International Journal of Computational Intelligence Systems (2022) 15:87
https://doi.org/10.1007/s44196-022-00150-0

RESEARCH ARTICLE

Semi-supervised Remote Sensing Image Scene Classification Based on Generative Adversarial Networks

Dongen Guo[1] · **Zechen Wu**[1] · **Yuanzheng Zhang**[1] · **Zhen Shen**[1]

Received: 8 July 2022 / Accepted: 12 October 2022

Abstract
With the availability of numerous high-resolution remote sensing images, remote sensing image scene classification has been widely used in various fields. Compared with the field of natural images, the insufficient number of labeled remote sensing images limits the performance of supervised scene classification, while unsupervised methods are difficult to meet the practical applications. Therefore, this paper proposes a semi-supervised remote sensing image scene classification method using generative adversarial networks. The proposed method introduces dense residual block, pre-trained Inception V3 networks, gating unit, pyramidal convolution, and spectral normalization into GANs to promote the semi-supervised classification performance. To be specific, the pre-trained Inception V3 network is introduced to extract semantic features to enhance the feature discriminant capability. The gating unit is utilized to capture the relationships among features. The pyramidal convolution is integrated into dense residual block to capture different levels of details to strengthen the feature representation capability. The spectral normalization is introduced to stabilize the GANs training to improve semi-supervised classification accuracy. Extensive experimental results on publicly available EuroSAT and UC Merced datasets show that the proposed method gains the highest overall accuracy, especially when only a few labeled samples are available.

Keywords Remote sensing image scene classification · Generative adversarial networks (GANs) · Semi-supervised learning · Gating unit · Pyramidal convolution · Spectral normalization

Abbreviations

GANs	Generative adversarial networks
FMGAN	Feature-matching GANs
tripleGAN	Triple GANs
BADGAN	BAD GANs
REGGAN	Regularization GANs
SFGAN	Semantic fusion GANs
PyConv	Pyramidal convolution
SN	Spectral normalization
SSGAN	Semi-supervised GANs
GU	Gating units
G	Generative network
D	Discriminative network
CatGAN	Categorical GANs
CNNs	Convolutional neural networks
BN	Batch normalization
PReLU	Parametric ReLU
GAP	Global average pooling
GSD	Ground sampling distances
OA	Overall accuracy
CM	Confusion matrix

✉ Dongen Guo
3161010@nyist.edu.cn

Zechen Wu
2672818006@qq.com

Yuanzheng Zhang
758949617@qq.com

Zhen Shen
652521268@qq.com

[1] School of Computer and Software, Nanyang Institute of Technology, 80 Changjiang Road, Nanyang 473004, Henan, China

1 Introduction

Scene classification of remote sensing image can automatically classify scene images into specified semantic categories based on their contents [1]. Currently, supervised methods based on deep neural networks are the mainstream, which usually rely on larger scale labeled samples to obtain higher classification accuracy [2]. However, labeling remote sensing images is often costly. Unsupervised scene

Published online: 20 October 2022

classification methods can learn directly from a large number of unlabeled samples, but their classification accuracy is difficult to meet practical applications because they cannot make full use of labels [3]. Semi-supervised scene classification is a combination of supervised and unsupervised, which can learn from a small number of labeled samples and a large number of unlabeled samples to obtain satisfactory classification accuracy [4].

In recent years, generative adversarial networks (GANs) [5] are introduced into semi-supervised image classification. GANs can learn the underlying data distribution from real training samples to compete with state-of-the-art semi-supervised image classification methods [6, 7]. Salimans et al. [8] proposed feature-matching GANs (FMGAN) by extending standard classifier. Li et al. [9] proposed a triple GANS (tripleGAN) to achieve excellent semi-supervised classification performance using a three-player's game. Dai et al. [10] proposed a new GANs-based semi-supervised classification model (BADGAN) to effectively improve the classification performance. Lecouat et al. [11] leveraged GANs with manifold regularization (REGGAN) for semi-supervised image classification. These GAN-based methods have achieved better semi-supervised classification performance using variants of the standard DCGAN [12] on CIFAR10 and SVHN datasets. GAN-based methods increase the number of training samples using generated samples for better classification performance. However, for complex remote sensing scene images, it is difficult for the discriminative network of standard DCGAN to extract more discriminative features, which affects the performance of semi-supervised classification. Therefore, it is a crucial challenge to further investigate more discriminative feature extraction to improve the semi-supervised classification performance.

Although some success has been achieved in the classification of low-resolution images, the GANs-based remote sensing scene classification still needs to be improved. More recently, Roy et al. [13] introduced a semantic branch into GANs (SFGAN) for semi-supervised satellite image classification to obtain better classification performance. However, SFGAN use a standard DCGAN structure, which is suitable for processing images with relatively simple scenes and low spatial resolution, the classification performance is more limited for high-resolution remote sensing images with complex scenes. Inspired by SFGAN, we further investigate methods to extract more discriminative features through discriminative network. Guo et al. [4] proposed a GAN-based semi-supervised remote sensing scene classification method, they introduced a gating unit and a self-attention gating (SAG) module into the discriminative network to improve semisupervised classification performance (SAGGAN). Ledig et al. [14] used GANs with dense residual block for super-resolution reconstruction of natural images to achieve more realistic image texture structure. Miech et al. [15] introduced a learnable nonlinear unit (named context gating) that aims to model the interdependencies between network activations for video classification. Liu et al. [16] proposed gated full convolutional blocks to improve micro video scene classification performance. Duta et al. [17] proposed pyramidal convolution (PyConv) with different sizes, types and depths of filters to extract different levels of details in scene image. Miyato et al. [18] introduced spectral normalization (SN) into GANs to stabilize the training process for improving the performance of GANs.

