IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 23, 2019, accepted July 3, 2019, date of publication July 8, 2019, date of current version July 31, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2927376

Spear and Shield: Attack and Detection for CNN-Based High Spatial Resolution Remote Sensing Images Identification

WENMEI LI[1,2,3], (Member, IEEE), ZHUANGZHUANG LI[2], JINLONG SUN[2], (Member, IEEE), YU WANG[2], (Student Member, IEEE), HAIYAN LIU[2], JIE YANG[2], (Member, IEEE), AND GUAN GUI[2], (Senior Member, IEEE)

[1]School of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
[2]College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
[3]Smart Health Big Data Analysis and Location Services Engineering Laboratory of Jiangsu Province, Nanjing 210023, China

Corresponding authors: Jinlong Sun (sunjinlong@njupt.edu.cn) and Guan Gui (guiguan@njupt.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61701258 and Grant 41401480, in part by the Jiangsu Specially Appointed Professor Program under Grant RK002STP16001, in part by the Summit of the Six Top Talents Program of Jiangsu under Grant XYDXX-010, in part by the Program for High-Level Entrepreneurial and Innovative Talents Introduction under Grant CZ0010617002, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20170906 and Grant BK20180765, in part by the Natural Science Foundation of Jiangsu Higher Education Institutions under Grant 17KJB510044, in part by the Nanjing University of Posts and Telecommunications Science Foundation (NUPTSF) under Grant 218085, in part by the Nanjing Technology Innovation Foundation for Selected Overseas Scientists under Grant 20180075, and in part by the 1311 Talent Plan, Nanjing University of Posts and Telecommunications.

ABSTRACT High spatial resolution remote sensing (HSRRS) images classification and identification is an important technology to acquire land surface information for land resource management, geographical situation monitoring, and global climate change. As the hottest deep learning method, convolutional neural network (CNN) has been successfully applied in HSRRS image classification and identification due to its powerful information extraction capability. However, adversarial perturbations caused by radiation transfer process or artificial or other unpredictable disturbances often deteriorate the stability of CNN. Under this background, we propose a robust architecture for adversarial attack and detection to classify and identify HSRRS images. First of all, two white-box attacks [i.e., large Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) and fast gradient sign method (FGSM)] are adopted respectively to generate adversarial images to confuse the model, and to assess the robustness of the HSRRS image classifier. Second, adversarial detection models based on support vector machine (SVM) with single or fused two level features are proposed to improve the detection accuracy. The features extracted from the testing CNN full connected layers contain adversarial perturbations and real information, from which SVM classifier and discriminate the real and the adversarial images. The adversarial attack model is evaluated in terms of overall accuracy (*OA*) and kappa coefficient (*kc*). The simulation results show that the *OA* decreases from 96.4% to 44.4% and 33.3% for L-BFGS and FGSM attacked classifier model, respectively. The adversarial detection is evaluated via *OA*, detection probability P_D, false alarm probability P_{FA}, and miss probability P_M. The simulation results indicate that the fused model with two different level features based on SVM can obtain the best *OA* (94.5%), P_D (0.933), P_{FA} (0.040), and P_M (0.067) among the detectors if the classifier is attacked by the FGSM. Meanwhile, when facing the L-BFGS attack, the fused model presents similar performance if the best single level features are utilized.

INDEX TERMS Convolutional neural network, attack detection, white-box attack, fast gradient sign method (FGSM), large Broyden-Fletcher-Goldfarb-Shanno (L-BFGS).

I. INTRODUCTION

Recently, many efforts on high spatial resolution remote sensing (HSRRS) images classification and identification have been made to acquire land surface information. The information can be used to facilitate researches on land resource management, geographical situation monitoring and global climate change. As the spectral statistical characteristics is not stable, and the same target show different spectral characteristics, the traditional spectral classification methods

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

could not obtain satisfactory results. With the development of computer technology, deep learning algorithms have been applied for classification and recognition of HSRRS images.

Deep learning has got a lot of attention for its great success in computer vision, pattern recognition, wireless communications, resource location, and automatic speech recognition [1]–[8]. It has also become the preferred method to settle a lot of challenging works in analysis of mutations in DNA [9] and natural language processing and understanding [10]. Convolutional Neural Networks (CNN) is one of the hottest topics in deep learning, and it has been applied into image classification, semantic segmentation [11], object detection, self driving cars and disease diagnosis [12]. A lot of algorithms based on CNN have also been successfully applied as powerful information extractors in HSRRS field, and have achieved good performances in classification and target detection [2], [6].

However, HSRRS images acquired by satellites or aircrafts are susceptible to various noises due to the influence of atmosphere and instruments. And recent works also show that CNN is highly sensitive to perturbations and could be easily fooled with small and human-imperceptible additive perturbations [13]–[16]. The adversarial perturbations could change the estimated label of the classifier and result in misclassifications. What's more, the same image perturbation could fool multiple classifiers based on CNN or other deep learning models. The profound effects attract a wide interest of researchers in adversarial attacks including detections or defenses [17]. In-depth understanding of the weakness of CNN classification model could lead to better attack or threat prevention strategies, and most importantly, more effective attack detections or defense methods [18], [19]. The adversarial images which are deliberately modified to attack a network are powerful to generate incorrect prediction and may lead to misclassification for the CNN based classification architecture [17], [20], [21]. All these researches have shown that state-of-the-art classifiers are likely to achieve limited robustness. Several researchers tried to make deep networks more robust to the adversarial attacks [22], such as defensive distillation [23] and adversarial training [24]. However, studies show that those approaches can only reduce the probability of generating adversarial images instead of solving the problem completely.

