remote sensing

MDPI

Article

Semi-Supervised DEGAN for Optical High-Resolution Remote Sensing Image Scene Classification

Jia Li [1], Yujia Liao [2], Junjie Zhang [1], Dan Zeng [1] and Xiaoliang Qian [3,*]

[1] Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Joint International Research Laboratory of Specialty Fiber Optics and Advanced Communication, Shanghai Institute of Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China
[2] School of Automation, Northwestern Polytechnical University, Xi'an 710072, China
[3] College of Electric and Information Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, China
* Correspondence: qxlzengli@zzuli.edu.cn

Abstract: Semi-supervised methods have made remarkable achievements via utilizing unlabeled samples for optical high-resolution remote sensing scene classification. However, the labeled data cannot be effectively combined with unlabeled data in the existing semi-supervised methods during model training. To address this issue, we present a semi-supervised optical high-resolution remote sensing scene classification method based on Diversity Enhanced Generative Adversarial Network (DEGAN), in which the supervised and unsupervised stages are deeply combined in the DEGAN training. Based on the unsupervised characteristic of the Generative Adversarial Network (GAN), a large number of unlabeled and labeled images are jointly employed to guide the generator to obtain a complete and accurate probability density space of fake images. The Diversity Enhanced Network (DEN) is designed to increase the diversity of generated images based on massive unlabeled data. Therefore, the discriminator is promoted to provide discriminative features by enhancing the generator given the game relationship between two models in DEGAN. Moreover, the conditional entropy is adopted to make full use of the information of unlabeled data during the discriminator training. Finally, the features extracted from the discriminator and VGGNet-16 are employed for scene classification. Experimental results on three large datasets demonstrate that the proposed scene classification method yields a superior classification performance compared with other semi-supervised methods.

Keywords: semi-supervised scene classification; optical high-resolution remote sensing image; Generative Adversarial Network; Diversity Enhanced Network; Improved Fisher Kernel; feature fusion

Citation: Li, J.; Liao, Y.; Zhang, J.; Zeng, D.; Qian, X. Semi-Supervised DEGAN for Optical High-Resolution Remote Sensing Image Scene Classification. *Remote Sens.* **2022**, *14*, 4418. https://doi.org/10.3390/rs14174418

Academic Editors: Xiangtao Zheng, Fulin Luo and Qi Wang

Received: 30 July 2022
Accepted: 1 September 2022
Published: 5 September 2022

1. Introduction

Automatically understanding and interpreting massive high-resolution remote sensing images is critical in various remote sensing applications. Remote sensing image scene classification, a process intended to tag remote sensing images with semantic categories based on image content, can provide valuable information for object recognition [1,2], image segmentation [3,4], and similar tasks and effectively improve the image interpretation performance [5]. To date, remote sensing image scene classification techniques have been widely applied to change detection [6,7], environmental monitoring [8], urban planning [9], and other fields [10–14].

Recently, the deep convolutional neural network (CNN) has achieved significant success in computer vision [15,16], as well as been widely applied to optical high-resolution remote sensing scene image classification [17,18]. Compared to the traditional hand-design [19,20] and coding features-based [21] scene classification methods, the deep learning-based methods relying on automatically extracting high-level semantic information from images achieve a promising scene classification effect and have become the mainstream

approach [22–28]. According to the composition of the training samples, deep learning-based methods mainly include two categories, i.e., supervised and semi-supervised. Currently, the training of most deep learning-based methods is supervised. Supervised learning methods consist of two forms, one of which involves extracting the image features from a pre-trained or fine-tuned network and then adopting a classifier such as a support vector machine (SVM) for classification. For example, Cheng et al. extracted the trained features from the convolutional layer and then encoded them using Bag of Visual Words (BoVW) to form the final image representation for classification [29]. The other is to design an end-to-end network for training and testing. For instance, Liu et al. proposed a model named SPP-net that used end-to-end training and testing for high-resolution remote sensing image scene classification [30]. This model employs spatial pyramid pooling to solve the problem that the convolutional neural network (CNN) training and testing require the input images to be a fixed size. Furthermore, in the latest study, the complex networks fused with other modules have been developed for scene classification, such as the attention mechanism including channel-attention [31] and self-attention [32], GCN [33], multimodal [34], etc.

However, a large number of labeled training samples are required to achieve high classification accuracy for the supervised methods based on CNN. It is quite tough and labor-intensive to annotate the remote sensing image [23]. Moreover, with insufficient labeled samples, the overfitting will appear in the deeper network training such that the model yields poor classification performance. The above issue is a significant constraint on the wide application of supervised classification methods. Therefore, with the assistance of unlabeled samples, semi-supervised methods relying on the advantage of fewer labeled samples have become one of the most important research directions in the scene classification field. The existing representative semi-supervised scene classification methods can be coarsely grouped into two categories. One of them serves to annotate the unlabeled samples by designing a self-labeling algorithm, then the generated samples are utilized to promote the classification performance with supervised training. Han et al. [35] presented a semi-supervised generative framework named SSGF to classify the remote sensing scene image. Several classifiers are first trained on the confusing categories using the validation set and the input images are classified by two different depth networks. Subsequently, the input image is labeled based on the consistency of the output results and the judging of the confusing categories, while the training set is also updated. The above steps are repeated until the sample labeling is complete. On the other hand, the combination of unsupervised feature extraction and supervised classifiers learning is the second category of semi-supervised methods. For example, Dai et al. [36] proposed to adopt a joint ResNet and integrated learning strategy to obtain the most effective representation of images, then supervised training is utilized for scene classification. However, the first approach suffers from two problems, including the inaccurate labeling of unlabeled data and the under-utilization of information. As for the second kind of method, since the stages of unsupervised and supervised are separated, the information of unlabeled and labeled data are not jointly exploited. Both the labeled and unlabeled data are beneficial to the unsupervised feature extraction and supervised classifier learning. Generally speaking, the above two types of methods do not effectively combine the labeled and unlabeled data during the training procedure.