Inspired by the above works, we propose a novel semi-supervised remote sensing scene classification model based on GANs (SSGAN), which uses gating units (GU), PyConv, pre-trained network branch, SN and dense residual block to enhance the feature extraction capability of discriminative network. GU is dedicated to capturing the dependencies between features and adaptively focusing on the important features of the input image; PyConv is dedicated to capturing the detailed features at different levels in scene images; the dense residual blocks are used to replace the convolution in GANs, improving the quality of the generated images and enhancing the feature discrimination of the discriminative network; and SN aims at stabilizing the training process of GANs to improve the performance of GANs. Compared with SAGGAN [4], in the proposed SSGAN, we introduced the dense residual blocks to replace the convolution in SAGGAN to enhance feature discrimination ability, integrated the PyConv into residual blocks to capture feature details of different levels, and introduced SN into both discriminative network and generative network to promote the performance of GAN. Extensive experiments on EuroSAT [19] and UCM Merced [20] datasets show that the proposed SSGAN achieves higher classification accuracy than other state-of-the-art semi-supervised methods based on GANs. The main contributions of this paper are summarized as follows.

1. A novel GAN-based semi-supervised SSGAN model is proposed, by enhancing the feature exacting ability of discriminative network for improving semi-supervised scene classification accuracy.
2. GU, dense residual block and PyConv are introduced into the discriminative network to adaptively focus on important features and capture the details at different levels for achieving more discriminative feature representation.
3. SN is integrated into both generative and discriminative network to stabilize the training of GANs for improving classification accuracy.
4. Dense residual block is introduced the generative network to enhance the quality of the generated images for augmenting the training samples.

The remainder of this paper is organized as follows. In Sect. 2, the related works of this paper is introduced. In Sect. 3, the architecture of SSGAN is presented. In Sect. 4, the experimental results and discussions are shown. Finally, the conclusions are drawn in Sect. 5.

2 Related Works

In this section, we review related works on gating mechanisms, PyConv, semi-supervised image classification based on GANs, and so on.

2.1 GANs-Based Semi-supervised Image Classification

GANs are widely used in semi-supervised classification in recent years because of their powerful generation ability to effectively augment training samples to improve classification performance. However, the traditional GANs-based semi-supervised classification methods use the standard DCGAN structure, which affects the feature discriminant ability of the discriminative network to hinder the performance of semi-supervised classification. Therefore, we improve the standard DCGAN to further enhance the discriminant ability of the discriminative network to improve the semi-supervised classification performance. The principle of GANs-based semi-supervised image classification is detailed below.

The standard GANs have two components, the discriminative network D and generative network G. G synthesizes fake samples $G(z)$ by random noise z, D distinguishes between real and fake samples [5], and GANs accomplish the related task by a min–max game between G and D. The value function $V(G, D)$ can be expressed as follows:

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

where z is a random noise vector that is generated following a priori distribution ($z \sim p_z(z)$), p_{data} denotes real data distribution, $G(z)$ is image generated from G, $D(x)$ represents class probability that x is from real sample, and $D(G(z))$ denotes the probability that the sample is generated by G. The goal of D is to maximize the probability of the real sample, and the goal of G is to increase the probability of the generated sample being classified as a real image.

Springenberg et al. [21] proposed categorical generative adversarial networks (CatGAN) for semi-supervised image classification by using a multi-classifier to substitute binary classifier. Salimans et al. [8] further extended standard classifier. They add images generated from G as a new category $y = K + 1$, then the output dimension of D becomes logits $= \{l_1, l_2, \ldots, l_{K+1}\}$. These logits are converted into the class probabilities using softmax function. Then, the probability that x is a real sample of the j-th category is as follows:

$$p_{\text{model}}(y = j|x, j < K + 1) = \frac{\exp(l_j)}{\sum_{K=1}^{K+1}(\exp(l_k))}, \tag{2}$$

and the probability that x is fake sample is as follows:

$$p_{\text{model}}(y = K + 1|x) = \frac{\exp(l_{K+1})}{\sum_{K=1}^{K+1}(\exp(l_k))}, \tag{3}$$

For the semi-supervised classification model, labeled samples are trained in a supervised manner, while unlabeled samples are trained in an unsupervised manner. The network framework for GANs-based semi-supervised classification is shown in Fig. 1. One can observe from Fig. 1 that the inputs of D consist of real labeled samples, real unlabeled samples and samples generated by G. Therefore, the loss object of D is as follows:

$$\begin{aligned} L_D &= -E_{x,y\sim p_{\text{data}}(x,y)}[\log(p_{\text{model}}(y|x, y < K + 1))] \\ &\quad - \{E_{x\sim p_{\text{data}}(x)} \log([1 - p_{\text{model}}(y = K + 1|x))] \\ &\quad + E_{x\sim G}[\log(p_{\text{model}}(y = K + 1|x))]\} \\ &= L_{\text{s}} + L_{\text{un}}, \end{aligned} \tag{4}$$

where L_{s} presents the supervised loss, L_{un} denotes the unsupervised loss. The first term in L_{un} indicates the loss of real unlabeled sample, and the second term is the loss of fake sample. For unsupervised learning, D only outputs

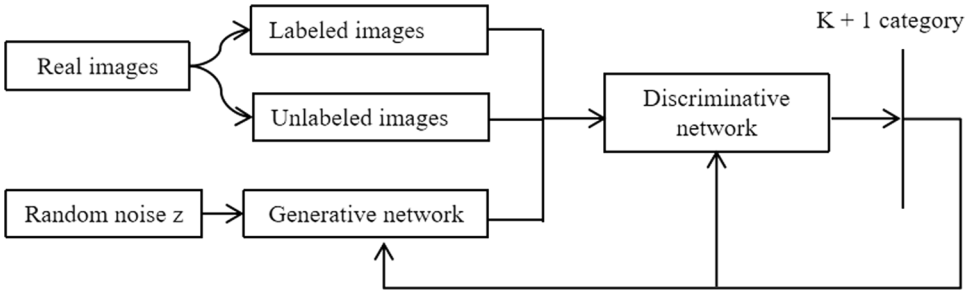

Fig. 1 The network framework of semi-supervised classification using GANs

true or false without distinguishing categories, so it can be expressed by Eq. (5).