In this work, we aim to propose an architecture for adversarial attack and detection on classifier designed for HSRRS images based on CNN. The first purpose is to evaluate the robustness of image classifier for HSRRS images with adversarial examples. And the second purpose is to propose a method to detect the adversarial images before classification. The fast gradient sign method (FGSM) [18] and large Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) [16] are two classic algorithms for generating adversarial examples. To assess the vulnerability of CNN classifier for HSRRS images, we applied FGSM and box constrained L-BFGS to generate adversarial examples. To enhance the robustness of HSRRS image classifier, machine learning based methods are presented to detect the adversarial images in CNN hidden layer. The contributions of this paper can be summarized as follows:

- Robustness evaluation of HSRRS image classifier. Recent works have shown that the classifier for HSRRS image performs well for multi-class classification without additional noise. So when the adversarial examples are input into the classifier, how will the robustness of the classifier be affected.
- Detection approach based on machine learning. Adversarial images fool the HSRRS image classifier and make it generate incorrect prediction. In order to reduce the effect of adversarial examples and enhance the robustness of the HSRRS classifier, machine learning based detection approach is proposed to detect the adversarial images before the features are fed into classifier.

The rest of the paper is organized as follows. An overview of HSRRS image classification based on CNN, adversarial attack generation, and attack detection is presented in Section II. The attacks for remote sensing scene classification model based on FGSM and L-BFGS are given in Section III, and the accuracy and precision of classification model before and after attacks are described. Then, the attack detection methods are provided in Section V. Finally, we draw some concluding remarks in Section VI.

II. RELATED WORKS

Recent years, CNN does a good job in image classification and computer vision. Based on CNN, many efforts have been made on HSRRS images classification and identification, and a lot of effective results have been accumulated. Meanwhile, several recent works have indicated that the CNN based classifier is vulnerable when adversarial examples are input into the classification architecture. In this section, the related works on HSRRS images classification, adversarial attack generation, and attack detection for image classifiers will be investigated, separately.

A. HSRRS IMAGE CLASSIFICATION BASED ON CNN

HSRRS images presenting higher spatial resolution, clearer texture and richer color information, have been an important development direction in the field of remote sensing. They provide detailed and specific information for classification [25], urban management and planning, environment management and precision agri-forestry assessment. In recent years, deep learning, especially CNN has been introduced into HSRRS images to make identification or classification. And extensive efforts have been made in developing feature representations and constructing classifiers for classification tasks [2], [26], [27]. For example, a remote sensing region-based CNN has been proposed to detect tiny object in GF dataset [28], a symmetrical dense shortcut deep fully CNN has been developed to make semantic segmentation [11], and a multi-task CNN (RoadNet) is developed to analyze road in complex urban scenes [29]. From the previous studies, it is found that CNN has become an effective and powerful tool

for HSRRS image classification, and that smaller number of parameters needed are needed compared with full-connected networks.

B. WHITE-BOX ATTACK ALGORITHMS

Different tools modeling adversarial attacks on images have been successfully developed to fool CNN. Based on whether the targeted model is known, there are two types of attacks, namely black-box attacks and white-box attacks. The black-box attacks feed a targeted model with adversarial images or examples which are generated without the knowledge of the model. The white-box attacks know the complete knowledge of the targeted model, such as its architecture, training method, training data, and so on. Both the black-box and white-box attacks can be applied to fool the CNN-based image classification model. The adversarial examples generated by white-box attacks typically occur in multi-category classification problems, misleading the CNN to make specific category decision.

The white-box attack algorithms contain L-BFGS [16], FGSM [18], BIM & ILCM [30], Deepfool [31], C&W attacks [22], and so on, among which the L-BFGS and FGSM are two representative white-box attacks. The box-constrained L-BFGS model was first demonstrated by Szegedy et al. to solve a box-constrained optimization problem [16]. It was observed that the perturbations generated by L-BFGS and added into the clean images could fool the neural network, but for human visual system the images appear similar to the real images. To enhance adversarial training efficiency, FGSM was proposed to generate un-targeted adversarial examples, and the approach could also be extended to iterative method for targeted or un-targeted attacks [18]. Kurakin et al.improved the FGSM approach by using ''one-step target class'' to generate adversarial examples from original images [32]. FGSM has been broadly applied to generate adversarial examples to attack CNN-based (or even deep learning) image classifiers. It is reported that the top-1 error rate on the adversarial examples generated by FGSM is around 63-69% for ImageNet [32].

C. DETECTION APPROACHES

There are different kinds of detection and defense methods against the adversarial attacks. Adversarial training is one of the most useful methods, and it enables the CNN-based networks to better generalize and reflect the features of adversarial images. Therefore, the robustness of the classifier model can be improved. However, as these optimized models overly based on confident linear responses to the input, and they can be easily fooled by noise that has not appeared in the training data [18]. Similarly, Papernot et al. also indicated that it is contradictory between the model accuracy and adaptability, which has to be calibrated for each application case [33]. Network distillation was introduced to reduce the sensitivity of network and enhance the robustness of network [23]. But attackers can apply logits layer output to gain adversarial gradients and bypass the Softmax layer and defeat it.

Detection of adversarial images is another topic to enhance the robustness of the classifier model. The strategies of detecting adversarial examples can be divided into three groups: training detector, sample statistics, and inconsistency predicting. The training detector is similar to the adversarial training, which uses adversarial examples to train detectors. Metzen *et al.* proposed a method by adding parallel branch to classifier and train it to detect the adversarial examples in the input images [34]. Grosse *et al.*s proposed a detection method by adding a new ''adversarial'' class in the last layer of the CNN model [35]. And it detects adversarial images based on the statistical distribution of the real images. Sample statistics is a method based on statistical test using maximum mean discrepancy. Liang *et al.* indicated that noise reduction methods could be selectively utilized to alleviate the influence of adversarial examples. And the fake images can be detected by comparing the label of image before and after adding perturbation [36]. Xu *et al.* also showed that local or non-local spatial smoothing and color bits squeezing can reach higher success rate on discriminating the fake images [37]. An idea of inconsistency predicting is to measure the inconsistency among several models to predict an unknown input, because an adversarial image may not fool all of the deep learning models. Feinman *et al.* proposed a detection method based on ''Bayesian neural network uncertainty'' to generate predictions for an input during test time [38]. However, existing studies show that many of these defense and detection approaches can be easily evaded by adding new perturbations or changing the types of perturbation slightly on the original attacks. Therefore, it is necessary and important to propose a more robust approach for adversarial examples detection and defense.