Compared with the supervised scene classification methods, the main advantage of semi-supervised methods is to utilize a large number of unlabeled samples to enhance the discriminative ability of the features originating from the network. However, as mentioned earlier, the supervised stage (labeled samples) is not effectively combined with the unsupervised stage (unlabeled samples) in existing semi-supervised methods. Therefore, we present a novel semi-supervised Diversity-Enhanced Generative Adversarial Network (DEGAN) for the optical high-resolution remote sensing scene classification. The supervised and unsupervised stages are deeply combined in the DEGAN training, subsequently, the features originating from the discriminator and VGGNet-16 are employed for final scene classification. In DEGAN, the unlabeled images are utilized to improve

the feature extraction ability of the discriminator by introducing the conditional entropy into the loss of the discriminator. In addition, considering the game relationship between two sub-networks, the discriminator is enhanced by strengthening the generator from the two aspects, e.g., a lot of unlabeled data are utilized to guide the fake image generation, and diversity-enhanced network (DEN) is presented to improve the diversity of fake images from the information entropy perspective. During generator training, a large number of unlabeled and few labeled samples together guide the generation of fake images, which promotes the network to obtain a more complete and accurate probability density space of fake images. Since the insufficient diversity of the generated images is a manifestation of low entropy in the information entropy theory, we design a DEN to maximize the information entropy and further increase the diversity of generated images. As for the discriminator training, the conditional entropy is adopted to make full use of the information of unlabeled data.

The framework of the proposed DEGAN is shown in Figure 1. DEGAN consists of a generator and a discriminator. The discriminator uses a multi-output structure and is responsible for the discrimination between the generated images and the various real-image categories. The generator contains two sub-networks, namely the fake-image generating network (FGN) and diversity enhancing network (DEN). In this scheme, the task of the FGN is to generate fake images and the DEN assists FGN in increasing the diversity of fake images by maximizing the information entropy.

The flowchart of the proposed scene classification method is shown in Figure 2. During training, DEGAN is first trained using a small number of labeled images and large amounts of unlabeled images, among which labeled images are also utilized to fine-tune the VGGNet-16. Then, the coding features are learned by Improved Fisher Kernel (IFK) using convolutional features extracted from two models. Finally, the fully connected layer features extracted from the two models and the coding features are fused to train SVM. Generally, the same labeled images are used throughout the entire training process, including codebook generation and SVM training. For testing, the images are input to the discriminator and VGGNet-16, respectively. With one-dimensional feature extraction and two-dimensional feature coding, four types of input image features are fused and further classified with SVM.

The major contributions of this paper are as follows.

1. We propose a semi-supervised DEGAN for optical high-resolution remote sensing image scene classification, in which the labeled and unlabeled images are effectively combined during the model training. A lot of unlabeled data can significantly improve the generator and further enhance the discriminator given the game relationship between two sub-networks in DEGAN.
2. We design a DEN in generator to increase the diversity of fake images by maximizing the information entropy.
3. We employ the conditional entropy in the discriminator training to make full use of the information of the unlabeled data.

The remainder of this paper is organized as follows. In Section 2, the related works concerning the proposed method are introduced. The proposed semi-supervised DEGAN for optical high-resolution remote sensing scene classification is described in Section 3. The experimental results and analysis are presented in Section 4. Finally, our conclusions are summarized in Section 5.

