attack	ϵ	RNN baseline	RNN baseline + Proposed Method
No Attack	-	**92.18**	88.78
FGSM	0.01	77.62	**84.88**
	0.02	48.35	**84.04**
	0.04	18.51	**82.42**
	0.06	13.04	**80.02**
PGD	0.01	77.41	**81.72**
	0.02	47.53	**75.69**
	0.04	17.15	**58.50**
	0.06	12.31	**43.22**

input for each step. Since RNN-based models have different network architectures with 2D and 3D CNNs, we construct RNN-based model using gated recurrent unit (GRU) following [53] and apply the proposed method. With the RNN baseline, TABLE 12 shows the experimental results regarding the robustness against adversarial attacks (*i.e.,* FGSM and PGD) on University of Pavia dataset. As described in the table, the RNN baseline seems vulnerable to adversarial attacks. However, the proposed method could improve the robustness against adversarial attacks, verifying that the proposed method could be performed with RNN-based model.

C. CONSISTENT NUMBER OF CONVOLUTION LAYERS AND FILTERS

In this section, we conduct the experiment using the 2D-CNN which is modified to have same number of convolution layers and filters with 3D-CNN model (please refer to TABLE 1 and 2). For 2D-CNN shown in TABLE 1, we remove the 6-th convolution layers and adjust the number of filters for each

TABLE 13. The classification accuracy (%) under various adversarial attacks on pavia dataset. The modified 2D-CNN is adopted, which has same number of convolution layers and filters.

Attack	ϵ	Modified 2D-CNN	Ours with Modified 2D-CNN
No Attack	-	**91.46**	86.89
FGSM	0.01	38.93	**85.79**
	0.02	18.73	**81.91**
	0.04	8.88	**69.51**
	0.06	6.53	**55.79**
PGD	0.01	25.50	**76.65**
	0.02	8.59	**50.09**
	0.04	5.51	**20.48**
	0.06	4.78	**14.63**
CW	-	9.15	**47.49**

convolution operator as same as 3D-CNN model. TABLE 13 shows the experimental results using the modified 2D-CNN on Pavia dataset. In this experiment, the proposed framework achieves the adversarial robustness against the various attacks.

VI. CONCLUSION

In this paper, we present a robust HSI classification framework, which shows the improved adversarial robustness under various adversarial attack scenarios. The proposed framework benefits from the unique spectral characteristic of HSIs by the spectral random sampling and the spectral shape feature encoding. The spectral random sampling selects spectral bands randomly in every testing time. Therefore, the proposed method make adversaries could not aware of which spectral bands to be used during each inference time. Also, we extract and encode the overall spectral shape of spectrum to acquire the shape feature, which is robust to adversarial attacks. To the best of our knowledge, it is the first work dealing with the adversarial attack scenarios, mitigating the adversarial vulnerability of the HSI classification framework, and alleviating the problems caused by applying the existing defense methods. Through comprehensive experiments, we demonstrate the effectiveness of the proposed approach, which can be applied into the other existing HSI classification frameworks.

ACKNOWLEDGMENT

(Sungjune Park and Hong Joo Lee contributed equally to this work.)

REFERENCES

[1] N. Akhtar and A. Mian, "Hyperspectral recovery from RGB images using Gaussian processes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 100–113, Jan. 2020.

[2] J. Li, X. Zhao, Y. Li, Q. Du, B. Xi, and J. Hu, "Classification of hyperspectral imagery using a new fully convolutional neural network," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 292–296, Feb. 2018.

[3] L. Zhang, W. Wei, Y. Zhang, C. Tian, and F. Li, "Reweighted laplace prior based hyperspectral compressive sensing for unknown sparsity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2274–2281.

[4] W. Li, Q. Du, and B. Zhang, "Combined sparse and collaborative representation for hyperspectral target detection," *Pattern Recognit.*, vol. 48, no. 12, pp. 3904–3916, Dec. 2015.

[5] L. He, J. Li, C. Liu, and S. Li, "Recent advances on spectral–spatial hyperspectral image classification: An overview and new guidelines," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1579–1597, Mar. 2018.

[6] J. Chi and M. M. Crawford, "Spectral unmixing-based crop residue estimation using hyperspectral remote sensing data: A case study at Purdue university," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2531–2539, Jun. 2014.

