Semi-Supervised Representation Learning for Remote Sensing Image Classification Based on Generative Adversarial Networks

PEIYAO YAN, FENG HE, YAJIE YANG, AND FEI HU
School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China
National Key Laboratory of Science and Technology on Multi-Spectral Information Processing, Wuhan 430074, China
Corresponding author: Feng He (hefeng@hust.edu.cn)

ABSTRACT In the existing studies on remote sensing image scene classification, the supervised learning methods which are fine-tuned from pre-trained model require a large amount of labeled training data and parameters, while unsupervised learning methods do not make full use of label information, and the classification performance could be improved. In this paper, we introduced semi-supervised learning into generative adversarial network (GAN), so the discriminator learned more discriminative features from labeled data and unlabeled data. Moreover, the mixup data augmentation method was introduced into our classification model to augment the data and stabilized the training process. We carried out extensive experiments for both UC-Merced and NWPU-RESISC45 datasets with a 5-fold cross-validation protocol using a linear SVM as classifier. We trained the proposed method on UC-Merced dataset and achieve an average overall accuracy of 94.05% under 80% training ratio. When trained on NWPU-RESISC45 dataset, the proposed method reached an average overall accuracy of 83.12% and 92.78% under the training ratios of 20% and 80% respectively, which achieves the state-of-the-art deep learning methods without pre-training.

INDEX TERMS Remote sensing scene classification, semi-supervised learning, generative adversarial networks, data augmentation.

I. INTRODUCTION

The currently available instruments (e.g. multispectral, hyperspectral, synthetic aperture radar, etc.) for earth observation generate more and more different types of airborne or satellite images with different resolutions [1]. With the development of remote sensing imaging technology, the quantity and quality of high resolution remote sensing images are increasing. In recent years, to interpret these images automatically and accurately has received considerable amount of attention. As for one of the hot topics, remote sensing image scene classification is to categorize scene images into a discrete set of meaningful land use and land cover classes according to the image contents [1]. Remote sensing image classification has important application requirements and is widely used in natural disaster detection, land resource utilization and coverage management, geospatial object detection, geographic image retrieval, vegetation mapping, environment monitoring, and urban planning [1]–[8].

Some remote sensing image datasets have been published, including UC-Merced [9], WHU-RS19 [10], NWPU-RESISC [1], etc. It can be observed from these datasets that objects in the same category usually have different size, colors and angles while other objects may be around the target area. These reasons cause high intra-class variance and low inter-class variance. Furthermore, many of the datasets have small scale of scene classes and images per class, which poses more challenges on data-driven algorithms. Therefore, learning robust and discriminative representations from remote sensing images is very challenging but crucial.

Many remote sensing image scene classification methods have been proposed. Most of those methods can broadly be categorized into two groups: feature encoding methods and feature learning methods [11]. The former relies on hand-crafted local image descriptors to represent images, while the
latter learns to extract features from data automatically. Handcrafted feature extraction needs expert experiences and may lack the flexibility to different scenes. Among the learning based methods, deep learning especially convolutional neural network has demonstrated superior performances in many computer vision tasks, such as image classification, object detection, and instance segmentation etc. The classification neural network shows very good ability after training by fine-tuning parameters.

In the existing studies on remote sensing image scene classification, the state-of-the-art results are achieved through end to end supervised feature representation learning fine-tuned from pre-trained model which require a large amount of labeled training data and parameters. As for unsupervised feature learning, generative adversarial network (GAN) [9] is one of the most potential unsupervised learning methods in recent years, which trains networks by means of adversarial training. The adversarial training method means to train the networks with both normal samples and adversarial samples which are generated by generator network. In this way, GAN not only generates a large number of samples and enlarge the training set, but also improves the feature extraction ability of the networks and the generalization performance of the classifier through the adversarial training. However, the existing GAN based remote sensing image classification methods are still unsupervised for feature extraction, because these methods do not use the labels of the training data and the generated images of the generator which are used to train GAN are unlabeled as well. Therefore, to make full use of label information of the real data and unlabeled generated data, we proposed a semi-supervised learning model based on GAN for remote sensing image scene classification.

The contributions of this paper are the following:

1) In combination with semi-supervised learning and GAN, the semi-supervised classification model based on GAN is established.

2) The mixup data augmentation method is introduced into our classification model to augment the data and stabilize the training process.

3) The results of experiments on UC-Merced dataset showed that the proposed method outperforms state-of-art models based on GAN in terms of overall classification accuracy. As for NWPU-RESISC45 dataset, our method outperforms state-of-art deep learning methods under the premise of no pre-training.