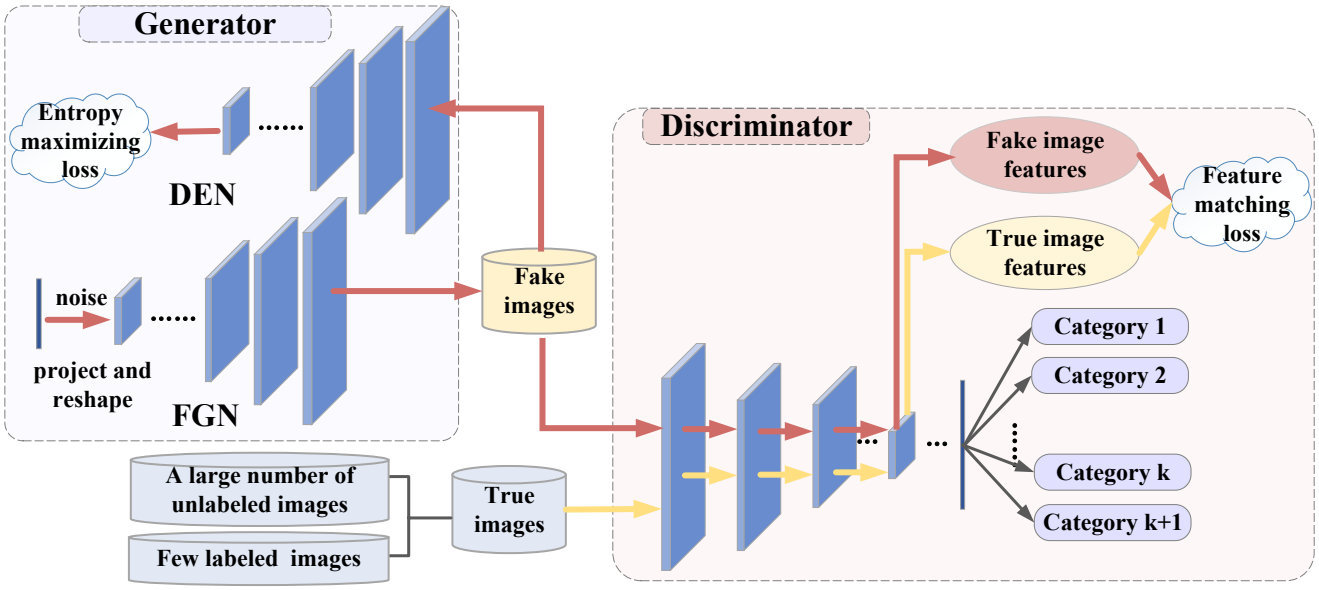

Figure 1. The framework of DEGAN.

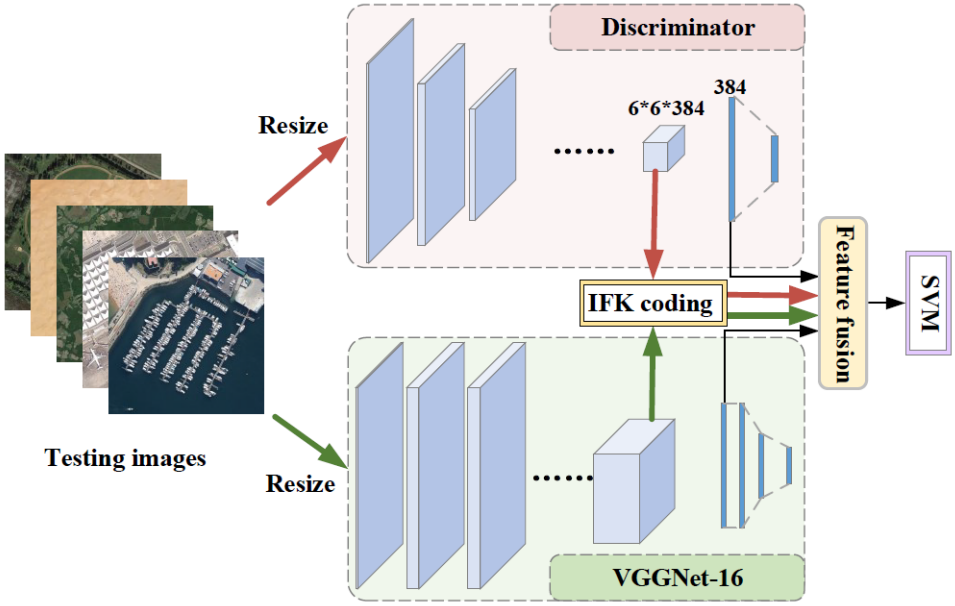

Figure 2. The flowchart of the proposed scene classification based on DEGAN.

2. Related Work

Since we apply the deep learning features and further code them in the proposed method, the existing coding feature-based and deep learning-based scene classification methods are introduced. Moreover, we also describe the relative semi-supervised learning methods in this section.

2.1. High-Resolution Remote Sensing Image Scene Classification

2.1.1. Coding Feature-Based Methods

The coding feature-based methods generate the dictionary clustered by low-level image features to map the representation of the image. These algorithms typically include four steps: local feature detection, codebook generation, global feature description, and image classification. The Bag of Visual words (BoVW) [21] is a typical coding-features

method which clusters the hand-design features to a dictionary and further codes the remote sensing image into a histogram according to the dictionary. Compared with the BoVW, the Spatial Pyramid Matching (SPM) [37] adds spatial information to image features to achieve more accurate representations of remote sensing images. Moreover, topic models, such as Probabilistic Latent Semantic Analysis (PLSA) [38] and Latent Dirichlet Allocation (LDA) [39] are also introduced to recognize scenes. These algorithms adopt topics obtained by low-level features to represent remote sensing images. Improved Fisher Kernel (IFK) [40] mapping the final image representation based on the Gaussian Mixture Model (GMM) has a promising effect on scene classification. The coding feature-based methods further integrated the low-level features such that the advanced scene classification performances are achieved.

2.1.2. Deep Learning-Based Methods

In recent years, with the emergence of CNNs, remote sensing image classification methods based on deep-learning features have made great strides [41,42]. Compared with the traditional methods, a CNN learns more discriminative features by training on large numbers of images without complex engineering work for feature descriptors, and the superiority of CNNs is obvious when faced with complex scene classification tasks. With the success of general models proposed for natural images' processing tasks, they are also utilized for scene classification. Hu et al. [43] first used pretrained networks to extract high-level semantic features, such as VGGNet [44] and AlexNet [45]. Most of the deep learning-based methods are targeted toward further improving the features of a CNN or designing a new end-to-end model. Wang et al. [46] proposed a novel end-to-end attention recurrent convolutional network (ARCNet) for scene classification. This model explores the use of an attention mechanism to improve scene classification. Some algorithms improve features by combining CNNs with other methods and are also successful at remote sensing scene classification. For instance, the local features of scene images extracted from different depth layers of CNNs are encoded to obtain global representations based on feature coding using the BoVW model and the Improved Fisher Kernel (IFK) [40] in [43]. Chaib et al. [47] adopted discriminant correlation analysis (DCA) to fuse image features derived from the fully connected layers of a pre-trained VGGNet, resulting in image features with much lower dimensions. Subsequently, the complex networks fused with other modules have been developed for scene classification, such as an attention mechanism including channel-attention [31] and self-attention [32], GCN [33], and multimodal [34].