In short, HSRRS image classifiers based on CNN have achieved abundant research results, and the robustness of CNN is indicated to be affected by adversarial perturbations. Meanwhile, the adversarial images detection approach need to be proposed to improve the robustness of the CNN-based classifiers.

III. METHODOLOGY

As multi-level representative-learning methods, deep learning are used to generate several level features. The state-of-the-art deep learning model (e.g., CNN) has been applied in classification tasks for clean and perturbed HSRRS images. The HSRRS image dataset for researches on adversarial attack and detection was constructed, where airport, avenue, bridge, building, marina, parking lot, residents, road, roadside tree, and storeroom images are extracted from the UC Merced and RSI-CB datasets. Each image in the obtained dataset is 128×128 in 3 channels, and some of them are shown in Fig. 1. The detailed information of the HSRRS image dataset and the CNN-based classifier architecture can be seen in [2].

A. ADVERSARIAL IMAGES GENERATION

As described in II, the L-BFGS and FGSM are two effective white-box adversarial attack algorithms for targeted model

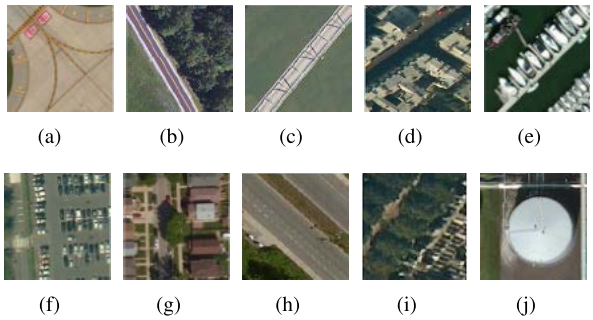

FIGURE 1. The HSRRS image dataset. (a-j) Airport, avenue, bridge, building, marina, parking lot, residents, road, roadside tree and storeroom, respectively.

based on deep learning. In this section, the principles and methods of generating adversarial images by using L-BFGS and FGSM will be described, respectively.

Since the HSRRS image dataset are in RGB scales, suppose the original image is x, the perturbation is ξ, and an image can be viewed as a vector $\boldsymbol{x} \in \boldsymbol{R}^{3mn}$. Crafting adversarial images can be formulated to find samples $\widetilde{x}$ as shown in Eq. (1) to fool the CNN models.

$$\widetilde{x} = x + \xi \tag{1}$$

L-BFGS is a classic adversarial attack algorithm to fool deep learning based models into making misclassification or misleading predictions. It is expressed as a box-constrained optimization approach, and its formula can be expressed as:

$$\begin{aligned} &\min_{\xi} A(x+\xi) \neq A(x) \\ &s.t.\, \widetilde{x} \in [0,1]^{3mn}, \quad m, n = 1, 2, 3 \ldots. \end{aligned} \tag{2}$$

For the classifier model the Eq. (2)can be written as:

$$\begin{aligned} &\min_{\xi} x\,|\xi| + \mathcal{L}(\widetilde{x}, \mathcal{C}(\widetilde{x})) \\ &s.t.\, \widetilde{x} \in [0,1]^{3mn}, m, n = 1, 2, 3 \ldots, \end{aligned} \tag{3}$$

where $\mathcal{L}$ represents the loss of the classifier and $\mathcal{C}$ is the deep learning based classifier. It is shown that Eq. (3) results in an exact solution for HSRRS image classifier with a convex loss function. This adversarial attack approach is able to compute perturbations, which can be easily added into the original images to fool the deep learning based classifier model successfully. However, this approach is computationally expensive for deep learning based classifier model.

To speed up the computation of the adversarial images, Goodfellow *et al.* [18] developed an approach to compute adversarial perturbations for original images by adopting

$$\xi = \epsilon \text{sign}(\nabla \mathcal{J}(\theta, x, \mathcal{C})), \tag{4}$$

where ϵ is a small scalar value restricting the norm of the perturbation, $\nabla \mathcal{J}(\ldots)$ computes the gradient of the cost function with current values of the model parameters θ with respect to x, and $sign(.)$ represents the sign function. The method solving Eq. (4) was first termed as FGSM, and it is based on linearity hypothesis. It perturbs images by increasing the loss of classifiers on the generated images. There are several improved FGSM algorithms by modifying the loss function, such as "one-step target class", "Fast Gradient L_2", and "Fast Gradient L_∞" [32], [39]. These algorithms use the "one-step" methods to compute gradient for adversarial examples generation to against deep learning based image classifier models.

Most of the white-box attacks generate adversarial examples by adversarial training, it is not only time-consuming and laborious, but also susceptible to the types of disturbance in the training model. Especially for multi-category classifier, these problems may easily lead to huge training burden and invalidation of attack algorithms. To reduce the training burden and test the robustness of classifier model, we apply the CNN classifier model trained by clean images, and generate the adversarial images using another clean image dataset by L-BFGS and FGSM, respectively.

B. ATTACKS ON THE HSRRS IMAGE CLASSIFIER

The flowchart involving adversarial white-box attacks on the HSRRS image classifier is shown in Fig. 2. The HSRRS image dataset is divided into training and testing datasets with a ratio about 5 : 1. And the training images are augmented by normalization, scaling, rotation, width shift, and height shift, as done in [2]. First, all of the HSRRS images are divided into two parts with the ratio 5:1 for training and testing, respectively. The HSRRS image classifier model is generated based on the clean training images. And the adversarial examples are crafted by adding perturbations to the clean testing images, where the perturbations are generated by L-BFGS or FGSM. Then the adversarial images and clean testing images are feed into the CNN classifier model to test its robustness. It is noted that the classifier model does not contain knowledge about the perturbations, which is consistent with the practical scenarios of classification. This evaluation scheme can better and objectively reflect the robustness of the CNN classifier.