The rest of this paper is organized as follows. In section II, the related works are illustrated. Section III introduces a remote sensing image scene classification method based on GAN. Section IV shows the experimental results and analysis. Finally, the conclusions are drawn in section V.

II. RELATED WORKS

Szegedy et al. [13] found that they can cause the network to misclassify an image by applying a certain hardly perceptible perturbation which was called adversarial sample. Subsequently, they proposed adversarial training [14] by using the adversarial samples. Adversarial training means the network trains both the samples from original dataset and the adversarial samples, so as to improve the generalization ability of the network. In order to make adversarial samples look more realistic, Yang and Newsam [9] proposed generative adversarial network which was inspired by two-player game theory, in which generator model (G) and discriminator model (D) were the two players. The purpose of G is to map random noise to samples which are real enough to fool D while D needs to discriminate real samples and generated samples. Therefore, the objective function of GAN is to find a Nash equilibrium to the following two-player min-max problem:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x \sim p_r(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z}[\log(1 - D(G(z)))]$$

where $z$ obeys a prior noise distribution $p_z$, $p_r(x)$ denotes the real distribution of the dataset.

Szegedy et al. [13] pointed out that the same adversarial samples would be misclassified by different classifiers, even if they used different training sets. However, the adversarial samples were added with certain perturbation. Not every random noise can make the original image be misclassified. Akhtar and Mian [15] reviewed the works which designed adversarial attacks, analyzed the existence of such attacks, proposed defenses against them and emphasized that adversarial attacks were possible in practical conditions. Therefore, adversarial examples matter because training a model to resist them can improve its accuracy on non-adversarial examples [16]. The neural network has defects on generalization at present and the purpose of adversarial training is to improve the robustness and generalization ability of the model. In the subsequent research, DCGAN [17] made some progress in image synthesis. This method introduced CNN into the GAN model for unsupervised training, optimized the network structure by adding convolutional layer and batch normalization layer. BigGANs [18] trained GAN at large scale and allowed fine control over the trade-off between sample fidelity and variety. What’s more, GAN has been applied to medical image processing [19], [20], semi-supervised classification [21] and semantic segmentation of remote sensing images [22] to improve the feature extraction ability of the network. To sum up, whether there is an actual adversarial sample to attack classification model, introducing GAN into image classification can improve the feature extraction ability of the model.

Special equipment or manual labeling is often needed to obtain the labels of the data which undoubtedly requires a lot of time and expertise knowledge. As for the unlabeled samples are much more than the labeled samples, semi-supervised learning is proposed to make full use of limited label data and Salimans T et al. introduced semi-supervised learning into GAN and applied it to classification by using both labeled and unlabeled data to train the classifier [21]. Tobias Springenberg et al. [23] proposed...
CatGAN and built loss function of the semi-supervised learning model with entropy and mutual information. The classification results of experiments on MNIST dataset and CIFAR-10 dataset were improved.

Cheng et al. [1] reviewed the existing remote sensing image classification datasets, and published the large-scale NWPU-RESISC45 dataset. In [1] the authors listed some results using handcrafted feature based methods, unsupervised feature learning based methods, deep CNN feature learning methods and fine-tuned deep CNN feature learning methods. The results of experiment show that the fine-tuning method has the best performance on the dataset. Zhou et al. [24] proposed two-pathway ResNet (ResNet-TP) in which the input images went through two paths of convolutional operations after a few layers and aggregated the contextual information. The method improved the discriminative ability of the features and the experiments on NWPU-RESISC45 dataset showed that the proposed mechanism achieved promising improvements over state-of-the-art methods. Due to the limitation of the labeled data and difficulty to apply supervised learning, GAN would be the excellent choice to tackle this issue because it is an unsupervised learning method in which the required quantities of training data would be provided by its generator. Lin et al. [2] proposed MARTA-GAN which applied GAN to extract features of remote sensing images through unsupervised representation learning at the first time. They proposed a multi-feature layer of concatenating last three convolutional layers for discriminator and perceptual loss for generator to learn better image representations. Recently, capsule network has been used in remote sensing image scene classification and shows promising performance [4], [5].