2.2. Semi-Supervised Learning

In the past decade, semi-supervised learning (SSL) has been successfully applied in many fields. When a small amount of labeled data are available, SSL can effectively utilize the unlabeled data to promote performance, among which consistency regularization and entropy minimization are representative methods. The consistency regularization assumes that a classifier should output the same class distribution for applying data augmentation to semi-supervised learning, which enforces that an unlabeled example should be classified the same as an augmentation of itself [48]. However, the domain-specific data augmentation strategies limit the effect of consistency regularization methods. To overcome the above drawback, virtual adversarial training (VAT) computes an additive perturbation to the input, which maximally changes the distribution of output categories [49]. There is a common assumption that the classifier's decision boundary should not pass through high-density regions of the marginal data distribution. Therefore, the low-entropy predictions from the classifier of unlabeled data are required in entropy minimization methods [50]. For example, Lee et al. trained the network with labeled and unlabeled data simultaneously, and the unlabeled data are arranged to the class which has the maximum predicted probability [51]. Moreover, Berthelot et al. proposed an SSL algorithm named MixMatch, which introduces a unified loss term for unlabeled data that seamlessly reduces entropy while maintaining consistency and remaining compatible with traditional regularization

techniques [48]. Kihyuk et al. introduces a simple SSL method FixMatch by retaining the pseudo-label with a high-confidence prediction [52].

2.3. Semi-Supervised High-Resolution Remote Scene Image Classification

In recent decades, a variety of semi-supervised methods dedicated to high-resolution remote sensing image scene classification have been proposed to address the problem of large training sample requirements in the supervised methods. According to the principle of using unlabeled images, these methods can be coarsely grouped into two categories. The first type is to generate the label for the unlabeled data, and then they are employed to improve the classification accuracy with supervised training. A semi-supervised generative framework named SSGF was proposed by Han et al. [35] to classify the remote sensing scene image. SSGF adopts several classifiers to determine the category of unlabeled data in the confusing categories, subsequently, the input images are classified by different classifiers. According to the yielded results and confusing categories, the unlabeled data are assigned a label, and the training set is simultaneously updated. The above steps are repeated until the training process is complete. Tian et al. [53] employed multiple models trained by simple samples to generate the pseudo-labels. Then, the labeled, pseudo labeled, and unlabeled samples are simultaneously utilized to train the model in a semi-supervised method. Unlike the above methods, there are some semi-supervised scene classification methods relying on unsupervised feature learning. Since Autoencoders and GAN can automatically learn valuable representations without any labeled data, these two models are generally adopted for scene classification. Cheng et al. [54] combined the autoencoder and a single hidden layer neural network to obtain a more effective sparse representation. A convolution sparse autoencoder was designed by Han et al. [55] to solve the issue of inadequate representation. To learn mid-level visual features, Cheng et al. [56] introduced a novel autoencoder and further improved the classification accuracy. Yao et al. [57] added the paired constraints for a stacked sparse autoencoder, which can provide more discriminative feature representation for scene classification. In addition to autoencoder, GAN is also applied to scene classification. Lin et al. [58] presented a multiple-layer feature-matching constraint for GAN to strengthen the model ability. An unsupervised attention-GAN was proposed by Yu et al. [59] to enhance the feature representation ability of the discriminator, in which the loss functions of the generator and discriminator are improved. Moreover, to obtain a more effective representation of images, Dai et al. [36] proposed to jointly adopt ResNet and integrate a learning strategy; then, supervised training was utilized for scene classification. The relative semi-supervised scene classification methods and their main features are listed in Table 1.

Table 1. Existing semi-supervised scene classification methods.

Method Categories	Basic Model	Mthods
Pseudo-label generation	/	Han [35], Tian [53]
Unsupervised feature learning	Autoencoder	Cheng [54],Han [55], Cheng [56], Yao [57]
	GAN	Lin [58], Yu [59]
	ResNet	Dai [36]

2.4. Generative Adversarial Network

In 2014, Ian Goodfellow [60] of Google Brain scientists proposed the Generative Adversarial Network (GAN) based on the idea of an adversarial game. It is mainly composed of two models, namely the generator and discriminator. The generator focuses on generating the new samples to learn the potential distribution in the real data samples, while the discriminator is responsible for determining whether the input data are from real data or generated data (fake data). During training, the parameters of the generator (G) and

discriminator (D) are updated alternately, and the optimization of GAN can be formulated as a min–max problem:

$$\min_G \max_D f(D,G) = \mathbf{E}_{x\sim P_{data}}(\log(\mathrm{D}(\mathrm{x}))) + \mathbf{E}_{\mathrm{z}\sim \mathrm{P}_{\mathrm{z}}}(\log(1-\mathrm{D}(\mathrm{G}(\mathrm{z})))) \quad (1)$$

Nowadays, many GAN variants have been presented to promote the ability of models or for different tasks. The deep convolutional GAN (DCGAN) [61] replaces the G and D with two CNN in the original structure of GAN. Considering that the Jensen–Shannon divergence is not suitable for measuring the distance of the distribution, the Wasserstein distance is employed in the Wasserstein GAN (WGAN) [62], making the training procedure more stable. Moreover, the WGAN-gradient penalty (WGAN-GP) [63] was introduced by Gulrajani to address the slow convergence problem of WGAN. In addition, there are many GAN variants aimed at the training procedure to achieve fast and stable convergence, such as the least square GAN [64], Loss-sensitive GAN [65], Energy-Based GAN (EBGAN) [66], Boundary Equilibrium GAN (BEGAN) [67], etc. Both the above GANs are based on the original foundation of GAN, hence, GANs with different structures are proposed in practice. Mirza et al. [68] presented the conditional GAN (CGAN) to obtain the samples with a unique category. The information maximizing GAN (InfoGAN) [69] decomposes the input noise vector into two parts, including z and c, in which z is considered incompressible noise, and c represents the significant semantic features of the real samples. Furthermore, some models are designed for other tasks. For instance, CycleGAN [70] is proposed for image translation, which does not require pairing data. Unlike CycleGAN, StarGAN [71] can train the same model by implementing joint training between multiple datasets. It aims at mapping multiple domains. With the development of deep learning, massive GAN models are designed and applied to different fields.