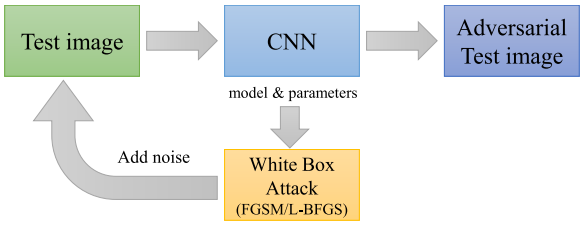

FIGURE 2. The architecture of adversarial attacks on HSRRS image classifier.

Confusion matrix is the most commonly applied indicator in classification issues. Table. 1 shows the structure of the confusion matrix for multi-category classification, where the predicated and real labels locate in horizontal and vertical, respectively. The overall accuracy is applied to evaluate the

TABLE 1. Confusion matrix.

		Predicted label				
		1	2	···	n	
True label	1	C_{11}	C_{12}	···	C_{1n}	C_{1+}
	2	C_{21}	C_{22}	···	C_{2n}	C_{2+}
	···	···	···	···	···	···
	n	C_{n1}	C_{n2}	···	C_{nn}	C_{n+}
		C_{+1}	C_{+2}	···	C_{+n}	N

proportion that the images are correctly classified, and it can be calculated by the confusion matrix. The calculation is expressed as Eq. 5, where *OA* represents the overall accuracy, n is the number of class, and N is the total number of the samples.

$$OA = \frac{1}{N}\sum_{k=1}^{n} C_{kk}$$
$$with \quad k = 1, 2, 3, \cdots, n. \tag{5}$$

What's more, an indicator named kappa coefficient, which is calculated from the confusion matrix, is often applied to evaluate the consistency and classification precision. As indicated in Eq. 6, the kappa coefficient not only considers the overall accuracy but also considers the imbalanced number of samples in each category,

$$kc = \frac{p_0 - p_e}{1 - p_e},$$
$$p_0 = \frac{1}{N}\sum_{k=1}^{n} C_{kk},$$
$$p_e = \frac{1}{N^2}\sum_{k=1}^{n}(C_{k+}C_{+k}), \quad k = 1, 2, 3, \cdots, n. \tag{6}$$

where kc is the kappa coefficient, p_0 is the overall accuracy (*OA* in Eq. 5).

C. ADVERSARIAL IMAGES DETECTION

As illustrated above, it is consuming time to train the adversarial examples or statistic the differences between legitimate and fake examples during the model training. SVM is a supervised learning method with stable effects and fast prediction capacity. It is designed to find a hyperplane to segment the positive and negative examples, which ensures the largest interval between the two classes. The SVM is very suitable for bi-classification. Therefore, we proposed an adversarial image detection method based on SVM. After the adversarial examples are generated, the features extracted in the first and second full connected layers of the HSRRS image classifier model are then fused and fed into the SVM to detect the fake images. And finally, the detection precision and accuracy are presented.

Acquiring the hyperplane is the key issue for the SVM. Given a linearly separable training dataset, the classification hyperplane can be obtained by maximizing or equivalently solving the corresponding convex quadratic programming problem; the decision formula and hyperplane representation are as follows.

$$f(x) = \text{sign}(w \cdot x + b)$$
$$w \cdot x + b = 0, \tag{7}$$

where the hyperplane divides the features into two parts: the positive class, and the negative class.

To enhance the utilization of multi-scale information, a hybrid model is developed. Its specific flowchart is shown in Fig. 3. From the figure, we can see that the features generated in the first (Dense 1024) and second (Dense 512) full connected layers are extracted to make a fusion. The fused features are then applied in the SVM classifier. The hybrid model synthesizing two different scales of information may improve the detection rate.

Similar to the evaluation of the robustness for the HSRRS image classifier before and after adversarial attacks. The performance of adversarial images detection is also evaluated by the confusion matrix. As the adversarial detection focuses on the detection rate of adversarial images, the confusion matrix is simplified into a binary classification. In table. 2, TP, TN,FP and FN are parameters for calculating detection probability (P_D), overall accuracy (*OA*), fake alarm probability (P_{FA}) and miss probability (P_M)(eq. 8),

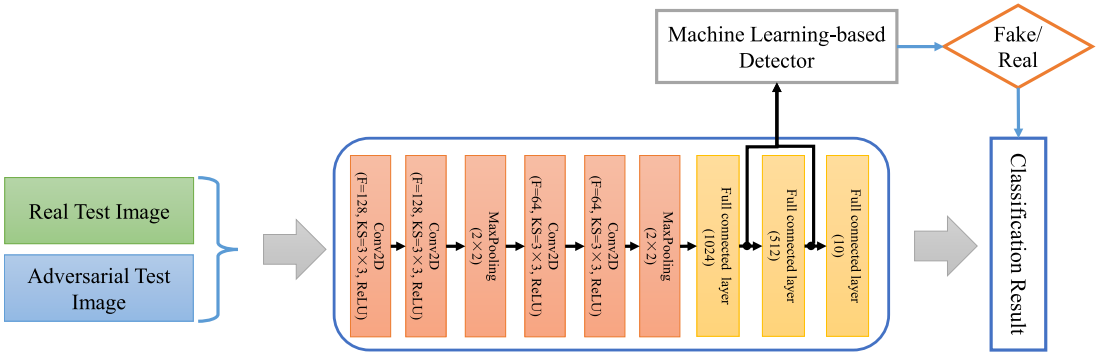

FIGURE 3. The flowchart of hybrid model for adversarial examples detection.