As for small-scale problems, like remote sensing image classification, using data augmentation to enlarge the dataset can improve the performance. The traditional data augmentation methods include rotation, cropping, changing image color difference, distorting image features and adding random noise (Gaussian noise, salt-pepper noise) which require professional knowledge [25] to determine the appropriate method for different datasets. Moreover, the traditional methods produce new samples which share the same class of the original dataset, thus do not consider the vicinity distribution between different classes. Zhang et al. [26] proposed mixup data augmentation method to enlarge the dataset, which weighted and summed the randomly selected images and corresponding labels respectively. The mixup data augmentation method encourages the model to behave linearly between training examples and reduces the inadaptation of the model when predicting outside training examples.

### III. PROPOSED METHODS

In the existing studies on remote sensing image scene classification, the supervised feature learning models are end-to-end trained from pre-trained model, like ImageNet, with a lot of fine-tunings. However, unsupervised learning methods, including GAN, do not make full use of label information, and the classification performance could be improved. Therefore, we introduce the semi-supervised learning into GAN to extract features of the remote sensing images, so as to use both the labeled training data and the unlabeled generated data to train GAN. In this paper, we propose a classification model based on GAN with semi-supervised learning, aiming to improve the feature extraction ability of the discriminator and the classification performance.

As data augmentation method can improve the generalization ability of neural networks, especially for small training datasets. In this paper, we introduce the mixup data augmentation method into GAN to relieve the problem of small dataset, and stabilize the training process.

#### A. CLASSIFICATION MODEL BASED ON SEMI-SUPERVISED FEATURE EXTRACTION

Consider an ordinary supervised learning classifier which takes in $x$ as input and outputs a $K$-dimensional vector $\{l_1, \ldots, l_K\}$ as the predict labels. The output vector can be turned into class probabilities by applying the softmax:

$$P(y = j|x) = \frac{\exp(l_j)}{\sum_{k=1}^{K} \exp(l_k)}.$$  \hspace{1cm} (2)

In supervised learning, we usually minimize the cross-entropy of class labels and predict probabilities to optimize parameters. In recent decades, semi-supervised learning has become a hot topic in order to train models with less or unlabeled data. The technique of training classifiers with both labeled data and unlabeled data is called semi-supervised learning. Usually, we use a small amount of labeled data and a large amount of unlabeled data, which are from the same domain, to train a neural network. For the purpose of image generation, the output of the discriminator $D$ is only one-dimensional, which is used to distinguish between real data and generated data. After the training, the generator $G$ is used to generate images while the discriminator $D$ is used to guide the generator $G$ during the training process. As for semi-supervised learning, we add $K$ dimensions to the output of the discriminator $D$. Letting $K$ denotes the number of classes of the dataset. After the training, we use the discriminator $D$ to extract features and the generator $G$ is used to generate unlabeled images to improve the ability of the discriminator $D$.

The goal of training the generator $G$ is to produce samples that fool the discriminator $D$. The output of the generator $G(z)$ is an image which will be used as the input for the discriminator $D$. Therefore, we want to maximize $D_{k+1}(G(z))$ which indicates the probability that $G(z)$ is judged as a real sample. Letting $D_{k+1}(G(z))$ denote the $k$-th dimension of the output of the discriminator $D$ when its input is generated image. The objective function of training generator is to minimize:

$$\min_{G} \ell_g = E_{z \sim P_z}[\log(1 - D_{k+1}(G(z))].$$  \hspace{1cm} (3)

To make the images generated by generator more similar to the real images, we train the generator to match the features in
multi-feature layer of the discriminator. Letting \( f(x) \) denotes activations on the multi-feature layer of the discriminator, the loss of feature matching for the generator is defined as:

\[
\min \|f(x) - f(G(z))\|_2^2.
\]

Therefore, the object function of the generator is:

\[
\min G V(G) = E_{x \sim P_{data}(x)}[\log (1 - D_{k+1}(G(z)))] + w_1 E_{z \sim P_{data}(z)}[\log (1 - D_{k+1}(G(z)))].
\]

where \( w_1 \) is a weighting factor. The objective function of the discriminator is composed of two parts, supervised learning loss and unsupervised learning loss. For unsupervised learning, the discriminator only needs to judge the input image is true or false. In this part, class labels are not needed. Therefore, the object function of unsupervised learning is to maximize:

\[
\max V(D)_{\text{unsupervised}} = E_{x \sim P_{data}(x)}[\log D_k(x)] + E_{z \sim P_{data}(z)}[\log (1 - D_{k+1}(G(z)))].
\]