3. Proposed Method

3.1. Overview

To effectively exploit the information of unlabeled and labeled data during the feature extraction and classification stages, the Diversity-Enhanced Generative Adversarial Network (DEGAN) is proposed to joint utilize the labeled and unlabeled remote sensing images in the whole model training procedure. In DEGAN, the unlabeled images are utilized to improve the feature extraction ability of the discriminator by introducing the conditional entropy into the loss of the discriminator. In addition, a diversity-enhanced network (DEN) is designed to enhance the generator from the information entropy perspective, which further promotes the discriminator according to the game relationship between the generator and discriminator. Moreover, to introduce the prior knowledge of natural images, the VGGNet-16 is employed and fine-tuned with the optical high-resolution remote sensing images. After the training of models, the convolutional features extracted from the discriminator and VGGNet-16 are encoded by Improved Fisher Kernel (IFK) due to its stronger ability to abstract the features. Finally, the fully connected features and coding features are concatenated as the representation of a remote sensing scene image, which is fed to the SVM for scene inference.

3.2. Modeling of DEGAN

3.2.1. Modeling of Generator

The generator is responsible for generating fake images to fool the discriminator, namely the distribution of real images is learned. However, the generator in conventional GAN usually cannot precisely learn the distribution of real images such that the diversity of fake images is insufficient. Therefore, the designed generator in DEGAN consists of two sub-networks, i.e., Fake-image Generating Network (FGN) and Diversity Enhanced Network (DEN), among which the FGN is responsible for generating fake images, and the DEN is designed to increase the diversity of fake images. Since the insufficient diversity of fake images is a direct manifestation of the low entropy of generated feature distribution,

maximizing the entropy by DEN can increase the diversity of generated images and further enhance the capacity of the generator.

Since the entropy is dynamic in the high-dimensional feature space through the training process and stable in the input space, we chose to increase the generator's entropy in the input space according to variational inference (VI), which is noted as $\mathbf{H}(p_{gen}(x))$. Inspired by [72], $\mathbf{H}(p_{gen}(x))$ can be maximized by minimizing the conditional entropy:

$$\mathbf{H}(p_{gen}(z|x)) = \mathbf{E}_{x\sim p_{gen}(x)}[\mathbf{E}_{z\sim p_{gen}(z|x)}[-\log p_{gen}(z|x)]]. \tag{2}$$

Considering the difficulty in calculating the posterior probability $\mathbf{H}(p_{gen}(z|x))$, that value can be replaced by minimizing a variational upper bound $\mathbf{U}(q_{gen})$ defined by an approximate posterior $q_{gen}(z|x)$:

$$\begin{aligned}\mathbf{H}(p_{gen}(z|x)) &= \mathbf{E}_{x\sim p_{gen}(x)}[\mathbf{E}_{z\sim p_{gen}(z|x)}[-\log q_{gen}(z|x)] - \mathbf{KL}(p_{gen}(z|x)||q_{gen}(z|x))] \\ &\leq \mathbf{E}_{x\sim p_{gen}(x)}[\mathbf{E}_{z\sim p_{gen}(z|x)}[-\log q_{gen}(z|x)]] \\ &= \mathbf{U}(q_{gen}).\end{aligned} \tag{3}$$

The variational upper bound $\mathbf{U}(q_{gen})$ can also be rewritten as follows:

$$\begin{aligned}\mathbf{U}(q_{gen}) &= \mathbf{E}_{x,z\sim p_{gen}(x,z)}[-\log q_{gen}(z|x)] \\ &= \mathbf{E}_{z\sim p_{gen}(z)}[\mathbf{E}_{x\sim p_{gen}(x|z)}[-\log q_{gen}(z|x)]].\end{aligned} \tag{4}$$

Consequently, $\mathbf{H}(p_{gen}(x))$ can be effectively maximized by minimizing the upper bound $\mathbf{U}(q_{gen})$ of the conditional entropy $\mathbf{H}(p_{gen}(z|x))$. In [72], the approximate posterior distribution $q_{gen}(z|x)$ is parameterized with a diagonal Gaussian distribution whose mean and covariance matrix are the output of a trainable inference network, i.e.,

$$\begin{aligned}q_{gen}(z|x) &= \mathcal{N}(\mu, \mathrm{I}\sigma^2) \\ \mu, \log\sigma &= f^{\text{infer}}(x),\end{aligned} \tag{5}$$

where f^{infer} denotes the inference network and I is the identity matrix. Therefore, the DEN is designed as the inference network in this paper that maximizes the entropy of the generated features to increase the diversity of fake images.