[7] Y. Yuan, Q. Wang, and G. Zhu, "Fast hyperspectral anomaly detection via high-order 2-D crossing filter," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 2, pp. 620–630, Feb. 2015.

[8] X. Wei, W. Li, M. Zhang, and Q. Li, "Medical hyperspectral image classification based on end-to-end fusion deep neural network," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 11, pp. 4481–4492, Nov. 2019.

[9] S. Mei, J. Ji, Y. Geng, Z. Zhang, X. Li, and Q. Du, "Unsupervised spatial–spectral feature learning by 3D convolutional autoencoder for hyperspectral classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6808–6820, Sep. 2019.

[10] H. Zhang, Y. Li, Y. Jiang, P. Wang, Q. Shen, and C. Shen, "Hyperspectral classification based on lightweight 3-D-CNN with transfer learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5813–5828, Aug. 2019.

[11] J. Yang, Y.-Q. Zhao, and J. C.-W. Chan, "Learning and transferring deep joint spectral–spatial features for hyperspectral classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4729–4742, Aug. 2017.

[12] H. Lee and H. Kwon, "Contextual deep CNN based hyperspectral classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 3322–3325.

[13] 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, p. 1068, Jul. 2018.

[14] 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.

[15] C. Mu, J. Liu, Y. Liu, and Y. Liu, "Hyperspectral image classification based on active learning and spectral–spatial feature fusion using spatial coordinates," *IEEE Access*, vol. 8, pp. 6768–6781, 2020.

[16] X. Ji, Y. Cui, H. Wang, L. Teng, L. Wang, and L. Wang, "Semisupervised hyperspectral image classification using spatial-spectral information and landscape features," *IEEE Access*, vol. 7, pp. 146675–146692, 2019.

[17] J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and J. Li, "Visual attention-driven hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 8065–8080, Oct. 2019.

[18] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

[19] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, early access, Aug. 18, 2020, 10.1109/TGRS.2020.3015157.

[20] J. Yao, D. Meng, Q. Zhao, W. Cao, and Z. Xu, "Nonconvex-sparsity and nonlocal-smoothness-based blind hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2991–3006, Jun. 2019.

[21] A. Fawzi, H. Fawzi, and O. Fawzi, "Adversarial vulnerability for any classifier," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1178–1187.

[22] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learn. Represent.*, 2018. [Online]. Available: https://openreview.net/forum?id=rJzIBfZAb

[23] J.-C. Burnel, K. Fatras, and N. Courty, "Generating natural adversarial hyperspectral examples with a modified Wasserstein GAN," 2020, *arXiv:2001.09993*. [Online]. Available: http://arxiv.org/abs/2001.09993

[24] 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.

[25] Q.-Z. Cai, C. Liu, and D. Song, "Curriculum adversarial training," in *Proc. 27th Int. Joint Conf. Artif. Intell. (IJCAI)*, Jul. 2018, pp. 3740–3747, doi: 10.24963/ijcai.2018/520.

[26] H. Lee, H. J. Lee, S. T. Kim, and Y. M. Ro, "Robust ensemble model training via random layer sampling against adversarial attack," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2020.

[27] J. Yuan and Z. He, "Adversarial dual network learning with randomized image transform for restoring attacked images," *IEEE Access*, vol. 8, pp. 22617–22624, 2020.

[28] T. Pang, K. Xu, C. Du, N. Chen, and J. Zhu, "Improving adversarial robustness via promoting ensemble diversity," in *Proc. Mach. Learn. Res.*, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. Long Beach, CA, USA: PMLR, Jun. 2019, pp. 4970–4979.

[29] 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.

[30] A. Rahiche and M. Cheriet, "Forgery detection in hyperspectral document images using graph orthogonal nonnegative matrix factorization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 662–663.

[31] R. Qureshi, M. Uzair, K. Khurshid, and H. Yan, "Hyperspectral document image processing: Applications, challenges and future prospects," *Pattern Recognit.*, vol. 90, pp. 12–22, Jun. 2019.

[32] S. Ortega, H. Fabelo, R. Camacho, M. de la Luz Plaza, G. M. Callicó, and R. Sarmiento, "Detecting brain tumor in pathological slides using hyperspectral imaging," *Biomed. Opt. Exp.*, vol. 9, no. 2, pp. 818–831, Feb. 2018.