In order to use the labels of training data, the loss function of supervised learning is defined in the form of cross-entropy. In this part, the class label is coded as one-hot vector. Letting \( y_i \) denote the ith dimension of the label, and \( D_i(x) \) denote the ith dimension of the output of discriminator \( D \) when its input is real image. Therefore, the objective function of supervised learning is to maximize:

\[
\max V(D)_{\text{supervised}} = E_{x, y \sim P_{data}(x,y)} \sum_{i=1}^{k} y_i \log D_i(x).
\]

The overall object function of discriminator is defined as:

\[
\max V(D) = w_2 V(D)_{\text{supervised}} + V(D)_{\text{unsupervised}} = w_2 E_{x, y \sim P_{data}(x,y)} \sum_{i=1}^{k} y_i \log D_i(x) + E_{x \sim P_{data}(x)}[\log D_k(x)] + E_{z \sim P_{data}(z)}[\log (1 - D_{k+1}(G(z)))],
\]

where \( w_2 \) is a weighting factor. Therefore, the whole training process (the combination of (5) and (8)) is represented as:

\[
\min \max V(D, G) = w_2 E_{x, y \sim P_{data}(x,y)} \sum_{i=1}^{k} y_i \log D_i(x)
+ E_{x \sim P_{data}(x)}[\log D_k(x)] + E_{z \sim P_{data}(z)}[\log (1 - D_{k+1}(G(z)))]
+ w_1 E_{z \sim P_{data}(z)}[\|f(x) - f(G(z))\|_2^2].
\]

\[\text{B. MIXUP DATA AUGMENTATION METHOD}\]

The mixup data augmentation method randomly selects images and combine the samples with weights sampled from a Beta distribution. The corresponding labels are combined with the same weights. Therefore, a generic vicinal distribution is proposed:

\[
\mu(x, y) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}[ \delta(x = \lambda \cdot x_i + (1 - \lambda) \cdot x_j, y = \lambda \cdot y_i + (1 - \lambda) \cdot y_j)],
\]

where \( \lambda \sim \text{Beta}(\alpha, \alpha) \), for \( \lambda \in [0, 1], \alpha \in (0, \infty) \). In our method, \((x_i, y_i)\) and \((x_j, y_j)\) are two samples drawn at random from the training data and the generated data respectively. \( x_i \) and \( x_j \) denote images while \( y_i \) and \( y_j \) denote the corresponding labels. The mixup hyper-parameter \( \alpha \) controls the strength of interpolation between image pairs or label pairs.

\[\text{C. ALGORITHM}\]

The procedure is formally presented in Algorithm 1.

After training, we use the discriminator \( D \) to extract features of training data and testing data. Then we use the extracted features and images’ labels to train a support vector machine (SVM), which is the classifier.

\[\text{IV. EXPERIMENTS AND ANALYSIS}\]

In this section, the datasets used in the experiments are first introduced in detail. And then, we analyze the effects of supervised loss, mixup and network structure respectively on two challenging high-resolution remote sensing image datasets. Finally, the experimental results achieved by the proposed method and the compared methods are showed and discussed.

\[\text{A. DATASETS}\]

Two different datasets are used for our experiments in this paper. The first one is the UC-Merced dataset [9]. UC-Merced dataset consists of 21 land-use classes. Each class consists of 100 aerial images measuring 256 \times 256 pixels. The images were manually extracted from large images from the USGS National Map Urban Area Imagery collection for various urban areas around the country. This dataset holds highly overlapping land-use classes such as dense residential, medium residential and sparse residential, which mainly differ in the density of structures and hence make the dataset richer and more challenging. The pixel resolution of this public domain imagery is 1 foot. Fig. 1 shows the example images of this dataset. In this paper, the training ratio of UC-Merced dataset is 80%, and the rest 20% is used as test set. Note that the training ratio is defined as the percentage of training data among the total dataset.

The second dataset is the NWPU-RESISC45 [1] dataset. This dataset consists of 45 land-use classes. Each class consists of 700 images measuring 256 \times 256 pixels. This scene classes above include the land use and land cover classes (e.g. commercial area, forest, industrial area, mountain, sparse residential), man-made object classes (e.g. airplane, airport, bridge, church) and natural landscape (e.g. island, cloud, beach, lake, river, sea ice). These classes contain a variety of
Algorithm 1 Training Process of the Model Proposed in This Paper

**Input:** set of training data \((X, Y)\)  
**Output:** Models: \(G, D\)  

In each training iteration:

1. Sample \(m\) positive examples \(X_r = \{x^1_r, x^2_r, \ldots, x^m_r\}\) and its corresponding labels \(Y_r = \{y^1, y^2, \ldots, y^m\}\) from real training dataset.
2. Sample \(m\) noise samples \(\{z^1, z^2, \ldots, z^m\}\) from a distribution \(P_z(z)\).
3. Obtain generated data \(X_g = \{x^1_g, x^2_g, \ldots, x^m_g\}\). Note that all the parameters of \(G\) and \(D\) have been randomly initialized before the training iterations.
4. Randomly sample real data \(x_r\) and generated data \(x_g\) to obtain generic vicinal distribution sample using mixup data augmentation method:
   
   \[
   \tilde{x}_{\text{mixup}} = \lambda x_r + (1 - \lambda) x_g \\
   \tilde{y}_{\text{mixup1}} = \lambda y_r + (1 - \lambda) y_g \\
   \tilde{y}_{\text{mixup2}} = \lambda y'_r + (1 - \lambda) y'_g
   \]

   where \(\lambda \sim \text{Beta} (\alpha, \lambda), \lambda \in [0, 1], \alpha \in (0, \infty)\), \(y_r\) is one and \(y_g\) is zero, which represent image’s real or fake property, \(y'_r\) is class label, \(y'_g\) is the predicted pseudo class label by discriminator. One-hot vector is used to code \(y'_r\) and \(y'_g\).
5. Update discriminator parameters \(\theta_d\) to maximize:
   
   \[
   \tilde{V}_d = \frac{1}{m} \sum_{i=1}^{m} \left( \log D_k + \log (1 - D_{k+1}(\tilde{x}^i_{\text{mixup}})) \right) \\
   + \frac{w_2}{m} \sum_{i=1}^{m} \sum_{j=1}^{k} (y'_j \log D_f(x^j_r)) \\
   + \frac{w_3}{m} \sum_{i=1}^{m} \left( \tilde{y}_{\text{mixup1}} \log D_{k+1}(\tilde{x}^i_{\text{mixup}}) \right) \\
   + (1 - \tilde{y}_{\text{mixup1}}) \log (1 - D_{k+1}(\tilde{x}^i_{\text{mixup}})) \\
   + \frac{w_4}{m} \sum_{i=1}^{m} \sum_{j=1}^{k} (\tilde{y}'_{\text{mixup2}} \log D_f(\tilde{x}^i_{\text{mixup}})).
   \]

   \[
   \theta_d \leftarrow \theta_d + \eta \nabla \tilde{V}(\theta_d), \\
   \theta_g \leftarrow \theta_g + \eta \nabla \tilde{V}(\theta_g),
   \]

   where \(w_2, w_3, w_4\) are weighting factors, and \(\eta\) is learning rate.
6. Sample \(m\) noise samples \(\{z^1, z^2, \ldots, z^m\}\) from a distribution \(P_z(z)\).
7. Update generator parameters \(\theta_g\) to minimize:
   
   \[
   \tilde{V}_g = \frac{1}{m} \sum_{i=1}^{m} \log(1 - D(G(z^i))) + w_1 \left\| f(x^i) - f(G(z^i)) \right\|^2,
   \]

spatial patterns, some homogeneous with respect to texture, some homogeneous with respect to color, others not homogeneous at all. The spatial resolution varies from about 30 m to 0.2 m per pixel for most of the scene classes except for the classes of island, lake, mountain, and snowberg that have lower spatial resolutions. Fig. 2 shows the example images of this dataset.

B. EXPERIMENTAL SETUP

We carried out experiments on both UC-Merced and NWPU-RESISC45 datasets with a 5-fold cross-validation protocol using a linear SVM as classifier. All the results are provided with the mean and standard deviation of these five times of training and testing. We trained our method using a modified architecture from MARTA-GAN’s [2]. The network architecture is shown in Fig. 3. The input of the generator is a 100-dimensional random noise. Then, the result is reshaped into a four-dimensional tensor. We use deconvolutional layers in generator to learn its own spatial upsampling and output 256 × 256 images. The input of discriminator is images, include real images and generated images. We use convolutional layers to learn its own spatial downsampling. After downsampling, we use the multi-feature layer to obtain features which will be sent into a fully connected layer. Then, the output of the discriminator is a batch of \(k + 1\)-dimensional vector. The 1st to \(k\)th dimensions denote the predicted scene class labels of the input images, while the \(k + 1\)th dimension predicts an image is real or fake.