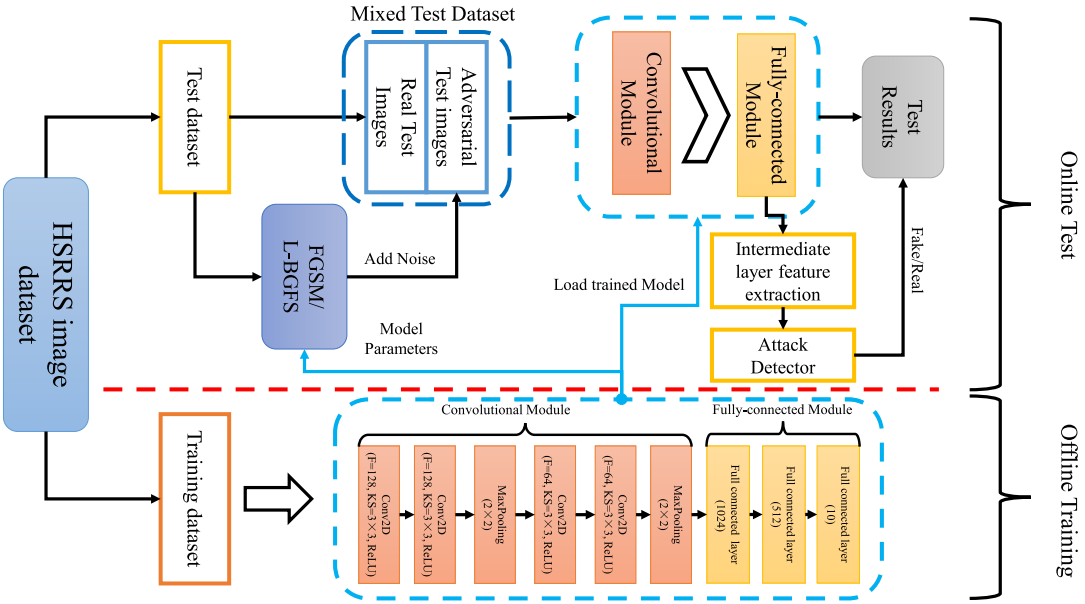

FIGURE 4. The flowchart of adversarial attack and detection for HSRRS image classifier.

TABLE 2. The confusion matrix of binary classification.

	True label	
Predicted label	Adversarial examples	True examples
Adversarial examples	TP(true positive)	FP(false positive)
True examples	FN(false negative)	TN(true negative)

which are applied to evaluate the overall performance of our detection method.

$$\begin{aligned} P_D &= \frac{TP}{TP+FN}, \\ OA &= \frac{TP+TN}{TP+TN+FP+FN}, \\ P_{FA} &= \frac{FP}{TP+FP}, \\ P_M &= \frac{FN}{TP+FN}. \end{aligned} \tag{8}$$

IV. EXPERIMENT RESULTS

The HSRRS image classifier based on CNN obtained a good performance in urban built-up areas. To evaluate the robustness of the classifier, we used two different white-box attacks: L-BFGS and FGSM, to generate adversarial images.The detailed flowchart is shown in Fig. 4. First, all of the HSRRS images are divided into two parts with the ratio 5:1 for training and testing, respectively. And the HSRRS image classifier model is generated with the clean training images. Second, the L-BFGS and FGSM algorithms are applied to generate adversarial examples, respectively. Both the clean test images and the adversarial examples are fed into the HSRRS image classifier model. Third, the features generated in the first and second full connected layers are extracted to make a fusion, the followed SVM is to detect the fake images. And finally, the evaluation parameters are presented. The architecture of the HSRRS image classifier is described in [2] in detail, and we focus on the adversarial attacks to the HSRRS image classifier and the corresponding detection problem in this paper.

A. ADVERSARIAL ATTACKS

The L-BFGS and FGSM are applied to generate adversarial examples, respectively. The white-box attack algorithms master the detailed information of the CNN-based HSRRS image classifier model, and generate adversarial images by adding noise. Fig. 5 shows the adversarial images generation procedure using L-BFGS and FGSM, respectively. And the images in the left are clean, and the adversarial images are generated after adding noises. It indicates that the adversarial image

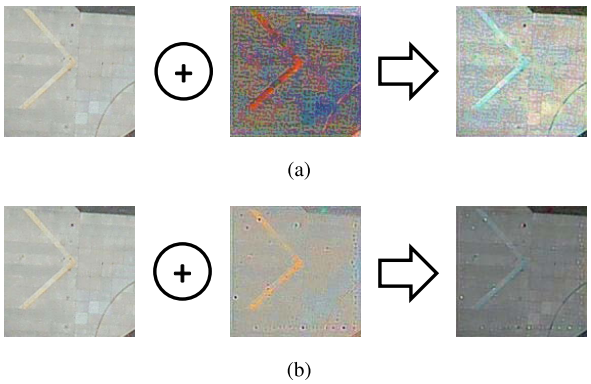

FIGURE 5. The adversarial images generation procedure: (a) FGSM, (b) L-BFGS.