Architecture

Figures 3 and 4 show the visualization of the designed FGN and DEN in the generator, respectively. Inspired by the encoder–decoder structure, the FGN and DEN are designed as symmetrical network structures. In addition, they are designed as networks with only a small number of layers to reduce the model parameters to facilitate training. In the FGN, 100-dimensional noise drawn from a Gaussian distribution is taken as input. Then, we reshape the input into a $4 \times 4 \times 512$ tensor, and six transposed convolutional layers are employed to generate images. Ultimately, a $256 \times 256 \times 3$ remote sensing image is obtained. In the DEN, we first downsample a $256 \times 256 \times 3$ fake image originated from the FGN to $4 \times 4 \times 512$ feature maps through the six convolution layers. Then, the feature maps are reshaped into a 8192-dimensional vector, and one fully connected layer is subsequently adopted to extract a 200-dimensional vector. Finally, the yielded vector is split into two 100-dimensional vectors, which are taken as the mean and variance of the Gaussian distribution, respectively.

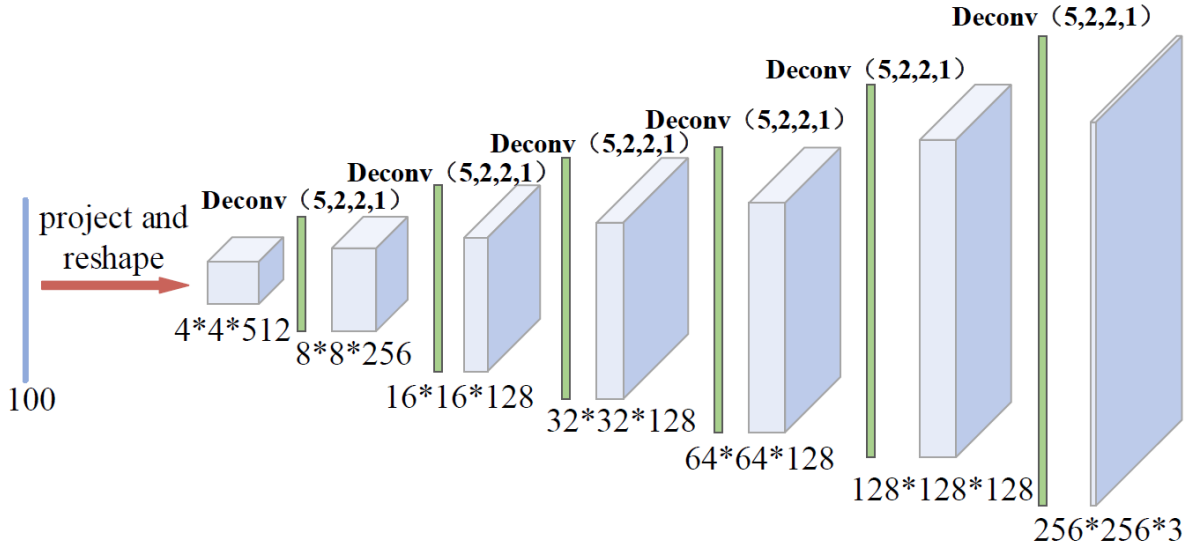

Figure 3. The architecture of the FGN.

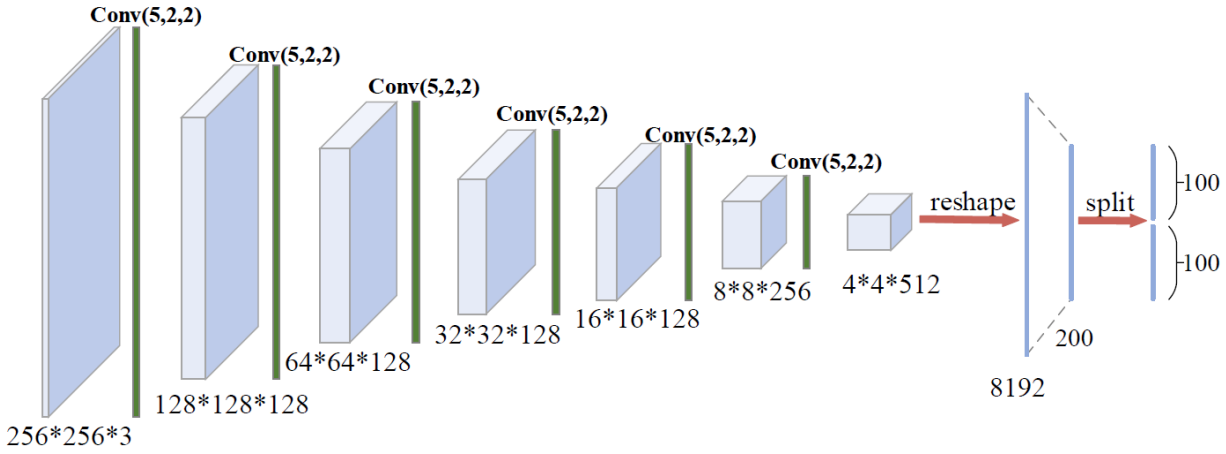

Figure 4. The architecture of the DEN.

Training Loss

Two principles are followed when designing the generator loss function: one is to make the generated images as similar to the real images as possible and the other is to increase the diversity of generated images. Therefore, $\mathbf{L}_G$ can be expressed as follows:

$$\mathbf{L}_G = \mathbf{L}_{FM} + \mathbf{L}_{EM}, \tag{6}$$

where $\mathbf{L}_{FM}$ and $\mathbf{L}_{EM}$ are designed for the first and second principles above, respectively. Then, the two parts of $\mathbf{L}_G$ are described in detail separately.