$$D(x) = 1 - p_{\text{model}}(y = K + 1|x), \tag{5}$$

Substituting Eq. (5) into L_{un}, we can obtain Eq. (6) as follows.

$$L_{\text{un}} = -\{E_{x\sim p_{\text{data}}} \log D(x) + E_{z\sim p_z(z)} \log(1 - D(G(z)))\}, \tag{6}$$

The loss L_G of G can be expressed as follows.

$$\begin{aligned} L_G = {} & E_{x\sim p_{\text{data}}(x)} f(x) - E_{z\sim G} f(\hat{x})_2^2 \\ & - E_{x\sim G} \log[1 - p_{\text{model}}(y = K + 1|x)], \end{aligned} \tag{7}$$

where $f(x)$ is the activation of middle layer from D to match the features between real samples and generated samples, and $\hat{x}$ denotes generated samples. The first term in Eq. (9) denotes the feature matching term, which drives G to generate an sample that matches the manifold of real samples, so that D can better distinguish real sample from sample generated by G.

More recently, Lecouat et al. [11] leveraged GANs with manifold regularization for semi-supervised image classification. Li et al. [9] proposed a Triple-GAN, which included G, D, and a separate classifier C to simultaneously achieve superior classification performance and a good image generation. Dai et al. [10] analyzed why good semi-supervised classification performance and good generator cannot be obtained at the same time. They proposed a BAD-GAN based on their analysis to improve classification performance on multiple benchmark datasets.

Traditional GANs-based semi-supervised classification methods use the standard DCGAN structure, which limits the scene classification performance of remote sensing images with complex scenes. Ledig et al. [14] used GANs with the dense residual structure for super-resolution reconstruction of natural images to achieve the more realistic image texture structure. Inspired by [14], we replace standard convolutional structure in the SFGAN with dense residual block.

2.2 Gating Mechanism

More recently, Srivastava et al. [22] leveraged adaptive gating units to train deep neural networks. Miech et al. [15] proposed a context gating unit to aim at capturing interdependencies among network activations for improving video classification performance. Liu et al. [16] introduced the gated fully convolutional blocks to improve micro-video venue classification performance. Guo et al. [4] proposed a self-attention gating module by combining a gating unit and a self-attention block to capture the long-range dependencies among feature maps to focus on crucial regions adaptively. Their experiments demonstrate that the GU can focus on important areas in the scene to eliminate the background effectively improving feature discrimination. Inspired by above mentioned works, we introduce the combination of gating units and residual blocks into GANs to further enhance feature discriminant ability for improving classification performance.

2.3 Pyramidal Convolution

Convolutional neural networks (CNNs) have become the core architecture for current computer vision applications. The core of CNNs are convolutional layers, which are used for visual recognition by learning spatial kernels. Typically, most CNNs utilize relatively smaller kernel sizes (e.g., 3×3) which can greatly reduce the number of parameters and computational complexity. However, smaller kernels limit receptive field of CNNs, which lost useful details to affect the performance of visual tasks. To address this issue, Yu et al. [23] used dilation convolution to aggregate multi-scale contextual information to effectively improve the accuracy of semantic segmentation. In addition, Zhao et al. [24] used a pyramid pooling module to interpret scenes to extract different levels of details. Dilated convolutions with irregular spatial pyramidal pooling are introduced into the literature [25] to encode global context using image-level features for improving semantic segmentation performance. However, these are additional blocks that need to be embedded in the CNNs, which remarkably increase model parameters and computational complexity. Duta et al. [17] introduced PyConv to process the input samples at multiple-scale filters. PyConv consists of a pyramidal kernel in which each layer contains of different types, sizes and depths of filters to capture different levels of details for enhancing feature discriminant ability. Recently, Guo et al. [3] introduce PyConv into each residual block of the discriminative network to capture the different levels of details from multiple-scale filters for enhancing the features discriminant ability. Their experiments show that PyConv is able to capture different levels of detailed features to effectively improve feature discrimination. Inspired by the above works, we replace the middle layer convolution in residual block of discriminative network with PyConv to capture more details for further enhancing the feature discriminant capability of discriminative networks.

3 Proposed Method

The proposed SSGAN is described in detail below.

3.1 Structure of Proposed SSGAN

As described in Sect. 2, the GAN-based semi-supervised classification model has the similar structure to the original GANs [5]. Following the structures in SRGAN [14], the dense residual block in SRGAN is used to replace the standard convolution in SFGAN construct the generative and discriminative network in SSGAN. Figure 2 illustrates the network framework of the proposed SSGAN. The generative network is composed of four residual blocks. The residual block is shown as the G_Block block in Fig. 2, which includes an upsampling layer, a batch normalization (BN), a parametric ReLU activation (PReLU), and two convolutional layers separated by a BN. In addition, a GU is added after the first residual block. The discriminative network contains five residual blocks and one global average pooling layer (GAP).

To improve the GANs-based semi-supervised classification performance, inspired by SFGAN [13], we extend the original discriminative network. First, a pre-trained Inception V3 network is introduced as a new branch to extend the discriminative network, which can extract semantic features by fine-tuning, and then a GAP operation is performed on the extracted feature maps. Second, the second convolutional layer of residual block in D is superseded by the PyConv to capture the longer range of contextual information. In addition, a GU is added to the discriminative network to adaptively focus on the crucial regions and filter the useless background. Specifically, the GU is added after the first residual block and GAP in the original discriminative network, and the GU is placed after the GAP in the Inception V3 branch. Finally, the two feature vectors from the Inception V3 and original discriminative network are concatenated and fed into the softmax function for semi-supervised scene classification.