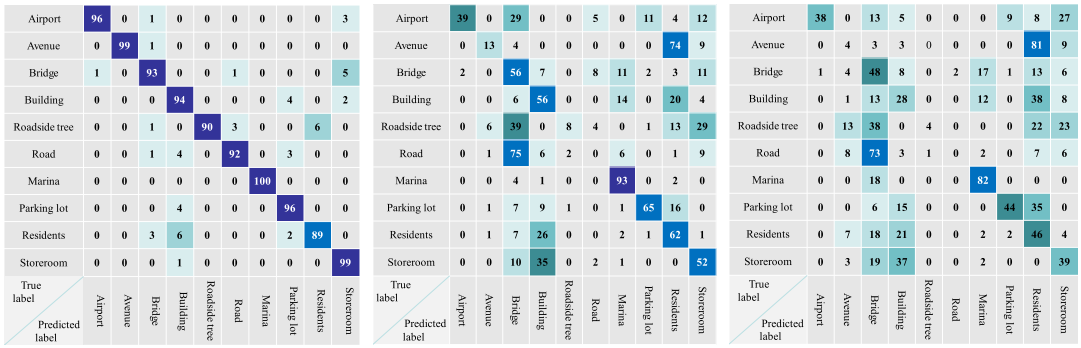

FIGURE 6. The confusion matrix of (a)original HSRRS image classifier, (b)L-BFGS attacked model and (c)FGSM attacked model.

generated by the FGSM are quite different with the original one. But the adversarial images generated by the L-BFGS are similar with the original ones except for the brightness. The adversarial images combined with the clean testing images are fed into the HSRRS image classifier. In order to access the robustness of the HSRRS image classifier, the accuracy indicators are calculated before and after the white-box attacks.

B. ADVERSARIAL DETECTION

As shown in Fig. 4, the test contains the original clean testing images and the adversarial images generated by L-BFGS or FGSM. To improve the robustness of the classifier model, the features generated by the full connected layers are extracted before they are fed into the softmax layer. And then the extracted features are fed into the SVM to detect the fake images. In order to comprehensively and effectively utilize the multi-scale information, the features extracted from two different level full connected layers are fused. And then the characteristics of the fusion is fed into the SVM classifier to detect the fake images.

V. RESULTS AND ANALYSIS

The results and analyses cover two parts: evaluation of the HSRRS image classifier before and after adversarial attacks, and assessment of the adversarial attack detection model.

A. ATTACKS ON HSRRS IMAGE CLASSIFIER

As described in section III, confusion matrix, overall accuracy, and kappa coefficient are used to evaluate the performance of the HSRRS classifier. Fig. 6 shows the confusion matrix of HSRRS image classifier before and after the white-box attacks. Fig. 6 (a) is the confusion matrix of the original HSRRS image classifier, which indicates that the diagonal elements occupies more than 90% of each category in test dataset. After the L-BFGS attack is introduced, the confusion matrix is shown as in Fig. 6 (b), where it can be seen that the large majority elements do not concentrate on the diagonal line. Fig. 6 (c) shows the confusion matrix after the FGSM attack is introduced, and it seems similar to that of 6 (b).

TABLE 3. Evaluation indicators for the classifier.

Models	Indicators	
	OA	kc
Original model	96.4%	0.960
L-BFGS	44.4%	0.385
FGSM	33.3%	0.259

The table 3 shows the overall accuracy (OA) and kappa coefficient (kc) of the three models. We can see that after the L-BFGS or FGSM attacks, both of the OA and kc decrease rapidly. It indicates that the CNN-based HSRRS image classifier is vulnerable when attacked by the L-BFGS or FGSM. Meanwhile, the OA and kc of the model attacked by the L-BFGS are larger than those of the FGSM attacked model. Table 3 also illustrates that for our HSRRS classifier model, the FGSM is more destructive to the model's recognition ability.

B. DETECTIONS ON ADVERSARIAL IMAGES

As described in V-A, the OA of the HSRRS image classifier declines from 96.4% to 44.4% (due to the L-BFGS) and 33.3% (due to the FGSM), respectively. Therefore, the adversarial detection evaluations here are carried out in two parts concerning the L-BFGS and FGSM, respectively.

1) EVALUATION OF DETECTION MODEL FOR CLASSIFIER ATTACKED BY L-BFGS

The evaluation of the detection model built for the HSRRS image classifier (attacked by L-BFGS) can be carried out as two aspects: one is the detection model based on single level features, and the other is the fused detection model based on features at two different levels. Both of the features are extracted from the full-connected layer. The evaluation indicators include overall accuracy OA, detection probability P_D, fake alarm probability P_{FA}, and miss probability P_M. And there are three detection models (DM), among which dense1 and dense2 represents the models based on single level features extracted from the first dense layer and the

second dense layer, respectively. And the fusion represents the model built on the basis of the two level features of dense1 and dense2. As shown in table 4, the *OA* of the detection model reaches 92.8% and 91.0% for the first and second level features, respectively. The P_D, P_{FA}, and P_M metrics also show that the dense1 model performs better than the dense2 model. After fusing the dense1 and dense2 models, we obtain the fusion model. It is very clear that the fusion model almost has no improvement in the detection compared with the dense1 model. But the fusion model is better than the dense2 model in all aspects.

TABLE 4. The detection evaluation for classifier attacked by L-BFGS.

DM \ EI	OA	P_D	P_{FA}	P_M
Dense1	92.8%	0.903	0.049	0.097
Dense2	91.0%	0.890	0.073	0.11
Fusion	92.8%	0.903	0.049	0.097

2) EVALUATION OF DETECTION MODEL FOR CLASSIFIER ATTACKED BY FGSM

Similar to the evaluation of the detection model for the classifier attacked by the L-BFGS, this section also presents the performance of the detection model based on single level features and two different level features, respectively.

The table. 5 illustrates the indicators for assessing the detection models. It can be seen that the adversarial detection models based on the single level features obtain accuracies of 94.0% and 93.3%, respectively. And the corresponding detection probabilities are 0.933 and 0.923, respectively. After fusing two different level features, the fusion model acquires an accuracy of 94.5% and a detection probability of 0.933. Besides, it is noted that the fake alarm probability of the fusion model decreases to 0.040 compared with the dense1 (0.054) and dense2 (0.058) models. All of these metrics indicate that the adversarial detection model based on the SVM is capable to detect the fake images with an accuracy above 93.3%. And the fusion model is improved in terms of accuracy and fake alarm probability when compared with the best single level features based model. Therefore, the SVM based fusion detection model performs the best in recognizing fake HSRRS images for the HSRRS image classifier attacked by FGSM.