Inspired by [73], the technique of feature matching is employed to help the generator generate images similar to the training images. Therefore,

$$\mathbf{L}_{FM} = \left\| \mathbf{E}_{x \sim I_{real}} f(x) - \mathbf{E}_{z \sim p_z(z)} f(\mathbf{G}(z)) \right\|_2^2, \tag{7}$$

where $x \sim I_{real}$ and $p_z(z)$ are real images and the distributions of generated images, respectively, $\mathbf{G}(z)$ represents generated images and $f(x)$ is the output from an intermediate layer of the discriminator.

$\mathbf{L}_{EM}$ is used to calculate the information entropy of the generated image. Therefore,

$$\begin{aligned} \mathbf{L}_{EM} &= -p\log(p) \\ p &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}, \end{aligned} \tag{8}$$

where z is the input noise, μ and σ are the variance and mean of the Gaussian distribution, respectively.

In summary, the entire loss function for training the generator is intended to minimize the following:

$$\begin{aligned} \mathbf{L}_G &= \mathbf{L}_{FM} - \mathbf{L}_{EM} \\ &= \left\| \mathbf{E}_{x\sim I_{real}} f(x) - \mathbf{E}_{z\sim p_z(z)} f(\mathbf{G}(z)) \right\|_2^2 + p\log(p) \\ p &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}. \end{aligned} \tag{9}$$

We replace the maximizing $\mathbf{L}_{EM}$ by minimizing the negative value of $\mathbf{L}_{EM}$ in the above formula.

3.2.2. Modeling of Discriminator

Architecture

The architecture is shown in Figure 5, in which different convolution layers are designed with different convolutional kernels. First, we fed the $256 \times 256 \times 3$ images into the discriminator. Then, the $6 \times 6 \times 384$ feature maps are obtained through the ten convolutional layers. The feature maps are subsequently transformed into a 384-dimensional vector by average pooling. Finally, the yielded vector is input to one fully connected layer followed by a softmax layer to produce the classification result. In the discriminator network, the input images are convoluted into smaller feature maps by the convolution kernels, which have larger strides in the first few layers. To increase the discriminative ability of the image features, some of the convolution layers do not alter the size of the feature maps, while the feature maps are abstractly expressed several times through these convolution layers.

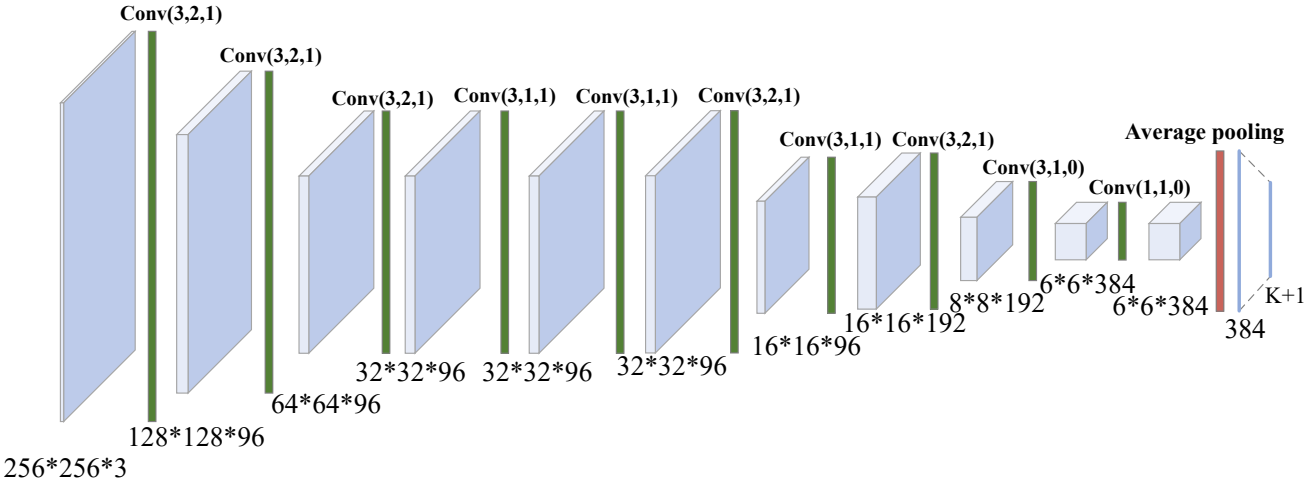

Figure 5. The architecture of the discriminator.

Training Loss

Three kinds of images are presented to the discriminator, namely real labeled images l, real unlabeled images u, and fake images G derived from the generator, where both u

and G are unlabeled images. Consequently, the loss $\mathbf{L}_D$ mainly includes the supervised $\mathbf{L}_{supervised}$ and unsupervised $\mathbf{L}_{unsupervised}$ parts corresponding to the labeled and unlabeled images in the training set, respectively. The discriminator outputs $K+1$ types, in which the real images correspond to the first K type of the output and the generated fake images correspond to the $K+1$ output. The loss function of the discriminator is introduced in detail as follows.