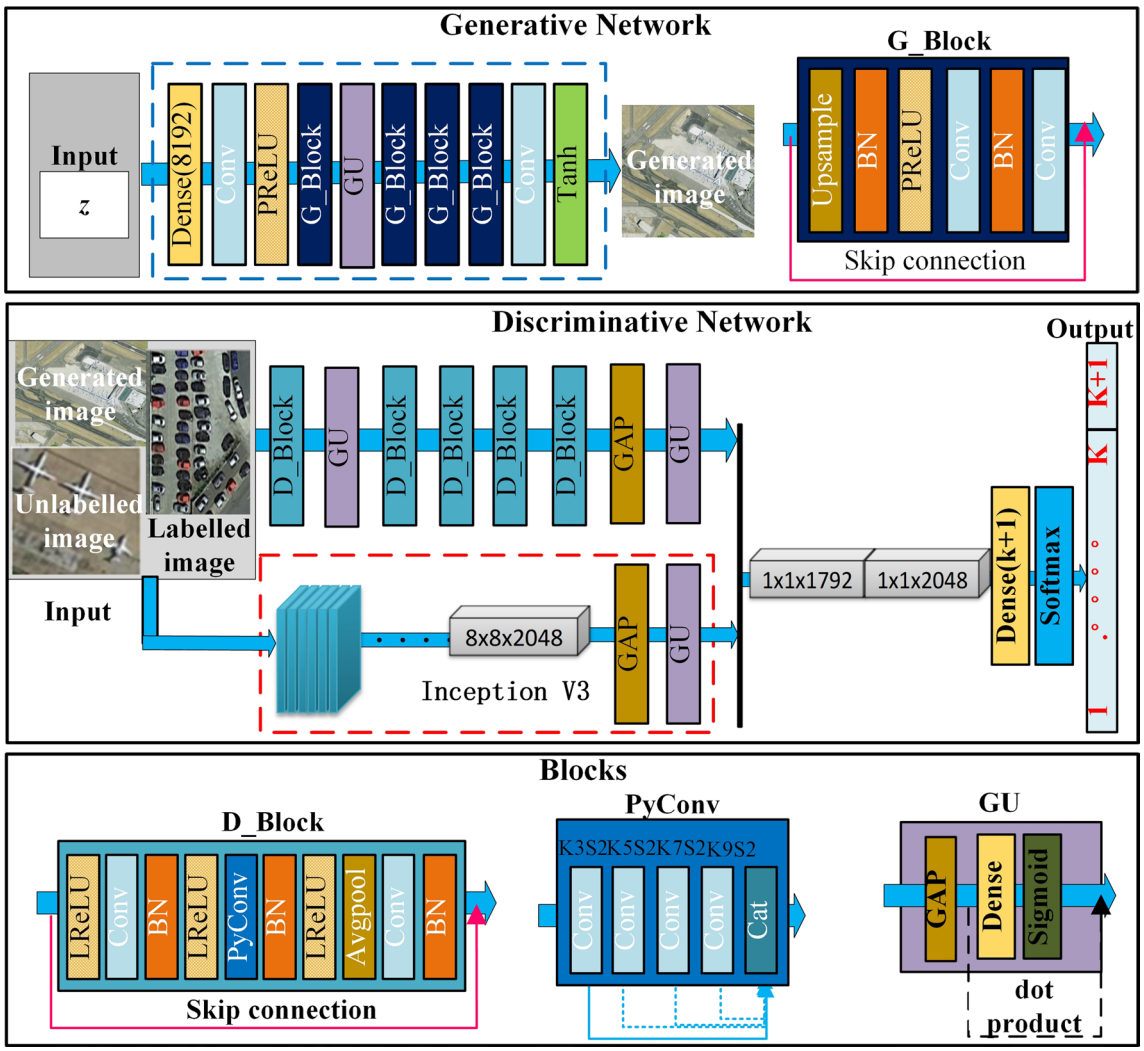

Fig. 2 The framework of the proposed SSGAN

In addition, the instability in the training of GANs is a major factor affecting performance. Miyato et al. [18] stabilized the training of GANs using SN to constrain Lipschitz constant of the discriminative network to satisfy 1-Lipschitz continuity. Zhang et al. [26] demonstrated that SN can achieve better performance when it is introduced into both generative and discriminative network. Inspired by [18, 26], we use spectral norm to achieve 1-Lipschitz continuity in residual blocks of both generative and discriminative network simultaneously to ensure the stability of SSGAN training.

3.2 Structure of Gating Unit

Most GANs-based semi-supervised methods use standard convolution to construct discriminative network. The convolution operation is limited by receptive field, and it is difficult to capture the dependencies among feature maps, which affects the semi-supervised classification performance.

Inspired by gating mechanism [16], the GU is designed and introduced after the first residual block of original discriminative network and after GAP of both branches to enhance the feature description capability. The Fig. 3 illustrates the structure of GU. The derivation of GU is briefly described below.

The input of GU can be any intermediate layer feature map F from the discriminative network, and GU can convert F into new feature F_{GU}. The derivation process is as follows.

$$f_{GU}(F) = \sigma(fc(F)), \tag{8}$$

$$F_{GU} = f_{GU}(F) \odot F, \tag{9}$$

where $\sigma(\cdot)$ presents sigmoid function, $f_{GU}(\cdot)$ denotes gating function, and after the sigmoid operation, the result is a weight matrix with values in the range [0,1]. $fc(\cdot)$ represents the fully connection, f_{GU} is the output of gating unit, and $\odot$ denotes the dot product operation. The gating unit can effectively extract the dependencies between feature maps, eliminate irrelevant background, and enhance feature representation.