TABLE 5. Detection evaluation for the classifier attacked by the FGSM.

DM \ EI	OA	P_D	P_{FA}	P_M
Dense1	94.0%	0.933	0.054	0.067
Dense2	93.3%	0.923	0.058	0.077
Fusion	94.5%	0.933	0.040	0.067

VI. CONCLUSION

In this paper, an architecture for adversarial attack generation and detection on the CNN-based HSRRS image classifier has been proposed. The robustness of the HSRRS image classifier has been assessed by white-box attacks (L-BFGS and FGSM), and the detection model based on the SVM has been proposed to detect the adversarial images. Meanwhile, two level features fusion can be combined to build an ensemble model to improve the detection ability. The robustness of the CNN-based classifier have been confirmed in terms of confusion matrix, *OA*, and *kc*. The results reveal that the HSRRS image classifier is vulnerable when attacked by L-BFGS or FGSM, and the *OA* decreases from 96.4% to 44.4% and 33.3%, respectively. The detection method for L-BFGS or FGSM attacked HSRRS image classifier model have also been assessed in terms of *OA*, P_D, P_{FA}, and P_M. The results show that the detection model based on the first level features performs better than that based on the second level features. For the HSRRS image classifier attacked by the L-BFGS, the evaluation indicators of the fusion model is not better than those of the dense1 model. Meanwhile, for the classifier attacked by the FGSM, the fusion model improves approximately 0.5% in *OA*, and decreases 0.014 in P_{FA} when compared with the dense1 model. Therefore, we can conclude that for the CNN-based HSRRS image classifier, the FGSM is more destructive for the stability of the classifier, and the adversarial images generated by the FGSM are more likely to be detected in our proposed methods.

There are still some problems to be studied further, for example, black-box attacks may be applied in the HSRRS image classifier, and other machine learning methods such as random forest (RF) and k-nearest neighbor (KNN) could be used to detect adversarial attacks. What is more, adversarial defence for HSRRS and other remote sensing image classifiers based on CNN could be considered.

VII. ACKNOWLEDGMENT

Portions of the research in this paper use the RSI-CB and UC Merced datasets.

REFERENCES

[1] R. Zhu, Z. Wang, Z. Ma, G. Wang, and J.-H. Xue, "LRID: A new metric of multi-class imbalance degree based on likelihood-ratio test," *Pattern Recognit. Lett.*, vol. 116, pp. 36–42, Dec. 2018.

[2] W. Li, H. Liu, Y. Wang, Z. Li, Y. Jia, and G. Gui, "Deep learning-based classification methods for remote sensing images in urban built-up areas," *IEEE Access*, vol. 7, pp. 36274–36284, 2019.

[3] Y. Wang, M. Liu, J. Yang, and G. Gui, "Data-driven deep learning for automatic modulation recognition in cognitive radios," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 4074–4077, Apr. 2019.

[4] M. Liu, T. Song, and G. Gui, "Deep cognitive perspective: Resource allocation for NOMA-based heterogeneous IoT with imperfect SIC," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2885–2894, Apr. 2019.

[5] H. Huang, W. Xia, J. Xiong, J. Yang, G. Zheng, and X. Zhu, "Unsupervised learning-based fast beamforming design for downlink MIMO," *IEEE Access*, vol. 7, pp. 7599–7605, 2018.

[6] Y. Zhong, X. Han, and L. Zhang, "Multi-class geospatial object detection based on a position-sensitive balancing framework for high spatial resolution remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 138, pp. 281–294, Apr. 2018.

[7] H. Huang, Y. Song, J. Yang, G. Gui, and F. Adachi, "Deep-learning-based millimeter-wave massive MIMO for hybrid precoding," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 3027–3032, Mar. 2019.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[9] H. Y. Xiong, B. Alipanahi, L. J. Lee, H. Bretschneider, D. Merico, R. K. C. Yuen, Y. Hua, S. Gueroussov, H. S. Najafabadi, T. R. Hughes, Q. Morris, Y. Barash, A. R. Krainer, N. Jojic, S. W. Scherer, B. J. Blencowe, and B. J. Frey, "The human splicing code reveals new insights into the genetic determinants of disease," *Science*, vol. 347, no. 6218, p. 144, 2015.

[10] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc Adv. Neural Inf. Process. Syst.*, 2014, pp. 1–9. [Online]. Available: https://arxiv.org/abs/1409.3215

[11] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[12] J.-G. Lee, S. Jun, Y.-W. Cho, H. Lee, G. B. Kim, J. B. Seo, and N. Kim, "Deep learning in medical imaging: General overview," *Korean J. Radiol.*, vol. 18, no. 4, pp. 570–584, 2017.

[13] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE Trans. Autom. Control*, vol. 59, no. 6, pp. 1454–1467, Jun. 2014.

[14] H. Kwon, Y. Kim, K.-W. Park, H. Yoon, and D. Choi, "Multi-targeted adversarial example in evasion attack on deep neural network," *IEEE Access*, vol. 6, pp. 46084–46096, 2018.

[15] 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.

[16] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, D. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014, pp. 1–10. [Online]. Available: https://arxiv.org/abs/1312.6199

[17] A. Fawzi, S.-M. Moosavi-Dezfooli, and P. Frossard, "The robustness of deep networks: A geometrical perspective," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 50–62, Nov. 2017.

[18] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014, pp. 1–11. [Online]. Available: https://arxiv.org/abs/1412.6572

[19] Y. Liu, X. Chen, L. Chang, and D. Song, "Delving into transferable adversarial examples and black-box attacks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–24. [Online]. Available: https://arxiv.org/abs/1611.02770

[20] Z. Wang, "Deep learning-based intrusion detection with adversaries," *IEEE Access*, vol. 6, pp. 38367–38384, 2018.

[21] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Trans. Evol. Comput.*, to be published. doi: 10.1109/TEVC.2019.2890858.