The models were trained with a batch size of 64, and we used Adam optimizer with a learning rate of 0.0002 and a momentum term \(\beta_1\) of 0.5. When using mixup data augmentation, the parameter \(\alpha\) was set to 1. All the weighting factors were set as 1. Before training, we adopted traditional data augmentation in the dataset via flipping images horizontally, vertically and rotating the by 90, 180, 270 degrees to increase the effective training set size. During the test phase, the test images are sent to the trained discriminator to extract features, then the features are sent to the trained SVM to make class prediction.

C. EFFECTIVENESS OF SUPERVISED LOSS

We evaluated the addition of a supervised loss, which is expressed in equation (7), to the total loss of discriminator. Table 1 shows that the addition of supervised loss increases the OA (overall accuracy) by 1% on UC-Merced dataset, and by 5.92% on NWPU-RESISC45 dataset.

| Loss Type                  | UC-Merced 80% training ratio | NWPU-RESISC45 20% training ratio |
|----------------------------|------------------------------|----------------------------------|
| Unsupervised loss          | 92.05 ± 0.60                | 72.56 ± 0.27                     |
| Supervised loss – unsupervised loss | 93.05 ± 0.81                | 78.48 ± 0.26                     |

TABLE 1. Overall accuracy (%) under different losses.
D. EFFECTIVENESS OF MIXUP LOSS
The mixup can be thought as a kind of data augmentation. There are two types of mixup shown in step 4 and 5 of Algorithm 1. The functions of mixup1 and the sum of mixup1 and mixup2 are evaluated through experiments. The mixup1 contributes to unsupervised learning, and mixup2 contributes to supervised learning. It is shown in Table 2 that adding mixup1 loss has slightly effect to the OA, while adding mixup1 together with mixup2 increases the OA by 1% on UC-Merced dataset, and by 1.25% on NWPU-RESISC45 dataset. The mixup2 data augmentation improves performance more significant than mixup1 means that the supervised learning data is relatively inadequate compared with unsupervised learning.

| Network structure | UC-Merced 80% training ratio | NWPU-RESISC45 20% training ratio |
|-------------------|-----------------------------|----------------------------------|
| Original network (Fig. 2) | 94.05±1.2 | 79.73±0.35 |
| Residual block structure | 93.90±0.60 | 83.12±0.26 |

E. EFFECTIVENESS OF BACKBONE NETWORK UPGRADE
To further improve the feature extraction ability of discriminator, we adopted residual learning block [27] to replace each of the original six convolutional layers. The residual block structure is shown in Fig. 4. Then we conducted experiments using the proposed algorithm 1 with upgraded backbone network. The results are displayed in Table 3. A 3.49% increase is achieved just by adopting a residual block structure. However, it does not improve the OA on UC-Merced dataset. We think the reason is that larger network usually need larger dataset to release its potential. On both datasets, adopting the residual structure decreases the standard deviation of OA when applying the 5-fold cross-validation protocol. The initial results show that it is promising to enhance performance by improving the backbone network structure further.

F. EXPERIMENTS ON BOTH DATASETS
Fig. 5 depicts the confusion matrix of classification results for our method. MARTA-GAN [2] and our method reach an overall accuracy of 92.05% and 94.05% respectively. Our method is 2% better because our model combines mixup and supervised learning. From the results of previous subsections, introduction of supervised loss contributes 1%, while mixup data augmentation contributes another 1%. The discriminator of MARTA-GAN use the multi-feature layer to merge the mid-level and global features. Based on MARTA-GAN, the discriminator of our model learn features of each class, thus learn more discriminative features.

The loss curve of the discriminator with mixup data augmentation is shown in Fig. 6 and the loss curve of the discriminator without mixup data augmentation is shown in Fig. 7. As we can see, the mixup data augmentation can stabilize...
FIGURE 2. Example images of NWPU-RESISC45 dataset: (1) airplane; (2) airport; (3) baseball diamond; (4) basketball court; (5) beach; (6) bridge; (7) chaparral; (8) church; (9) circular farmland; (10) cloud; (11) commercial area; (12) dense residential; (13) desert; (14) forest; (15) freeway; (16) golf course; (17) ground track field; (18) harbor; (19) industrial area; (20) intersection; (21) island; (22) lake; (23) meadow; (24) medium residential; (25) mobile home park; (26) mountain; (27) overpass; (28) palace; (29) parking lot; (30) railway; (31) railway station; (32) rectangular farmland; (33) river; (34) roundabout; (35) runway; (36) sea ice; (37) ship; (38) snowberg; (39) sparse residential; (40) stadium; (41) storage tank; (42) tennis court; (43) terrace; (44) thermal power station; (45) wetland.