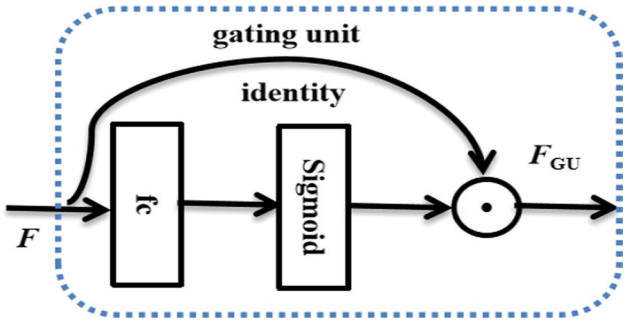

Fig. 3 The structure of GU

3.3 Structure of Pyramidal Convolution

Owing to the complexity of remote sensing scene images, different ground objects present different sizes in different scenes, and even the same ground objects in the same scene may display different sizes, so it is difficult to capture the diversity effectively using the traditional 3×3 convolution. Duta et al. [17] proposed the PyConv, which introduces pyramidal kernels with different filters to extract different levels of details. PyConv with n groups of different kernels can be represented as shown in Fig. 4. The residual block D_Block of discriminative network is constructed following PyConv in this paper, and the structure is shown as D_block

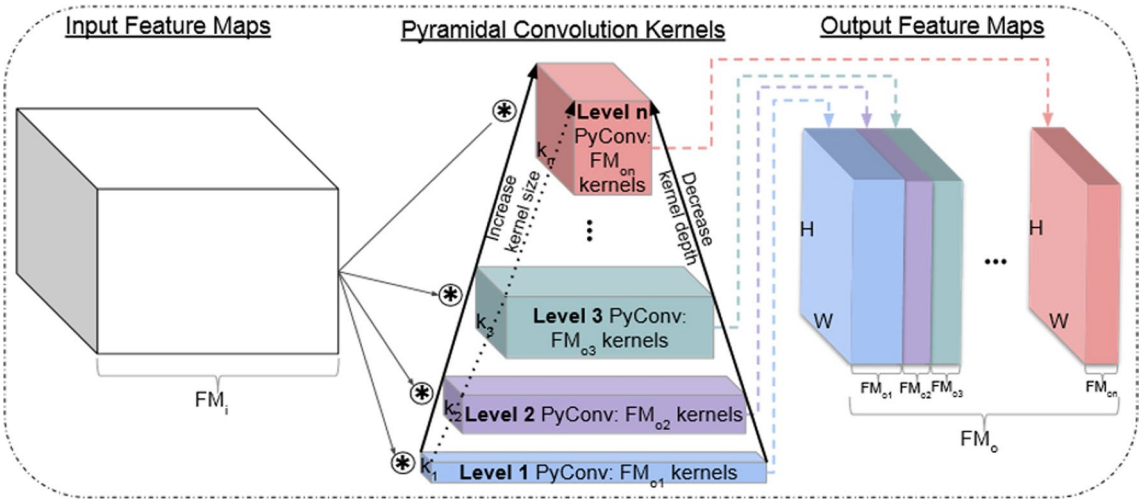

Fig. 4 The structure of PyConv [17]

in the bottom leftmost corner of Fig. 2, where the second convolution is replaced with PyConv. In this paper, we refer to the structure of PyConvHGResNet and use four groups of PyConv with different filter sizes to capture different levels of details.

where FMi denotes the feature map of the middle layer as the input of PyConv and FMo denotes the output of PyConv. PyConv consists of four groups of filters with different sizes and depths, and the residual blocks with PyConv can capture different levels of feature details to enhance feature representation ability in discriminative network.

4 Experimental Results and Analysis

In this section, comprehensive experiments are conducted on the EuroSAT and UCM Merced benchmark datasets to validate the effectiveness of the proposed SSGAN method.

4.1 DataSet Description

EuroSAT dataset: EuroSAT [19] is a recently released remote sensing image dataset acquired by Sentinel-2 satellite, which includes 10 different categories. In total, the dataset consists of 27,000 images with 64×64 pixels, which ground sampling distances (GSD) ranges from 10 to 60 m.

UC Merced dataset: The UC Merced [20] includes 21 different categories and has become benchmark dataset for remote sensing images classification. Each category has 100 images, and all images are 256×256 pixels in size.

To validate the semi-supervised classification performance of SSGAN, the EuroSAT dataset was split into three pieces as suggested in [13, 19]: the training samples are 80% and the rest are further divided into 90% for testing and 10% for validation, i.e., 216,00 samples for training set and the rest 5,400 samples are further divided into 4860 for testing and 540 for validation. UC Merced dataset was divided in same ratio. The number ($M = X_l$) of tagged training samples is set in accordance with SAGGAN [5]. More specifically, the M is set to 100, 1000, 2000, and 21,600 at EuroSAT dataset, the M was set to 100, 200, 400, and 1680 at UC Merced dataset, and the rest are treated as unlabeled samples (X_u).

4.2 Experimental Setup and Evaluation Metrics

All experiments were performed in PyTorch framework on a 64-bit Ubuntu 16.04 server with an 8-core Intel Gold 6048 CPU and four TITAN V GPUs. During training, the parameters are set following SFGAN [13] and SRGAN [14]. The SSGAN is trained by adaptive moment estimation optimization algorithm (Adam) with parameters $\beta\_1 = 0.5$ and $\beta\_2 = 0.9$. The minimum batch is set to 128, and the epoch is set to 200. The initial learning rate is set to 0.0003, and the decay rate is 0.9. All comparison methods follow the original settings to ensure impartiality and objectivity.

For the following experiments, the proposed SSGAN will be evaluated using the overall accuracy (OA) and confusion matrix (CM) on the EuroSAT and UCM Merced datasets.

OA: OA indicates the number of correctly classified samples divided by the total number of ones. The formula can be derived as follows:

$$\mathrm{OA} = \frac{T}{T+F}, \tag{10}$$

where T represents the sample number correctly classified, and F indicates the sample number misclassified.