[33] R. Pike, G. Lu, D. Wang, Z. G. Chen, and B. Fei, "A minimum spanning forest-based method for noninvasive cancer detection with hyperspectral imaging," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 3, pp. 653–663, Mar. 2016.

[34] H. Fabelo *et al.*, "*In-vivo* hyperspectral human brain image database for brain cancer detection," *IEEE Access*, vol. 7, pp. 39098–39116, 2019.

[35] P. W. T. Yuen and G. Bishop, "Hyperspectral multiple approach fusion for the long-range detection of low observable objects: MUF2," *Proc. SPIE*, vol. 6396, Oct. 2006, Art. no. 63960C.

[36] G. Suganthi and K. Reeba, "Discrimination of mine-like objects in infrared images using artificial neural network," *Indian J. Appl. Res.*, vol. 4, no. 12, pp. 206–208, 2014.

[37] 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

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

[39] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[40] X. Liu, M. Cheng, H. Zhang, and C.-J. Hsieh, "Towards robust neural networks via random self-ensemble," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 369–385.

[41] G. S. Dhillon, K. Azizzadenesheli, J. D. Bernstein, J. Kossaifi, A. Khanna, Z. C. Lipton, and A. Anandkumar, "Stochastic activation pruning for robust adversarial defense," in *Proc. Int. Conf. Learn. Represent.*, 2018. [Online]. Available: https://openreview.net/forum?id=H1uR4GZRZ

[42] 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.

[43] 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.

[44] 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.

[45] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019.

[46] J. Nalepa, M. Myller, and M. Kawulok, "Validating hyperspectral image segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1264–1268, Aug. 2019.

[47] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *Proc. NIPS Autodiff Workshop, Future Gradient-Based Mach. Learn. Softw. Techn.*, Long Beach, CA, USA, Dec. 2017.

[48] G. Bilgin, S. Erturk, and T. Yildirim, "Unsupervised classification of hyperspectral-image data using fuzzy approaches that spatially exploit membership relations," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 4, pp. 673–677, Oct. 2008.

[49] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin, "On evaluating adversarial robustness," 2019, *arXiv:1902.06705*. [Online]. Available: http://arxiv.org/abs/1902.06705

[50] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *Proc. Int. Conf. Mach. Learn.*, J. Dy and A. Krause, Eds., vol. 80. Stockholm, Sweden: PMLR, Jul. 2018, pp. 284–293. [Online]. Available: http://proceedings.mlr.press/v80/athalye18b.html

[51] 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.

[52] 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.

[53] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "Scalable recurrent neural network for hyperspectral image classification," *J. Supercomput.*, vol. 76, no. 11, pp. 8866–8882, Nov. 2020.

[54] 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.

SUNGJUNE PARK received the B.S. degree from the School of Electrical and Electronic Engineering, Yonsei University, Seoul, South Korea, in 2019. He is currently pursuing the Ph.D. degree in electrical engineering with the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea. His research interests include deep learning, machine learning, adversarial robustness, and object detection.

HONG JOO LEE received the B.S. degree from Ajou University, Suwon, South Korea, in 2016, and the M.S. degree from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2018, where he is currently pursuing the Ph.D. degree. His research interests include deep learning, machine learning, medical image segmentation, and adversarial robustness.

YONG MAN RO (Senior Member, IEEE) received the B.S. degree from Yonsei University, Seoul, South Korea, and the M.S. and Ph.D. degrees from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea. He was a Researcher at Columbia University, a Visiting Researcher at the University of California at Irvine, Irvine, CA, USA, and a Research Fellow of the University of California at Berkeley, Berkeley, CA, USA. He was a Visiting Professor with the Department of Electrical and Computer Engineering, University of Toronto, Canada. He is currently a Professor with the Department of Electrical Engineering and the Director of the Center for Applied Research in Artificial Intelligence (CARAI), KAIST. Among the years, he has been conducting research in a wide spectrum of image and video systems research topics. Among those topics, his interests include image processing, computer vision, visual recognition, multimodal learning, video representation/compression, and object detection. He received the Young Investigator Finalist Award of ISMRM, in 1992, and the Year's Scientist Award (Korea), in 2003. He served as an Associate Editor for IEEE Signal Processing Letters. He currently serves as an Associate Editor for IEEE Transactions on Circuits and Systems for Video Technology. He served as a TPC in many international conferences, including the program chair and organized special sessions.