FIGURE 3. Network architectures of generator and discriminator.

training process and make the loss curve of discriminator tend to convergence. Mixup data augmentation randomly selects images and combine them with weights sampled from a Beta distribution in each iteration. The weights are different in each iteration and each batch while the augmented images are also different. In this way, this method augments the data and thus can help to smooth the training process. Fig. 8 shows some exemplary images synthesized by the generator trained on both datasets.

From the experimental results, the overall accuracy of our model trained on NWPU-RESISC45 dataset reached 92.78% under the training ratios of 80% and 83.12% under the
training ratios of 20%. Table 4 shows the results obtained from references and proposed method. In general, the pre-trained models have stronger ability of feature extraction and obtain better classification accuracy. On UC-Merced dataset, our method achieves comparable performance among all models.

On NWPU-RESISC45 dataset, the proposed method achieves the state-of-the-art performance among no pre-training models.
V. CONCLUSION
Remote sensing image scene classification plays an important role in a wide range of applications, such as natural disaster detection, land resource utilization and coverage management etc. The generative adversarial network is the most potential deep learning method in recent years, and it is a new idea to introduce it into remote sensing image classification. This paper proposed a classification model based on GAN and supervised learning, then it can learn more discriminative features from labeled data and unlabeled data. Extensive experiments show that the introduction of supervised loss enhances the feature extracting ability. Meanwhile, mixing up generated fake images and real images can be thought as a type of data augmentation. Compared to previous approaches which are supervised or unsupervised, our method builds a more generalized classification model for remote sensing image scene classification and reaches the state-of-the-art performance among those without pre-training on ImageNet.

In this paper, it has been proved that the enhancement of the GAN backbone network structure can improve performance significantly especially for large dataset. In the future, it is considerable to enhance the network structure further. In addition, the proposed method in this paper generates images without labels, and the categories of the images are uncertain. Therefore, the improvement of feature extraction ability of each category is uniform. For some categories, the classification accuracy is almost 100%, while the results of others need to be improved. It is worth studying on generating images directionally to augment data and improve the classification results for those categories whose classification results need to be enhanced.

REFERENCES
[1] G. Cheng, J. Han, and X. Lu, “Remote sensing image scene classification: Benchmark and state of the art,” Proc. IEEE, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.
[2] D. Lin, K. Fu, Y. Wang, G. Xu, and X. Sun, “MARTA GANs: Unsupervised representation learning for remote sensing image classification,” IEEE Geosci. Remote Sens. Lett., vol. 14, no. 11, pp. 2092–2096, Nov. 2017.
[3] K. Nogueira, O. A. B. Penatti, and J. A. dos Santos, “Towards better exploiting convolutional neural networks for remote sensing scene classification,” 2016, arXiv:1602.01517. [Online]. Available: http://arxiv.org/abs/1602.01517
[4] W. Zhang, P. Tang, and L. Zhao, “Remote sensing image scene classification using CNN-CapsNet,” Remote Sens., vol. 11, no. 5, p. 494, 2019.
[5] 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.
[6] O. A. B. Penatti, K. Nogueira, and J. A. dos Santos, “Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2015, pp. 44–51.
[7] F. Zhang, B. Du, L. Zhang, and M. Xu, “Weakly supervised learning based on coupled convolutional neural networks for aircraft detection,” IEEE Trans. Geosci. Remote Sens., vol. 54, no. 9, pp. 5553–5563, Sep. 2016.
[8] A. Romero, C. Gatta, and G. Camps-Valls, “Unsupervised deep feature extraction for remote sensing image classification,” IEEE Trans. Geosci. Remote Sens., vol. 54, no. 3, pp. 1349–1362, Mar. 2016.
[9] Y. Yang and S. Newsam, “Bag-of-visual-words and spatial extensions for land-use classification,” in Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst. (GIS), 2010, pp. 270–279.
[10] G. Sheng, W. Yang, T. Xu, and H. Sun, “High-resolution satellite scene classification using a sparse coding based multiple feature combination,” Int. J. Remote Sens., vol. 33, no. 8, pp. 2395–2412, Apr. 2012.
[11] X. Lu, X. Zheng, and Y. Yuan, “Remote sensing scene classification by unsupervised representation learning,” IEEE Trans. Geosci. Remote Sens., vol. 55, no. 9, pp. 5148–5157, Sep. 2017.
[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 2672–2680.
[13] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," 2013, arXiv:1311.3619. [Online]. Available: http://arxiv.org/abs/1311.3619

[14] 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

[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] I. Goodfellow. (2015).