CM: CM is an information table to represent the confusion ratio among different categories. The row denotes the real category, and the column indicates the predicted category. The row-column intersection indicates the proportion of real categories classified as column categories, from which it is easy to observe whether these categories are confused and confusion ratio.

In addition, to ensure the reliability of experiments, all experimental results are the mean of 10 replicate experiments with randomly selected samples.

4.3 Experimental Analysis

In this section, classification accuracies of the proposed SSGAN and several representative methods are compared on EuroSAT and UCM Merced datasets. CNNs (from scratch) is the supervised classification method, Inception V3 is the method based on transfer learning. The rest are semi-supervised classification methods based on GANs, which include SAGGAN [4], tripleGAN [9], BADGAN [10], SFGAN [13], FMGAN [8], and REGGAN [11]. Overall accuracies of several methods are presented in Table 1.

One can see the following results from Table 1.

1. The proposed SSGAN achieves the highest OA on two datasets because of the introduction of GU, PyConv, SN and Inception V3 branch and dense residual block to further enhance feature discriminant capability. SAGGAN ranks second because of the introduction of self-attention gating module and gating unit to enhance feature discriminative capability. SFGAN [13] ranks third in performance, it is probably owing to the use of pre-trained Inception V3 network to strengthen feature discriminant capability of discriminative network. Inception V3 method outperforms other methods besides SAGGAN, SFGAN and proposed SSGAN because Inception V3 is pre-trained on the large-scale ImageNet dataset and able to extract more discriminative features in remote sensing images by fine-tuning the network. SAGGAN, SFGAN and SSGAN outperform Inception

Table 1 OA(%) of SSGAN and other compared approaches. The bold shows the highest, and the underlined denotes the second

Methods	Number of tagged samples on EuroSAT				Number of tagged samples on UC Merced			
	100	1000	2000	21,600	100	200	400	1680
CNNs (from scratch) [13]	29.30	46.10	59.00	83.20	18.45	32.75	43.55	62.08
InceptionV3 [13]	63.90	84.60	87.90	91.50	55.35	71.11	81.05	85.39
FMGAN [8]	63.05	75.81	78.36	86.92	43.55	69.17	74.48	80.22
tripleGAN [9]	56.32	83.26	85.71	88.83	39.96	70.52	80.59	84.13
BADGAN [10]	59.03	76.02	78.13	86.76	18.45	32.75	43.55	62.08
REGGAN [11]	64.71	72.82	76.35	82.28	40.36	55.39	63.54	72.30
SFGAN [13]	68.60	86.10	89.00	93.20	55.48	72.49	82.56	82.34
SAGGAN [4]	<u>76.79</u>	<u>88.72</u>	<u>90.66</u>	<u>94.32</u>	<u>57.10</u>	<u>75.69</u>	<u>83.33</u>	<u>90.48</u>
SSGAN	**78.56**	**89.02**	**91.53**	**95.50**	**59.52**	**77.13**	**84.86**	**91.32**

V3 because such three methods introduce a pre-trained Inception V3 branch into the discriminative network. CNNs (from scratch) has the lowest OA, which might be since that CNNs trained from scratch can be trained only using labeled samples in a fully supervised manner.

2. The four methods FMGAN, triple GAN, BADGAN and REGGAN use the standard DCGAN structure, the OA is significantly lower than SSGAN, SAGGAN and SFGAN. However, tripleGAN achieves higher OA due to the introduction of a separate classifier network. FMGAN effectively improves the OA by using the feature matching term, and the OA is slightly lower than tripleGAN. The OA of REGGAN is the lowest among these four methods, which indicates that manifold regularization is less effective for remote sensing image scene classification with complex scenes.
3. For all methods, the higher the number of labeled samples, the higher the OA. Furthermore, the proposed SSGAN at $M = 1000$ exceeds CNNs (from scratch) at $M = 21{,}600$ on EuroSAT dataset. Interestingly, SSGAN still has higher accuracy than pre-trained Inception V3 network even when $M = 21{,}600$. It may be because SSGAN utilizes the samples generated by G for additionally training, while these generated samples are not available for CNNs (from scratch) and pre-trained Inception V3. The same trend is found on UCM Merced dataset. These demonstrate that SSGAN can achieve higher OA using fewer labeled samples.
4. On the EuroSAT dataset, the OA of SSGAN reaches 78.56.2%, 89.02%, 91.53% and 95.50% at $M = 100$, 1000, 2000 and 21,600, which is 1.77%, 0.30%, 0.87% and 1.18% higher than SAGGAN, respectively.This is probably mainly due to the introduction of dense residual blocks, PyConv, and SN in SSGAN. Similarly, the OA is 9.96%, 2.92%, 2.53% and 2.30% higher than SFGAN, respectively. These show that SSGAN is indeed effective. In particular, the overall accuracies of SSGAN at $m = 100$ is 9.96% and 1.77% higher than SFGAN and SAGGAN, which indicates that SSGAN can obtain higher performance with fewer labeled samples. Similar results can be seen on UC Merced dataset.
5. On the UC Merced dataset, the OA of SSGAN is only 59.52%, 76.13%, 83.86% and 91.02% at $M = 100$, 200, 400 and 1680, which is because the total training samples is insufficient, and that limits GANs-based semi-supervised classification performance. But, SSGAN shows the highest OA compared to other methods.

To further evaluate the performance of proposed SSGAN, confusion matrices were generated at $M = 100$, 1000, 2000 and 21,600 on EuroSAT dataset, respectively. The following observations can be obtained from the confusion matrix in Fig. 5.