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

[23] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2016, pp. 582–597.

[24] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proc. IEEE Eur. Symp. Secur. Privacy Z (EuroS P)*, Mar. 2016, pp. 372–387.

[25] L. Zhang, X. Huang, B. Huang, and P. Li, "A pixel shape index coupled with spectral information for classification of high spatial resolution remotely sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 10, pp. 2950–2961, Oct. 2006.

[26] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

[27] Z. Hong, D. Ming, K. Zhou, Y. Guo, and T. Lu, "Road extraction from a high spatial resolution remote sensing image based on richer convolutional features," *IEEE Access*, vol. 6, pp. 46988–47000, 2018.

[28] J. Pang, C. Li, J. Shi, Z. Xu, and H. Feng, "R^2-CNN: Fast tiny object detection in large-scale remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, to be published.

[29] Y. Liu, J. Yao, X. Lu, M. Xia, X. Wang, and Y. Liu, "Roadnet: Learning to comprehensively analyze road networks in complex urban scenes from high-resolution remotely sensed images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2043–2056, Apr. 2019.

[30] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2016, pp. 1–13. [Online]. Available: https://arxiv.org/abs/1607.02533

[31] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2574–2582.

[32] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nov. 2016, pp. 1–17. [Online]. Available: https://arxiv.org/abs/1611.01236

[33] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman, "Towards the science of security and privacy in machine learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nov. 2016, pp. 1–19. [Online]. Available: https://arxiv.org/abs/1611.03814

[34] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff, "On detecting adversarial perturbations," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Feb. 2017, pp. 1–12. [Online]. Available: https://arxiv.org/abs/1702.04267

[35] K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. McDaniel, "On the (statistical) detection of adversarial examples," Feb. 2017, *arXiv:1702.06280*. [Online]. Available: https://arxiv.org/abs/1702.06280

[36] B. Liang, H. Li, M. Su, X. Li, W. Shi, and X. Wang, "Detecting adversarial image examples in deep neural networks with adaptive noise reduction," *IEEE Trans. Depend. Sec. Comput.*, to be published. doi: 10.1109/TDSC.2018.2874243.

[37] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," in *Proc. Netw. Distrib. Syst. Secur. Symp. (NDSS)*, 2017, pp. 1–15. [Online]. Available: https://arxiv.org/abs/1704.01155

[38] R. Feinman, R. R. Curtin, S. Shintre, and A. B. Gardner, "Detecting adversarial samples from artifacts," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 1–9. [Online]. Available: https://arxiv.org/abs/1703.00410

[39] 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. doi: 10.1109/TPAMI.2018.2858821.

WENMEI LI (M'18) received the M.S. degree from Nanjing University, in 2010, and the Ph.D. degree from the Chinese Academy of Forestry, in 2013. She is currently pursuing the Ph.D. degree with the Nanjing University of Posts and Telecommunications, where she is currently an Associate Professor with the School of Geographic and Biologic Information. Her research interests include deep learning, optimization, and image reconstruct and their application in land remote sensing.

ZHUANGZHUANG LI received the B.E degree in communication engineering from Xi'an Technological University, China, in 2018. He is currently pursuing the master's degree with the Nanjing University of Posts and Telecommunications, Nanjing, China. His research interests include machine learning, optimization, and its application in remote sensing image processing.

JINLONG SUN (M'18) received the M.S. and Ph.D. degrees from the Harbin Institute of Technology (HIT), Harbin, China, in 2014 and 2018, respectively. He is currently an Assistant Professor with the Nanjing University of Posts and Telecommunications (NUPT), Nanjing, China. His current research interests include signal processing for wireless communications, machine learning, and integrated navigation systems.

YU WANG (S'18) received the B.S. degree in communication engineering from the Nanjing University of Posts and Telecommunications (NUPT), Nanjing, China, in 2018, where he is currently pursuing the Ph.D. degree. His research interests include deep learning, optimization, and its application in wireless communications.

HAIYAN LIU received the B.E degree in electronic information engineering from Hefei Normal University, Hefei, China, in 2018. She is currently pursuing the master's degree with the Nanjing University of Posts and Telecommunication, Nanjing, China. Her research interests include deep learning, optimization, and its application in remote sensing image processing.

JIE YANG received the M.Sc. and Ph.D. degrees in communication engineering from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2006 and 2018, respectively. Her research interests include deep learning, wireless communication systems, and UAV communications.

GUAN GUI (M'11–SM'17) received the Dr.Eng. degree in information and communication engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2012. From 2009 to 2012, with the financial supported from the China Scholarship Council (CSC) and the Global Center of Education (ECOE), Tohoku University, he joined the Wireless Signal Processing and Network Laboratory (Prof. Adachi Laboratory), Department of Communications Engineering, Graduate School of Engineering, Tohoku University, as a Research Assistant and a Postdoctoral Research Fellow. From 2012 to 2014, he was supported by the Japan Society for the Promotion of Science (JSPS) Fellowship as a Postdoctoral Research Fellow with the Wireless Signal Processing and Network Laboratory (Prof. Adachi Laboratory). From 2014 to 2015, he was an Assistant Professor with the Department of Electronics and Information System, Akita Prefectural University. Since 2015, he has been a Professor with the Nanjing University of Posts and Telecommunications (NUPT), Nanjing, China. His current research interests include deep learning, compressive sensing, and advanced wireless techniques. He received several best paper awards such as CSPS2018, ICNC2018, ICC2017, ICC2014, and VTC2014-Spring. He was also selected as for Jiangsu Specially Appointed Professor, Jiangsu High-Level Innovation and Entrepreneurial Talent and Nanjing Youth Award. He was an Editor of *Security and Communication Networks*, from 2012 to 2016, has been an Editor of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, since 2017, *KSII Transactions on Internet and Information System*, since 2017, of IEEE ACCESS, since 2018.