[17] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," 2018, arXiv:1710.09412. [Online]. Available: http://arxiv.org/abs/1710.09412

[18] T. Salimans, I. Goodfellow, and W. Zaremba, "Improved techniques for generative adversarial networks," 2015, arXiv:1511.06434. [Online]. Available: http://arxiv.org/abs/1511.06434

[19] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis." 2018, arXiv:1809.11096. [Online]. Available: http://arxiv.org/abs/1809.11096

[20] H. Salehinejad, S. Valaee, T. Dowdell, E. Colak, and J. Barfett, "‘‘Gen–A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis.” 2018, arXiv:1809.11096. [Online]. Available: http://arxiv.org/abs/1809.11096

[21] Y. Tao, M. Xu, Y. Zhong, and Y. Cheng, "GAN-assisted two-stream neural network for high-resolution remote sensing image classification," Remote Sens., vol. 9, no. 12, p. 1328, 2017.

[22] J. Tobias Springenberg, "Unsupervised and semi-supervised learning with categorical generative adversarial networks," 2015, arXiv:1511.06390. [Online]. Available: http://arxiv.org/abs/1511.06390

[23] Z. Zhou, Y. Zheng, and H. Ye, “Satellite image scene classification via ConvNet with context aggregation,” in Proc. Pacific Rim Conf. Multimedia, 2018, pp. 329–339.

[24] C. Nader Vasconcelos and B. Nader Vasconcelos, “Convolutional neural network committees for melanoma classification with class-cal and expert knowledge based image transforms data augmentation,” 2017, arXiv:1702.07025. [Online]. Available: http://arxiv.org/abs/1702.07025

[25] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “Mixup: Beyond empirical risk minimization,” 2017, arXiv:1710.09412. [Online]. Available: http://arxiv.org/abs/1710.09412

[26] 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.

[27] G. Cheng, Z. Li, X. Yao, L. Guo, and Z. Wei, “Remote sensing image scene classification using bag of convolutional features,” IEEE Geosci. Remote Sens. Lett., vol. 14, no. 10, pp. 1735–1739, Oct. 2017.

[28] A. Goel, B. Banerjee, and A. Pizurica, "Hierarchical metric learning for optical remote sensing scene categorization," IEEE Geosci. Remote Sens. Lett., vol. 16, no. 6, pp. 952–956, Jun. 2019.

PEIYAO YAN was born in Wuhan, China, in 1994. She received the B.S. and M.S. degrees in electronic engineering from the Huazhong University of Science and Technology, Wuhan, in 2016 and 2019, respectively. She is currently employed with Wuhan Nari Limited Liability Company of State Grid Electric Power Research Institute. Her current research interests include remote sensing, image classification, and deep learning.

FENG HE received the B.S. degree in electronic engineering from Jiangxi Science and Technology Normal University, Nanchang, China, in 1998, and the M.S. and Ph.D. degrees in communication and systems from the Huazhong University of Science and Technology, Wuhan, China, in 2002 and 2009, respectively. From 2002 to 2005, he was a Software Engineer with the FiberHome Technologies Group, Wuhan, and the UT Starcom Shenzhen Research and Development Center, Shenzhen, China. Since 2005, he has been a Lecturer, and then an Associate Professor with the School of Electronic Information and Communications, Huazhong University of Science and Technology, since 2016. His current research interests include remote sensing image interpretation, deep learning, and the imaging of microwave interferometric radiometers.

YAJIE YANG received the B.E. degree from the Nanjing University of Information Science and Technology, Nanjing, China, in 2018. She is currently pursuing the master’s degree in electromagnetic and microwave technology with the Huazhong University of Science and Technology. Her research interests include remote sensing image scene classification and deep learning.

FEI HU received the M.S. degree in experimental mechanics and the Ph.D. degree in communication and information systems from the Huazhong University of Science and Technology, Wuhan, China, in 1998 and 2002, respectively. He is currently a Professor with the School of Electronic Information and Communications, Huazhong University of Science and Technology. He also served as a Senior Researcher and then the Deputy Director of the National Key Laboratory of Science and Technology on Multi-Spectral Information Processing Technologies. He has published more than 50 technical papers and reports in his research areas. His research interests include microwave technique, microwave interferometric radiometers, microwave remote sensing, and passive microwave imaging. * * *