1. As the number M of labeled samples increases, the accuracy of each category increases accordingly, while confusion ratio decreases. The accuracy of 8 among 10 categories is higher than 80% at $M = 100$, which indicates that the proposed SSGAN obtains higher classification accuracy with few labeled data.
2. By comparing the two confusion matrices at $M = 100$, 2000, the accuracy of categories 1, 2, 6, 7, and 8 improves by 23%, 34%, 28%, 21%, and 21%, respectively, which indicates that as the number of labeled samples increases, the classification accuracy of each category increases significantly.
3. In the case of $M = 21{,}600$, the classification accuracy of 9 among all 10 categories is higher than 95%, which shows that the proposed SSGAN can achieve good semi-supervised classification performance.

In short, the proposed SSGAN is effective.

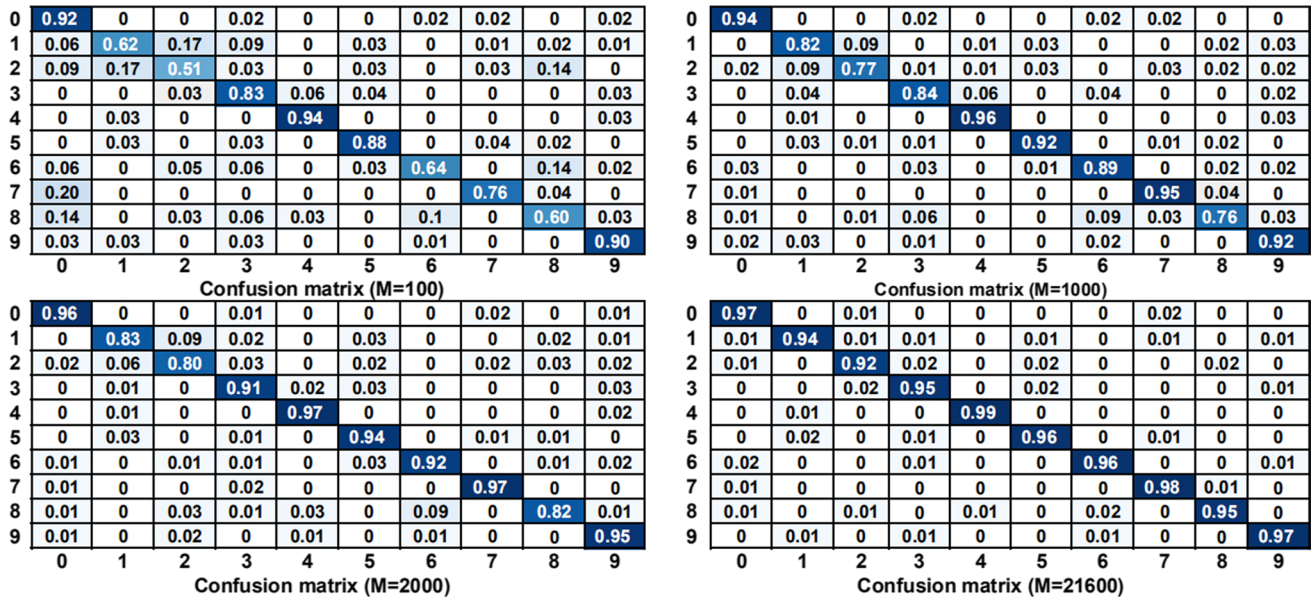

Fig. 5 The confusion matrices generated by proposed SSGAN on EuroSAT dataset at $M = 100$, 1000, 2000 and 21,600

4.4 Ablation Experiments

Compared with other semi-supervised image classification methods, the proposed SSGAN achieves the best performance by enhancing discriminative networks. In this section, the effectiveness of pre-trained Inception V3 branch, GU, PyConv and SN is verified. Four variants of proposed SSGAN are investigated individually: (1) SSGAN-GU is the variant without GU, (2) SSGAN-I is the variant without pre-trained Inception V3 branches, (3) SSGAN-P is the variant without PyConv, and (4) SSGAN-SN is the variant without SN.

For a fair comparison, extensive experiments were conducted in this paper at same experimental setup on same datasets. As shown from the experimental results in Table 2, Inception V3 branch, SN, PyConv, and GU all contribute to improving SSGAN performance on two datasets. Among them, SSGAN-I has the lowest OA in two datasets, which indicates that Inception V3 branch is the most effective because it can extract high-level semantic information from scene images and then feed extracted semantic features into GU to further enhance feature discriminative ability. The second most effective one is PyConv, because it can obtain different levels of details through multiple groups of PyConv operations with different kernel sizes to enhance feature representation. The third effective one is GU and the least effective one is SN, but the accuracy is also improved significantly on both datasets. Interestingly, SSGAN-SN has significantly lower accuracy than SSGAN when the number of labeled M is larger, which suggests that SN performs better under more labeled training samples.

5 Conclusion

In this paper, we propose a new GANs-based semi-supervised method for remote sensing scene classification using dense residual block, GU, PyConv, pre-trained Inception V3 network and SN. The proposed method achieves higher semi-supervised classification accuracy using a few labeled

Table 2 The comparison results of ablation study

Methods	The labeled sample number on EuroSAT				The labeled sample number on UC Mereced			
	100	1000	2000	21,600	100	200	400	1680
SSGAN-I	69.57	84.22	87.31	92.45	52.34	71.31	80.36	86.65
SSGAN-P	71.88	85.48	88.53	93.62	54.92	73.03	81.76	87.81
SSGAN-GU	73.65	87.24	89.73	94.38	57.88	75.37	82.72	88.51
SSGAN-SN	77.83	88.06	90.51	94.17	58.82	76.57	83.53	89.89
SSGAN	**78.56**	**89.02**	**91.53**	**95.50**	**59.52**	**77.13**	**84.86**	**91.32**

The bold indicates the highest accuracy, and the italic indicates the second highest accuracy

and numerous unlabeled samples. Specially, the pre-trained Inception V3 network is introduced into the discriminative network as a new branch to extract semantic features; the GU and PyConv are integrated into dense residual block to strengthen the feature discriminant capability of discriminative network; and SN is introduced into both generative and discriminative network to stabilize the training of GANs to improve semi-supervised classification accuracy. Comprehensive experimental results illustrate the proposed approach gains higher overall accuracy compared with other comparison methods, especially, when only there are a few labeled samples. In the future, it is planned to investigate unsupervised remote sensing image scene classification based on GANs, which is more difficult in the field of computer vision.

Acknowledgements The authors would like to thank the editors and reviewers for their outstanding comments and suggestions, which provide us with momentum and guidance to make deeper research into our subject matter and further improve our paper.

Author Contributions All authors contributed to the study conception and design. The study was mainly conceived and designed by DG. The experiments were performed by ZW an YZ. The frst draft of the manuscript was written by DG and all authors commented on previous versions of the manuscript. ZS edited the manuscript. All authors read and approved the fnal manuscript.

Funding This work was supported by the National Natural Science Foundation of China (Nos. 62102200, 41571401), the Science and Technology Research Project of Henan Province under Grants 212102210492, and the Key Research Projects of Henan Higher Education Institutions under Grants 23A520053, and the Science and Technology Research Project of Nanyang City under Grants KJGG102.

Availibility of Data and Materials The datasets used during the study are available at http://weegee.vision.ucmerced.edu/datasets/landuse.html and https://github.com/phelber/eurosat.

Declarations

Conflict of Interest The authors declare that they have no competing interests.

Ethics Approval and Consent to Participate Not applicable.

Consent for Publication Not applicable.

References

1. Cheng, G., Xie, X., Han, J., Guo, L., Xia, G.-S.: Remote sensing image scene classification meets deep learning: challenges, methods, benchmarks, and opportunities. IEEE J. Selected Topics Appl. Earth Observations Remote Sens. **13**, 3735–3756 (2020)
2. He, N., Fang, L., Li, S., Plaza, A., Plaza, J.: Remote sensing scene classification using multilayer stacked covariance pooling. IEEE Trans. Geosci. Remote Sensing **56**(12), 6899–6910 (2018)
3. Guo, D., Xia, Y., Luo, X.: Self-supervised gans with similarity loss for remote sensing image scene classification. IEEE J. Selected Topics Appl. Earth Observations Remote Sensing **14**, 2508–2521 (2021)
4. Guo, D., Xia, Y., Luo, X.: Gan-based semisupervised scene classification of remote sensing image. IEEE Geosci. Remote Sensing Lett. **18**(12), 2067–2071 (2021)
5. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Neural Information Processing Systems (2014)
6. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. In: Int. Conf. Learn. Representations. 1–13 (2017)
7. Miyato, T., Maeda, S.-I., Koyama, M., Ishii, S.: Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE Trans. Pattern Anal. Mach. Intell. **41**(8), 1979–1993 (2019)
8. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: Adv. Neural Inf. Proc. Syst. 1–10 (2016)
9. Li, C., Xu, K., Zhu, J., Zhang, B.: Triple generative adversarial nets. In: Adv. Neural Inf. Proc. Syst. 1–9 (2017)
10. Dai, Z., Yang, Z., Yang, F., Cohen, W.W., Salakhutdinov, R.: Good semi-supervised learning that requires a bad gan. In: Adv. Neural Inf. Proc. Syst. 1–8 (2017)
11. Lecouat, B., Foo, C.-S., Zenati, H., Chandrasekhar, V.R.: Semi-supervised learning with gans: Revisiting manifold regularization. In: Int. Conf. Learn. Representations. 1–11 (2018)
12. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: Int. Conf. Learn. Representations. 1–10 (2015)
13. Roy, S., Sangineto, E., Sebe, N., Demir, B.: Semantic-fusion gans for semi-supervised satellite image classification. In: 2018 25th IEEE International Conference on Image Processing (ICIP) (2018)
14. Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z.: Photo-realistic single image super-resolution using a generative adversarial network. In: IEEE Conference on Computer Vision and Pattern Recognition, 105–114 (2017)
15. Miech, A., Laptev, I., Sivic, J.: Learnable pooling with context gating for video classification. In: IEEE Conf. Comput. Vis. Pattern Recogn. 1–6 (2017)
16. Liu, W., Huang, X., Cao, G., Zhang, J., Yang, L.: Multi-modal sequence model with gated fully convolutional blocks for micro-video venue classification. Multimed. Tools Appl. **79**(2) (2020)
17. Duta, I.C., Liu, L., Zhu, F., Shao, L.: Pyramidal convolution: rethinking convolutional neural networks for visual recognition. In: IEEE Conf. Comput. Vis. Pattern Recogn. 1–6 (2020)
18. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. In: Int. Conf. Learn. Representations. 1–11 (2018)
19. Helber, P., Bischke, B., Dengel, A., Borth, D.: Eurosat: a novel dataset and deep learning benchmark for land use and land cover classification. IEEE J. Selected Topics Appl. Earth Observat. Remote Sensing **12**(7), 2217–2226 (2019)
20. Yang, Y., Newsam, S.: Bag-of-visual-words and spatial extensions for land-use classification. In: Sigspatial International Conference on Advances in Geographic Information Systems, 270 (2010)

21. Springenberg, J.T.: Unsupervised and semi-supervised learning with categorical generative adversarial networks. In: Int. Conf. Learn. Representations. 1–9 (2016)
22. Srivastava, R.K., Greff, K., Schmidhuber, J.: Training very deep networks. In: Adv. Neural Inf. Proc. Syst. 1–10 (2015)
23. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: Int. Conf. Learn. Representations. 1–10 (2016)
24. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6230–6239 (2017)
25. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. In: IEEE Conf. Comput. Vis. Pattern Recogn. 1–6 (2017)
26. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: IEEE Conf. Comput. Vis. Pattern Recogn. 1–6 (2018)