As the case is under common supervised training, we employ the cross-entropy to enable the discriminator to accurately assign labeled images to their respective categories in the first K output of the discriminator in the proposed method. Here, $\mathbf{L}_{supervised}$ is also denoted as $\mathbf{L}_l$ for the training of labeled images l, that is,

$$\mathbf{L}_{supervised} = \mathbf{L}_l = \mathbf{E}_{x,y\sim l}\log p_D(y|x, y \leq K). \tag{10}$$

For the unsupervised part $\mathbf{L}_{unsupervised}$, $\mathbf{L}_u$ and $\mathbf{L}_{AD}$ represent the real unlabeled images and the generated fake images, respectively. The loss $\mathbf{L}_{AD}$ encourages the fake images to be classified into the $K+1$ category and is defined as follows:

$$\mathbf{L}_{AD} = \mathbf{E}_{x\sim G}\log p_D(K+1|x). \tag{11}$$

$x \sim G$ stands for the fake images and $\log p_D(K+1|x)$ represents the predicted output of the discriminator in the $K+1$ category. For the input real unlabeled images, the loss $\mathbf{L}_u$ is designed as follows:

$$\mathbf{L}_u = \mathbf{E}_{x\sim u}\log p_D(y \leq K|x) + \mathbf{E}_{x\sim u}\sum_{k=1}^{K} p_D(k|x)\log p_D(k|x), \tag{12}$$

where $x \sim u$ represents the real unlabeled images, $y \leq K$ represents any category in the first K categories, and $\log p_D(y \leq K|x)$ represents the predicted output of the discriminator in any one in the first K categories. In addition, to further exploit the information of unlabeled data in the discriminator, we add a conditional entropy [49,74] to the unsupervised part $\mathbf{L}_u$ for the real unlabeled samples, which guarantees that the discriminator will have a strong ability to discriminate between real and fake images. The conditional entropy is $\mathbf{E}_{x\sim u}\sum_{k=1}^{K} p_D(k|x)\log p_D(k|x)$, where k represents each category in the first K categories. Consequently,

$$\begin{aligned}\mathbf{L}_{unsupervised} &= \mathbf{L}_u + \mathbf{L}_{AD} \\ &= \mathbf{E}_{x\sim u}\log p_D(y \leq K|x) + \mathbf{E}_{x\sim u}\sum_{k=1}^{K} p_D(k|x)\log p_D(k|x) + \mathbf{E}_{x\sim G}\log p_D(K+1|x).\end{aligned} \tag{13}$$

Finally, the discriminator training is realized by minimizing:

$$\begin{aligned}\mathbf{L}_D &= \mathbf{L}_{supervised} + \mathbf{L}_{unsupervised} \\ &= \mathbf{L}_l + \mathbf{L}_u + \mathbf{L}_{AD} \\ &= \mathbf{E}_{x,y\sim l}\log p_D(y|x, y \leq K) + \mathbf{E}_{x\sim u}\log p_D(y \leq K|x) \\ &+ \mathbf{E}_{x\sim u}\sum_{k=1}^{K} p_D(k|x)\log p_D(k|x) + \mathbf{E}_{x\sim G}\log p_D(K+1|x).\end{aligned} \tag{14}$$

During DEGAN training, the parameters of the discriminator are fixed when the generator is trained and the parameters of the generator are fixed when the discriminator is trained. The two training processes above were implemented alternatively until the training was complete. In the iterative training process, both the generator and discriminator can be assigned different training times, in which the training times of the generator and discriminator are both set to 1.

3.3. Fine-Tuning of VGGNet-16

Inspired by transfer learning, we fine-tuned a model pre-trained on the ImageNet dataset which contains extensive knowledge of natural images to assist the DEGAN discriminator in improving the classification results [75]. The VGGNet-16 is employed given its wide application in the high-resolution remote sensing image scene classification. There are two ways of fine-tuning VGGNet-16. One is to change the output from the 1000 ImageNet classes to the number of scene categories. The other is to add a classification layer that reduces the output from 1000 to the number of scene categories after the last layer of the model. We chose the latter approach, among which a better classification effect is achieved according to the experimental comparison. The same labeled samples used in DEGAN training are utilized to fine-tune the VGGNet-16 network.

3.4. Training of IFK Codebook and SVM

The $m \times m \times n$ convolutional features can be regarded as n-dimensional local features with the number of $m \times m$. Therefore, these n-dimensional features, which are similar to hand-designed features, are prepared for encoding algorithms such as BOW, IFK, and so on [29]. In this study, we adopt IFK as an encoding algorithm because of its stronger ability to further abstract the features.

The convolutional features are used to train the codebook. Then, the codebook is used to obtain the coding features. Subsequently, the fully connected features and coding features are fused and input into an SVM for classification. The details of the feature extraction and combination are described in Section 3.5. The same training samples used to train DEGAN are utilized for codebook and SVM training.

3.5. Inference the Scene Category

The proposed scene classification method contains two parts, namely feature extraction and scene classification. After training the networks, the testing images are input to the discriminator and the VGGNet-16 to obtain the depth features; then, the features are fused and classified. The details of the proposed method are as follows.

First, the fully connected feature f_{fc-dis} with a size of 384 is extracted from the discriminator. In addition, the $6 \times 6 \times 384$ convolutional features $f_{conv-dis}$ of the 10th convolution layer are also extracted and encoded according to:

$$f_{enc-dis} = ifk(f_{conv-dis}), \tag{15}$$

where $f_{enc-dis}$ represents the features after the encoding, and ifk denotes the IFK coding method used in this paper. Then, the image features are extracted from VGGNet-16: the fully connected features f_{fc-vgg} of the first fully connected layer (with a size of 4096) and the convolutional features $f_{conv-vgg}$ of the 13th convolution layer (with a size of $14 \times 14 \times 512$) are used. Similarly to $f_{conv-dis}$, $f_{conv-vgg}$ is also encoded to $f_{enc-vgg}$ according to Formula (15). Finally, the existing four features are concatenated to form the final image representation as follows:

$$\mathbf{F} = f_{fc-dis} © f_{enc-dis} © f_{fc-vgg} © f_{enc-vgg}. \tag{16}$$

F is input to SVM for the inference of a scene category. © represents the feature concatenation.

4. Experiments

4.1. Experimental Setting

4.1.1. Dataset

To verify the performance of the proposed method, three datasets of high-optical-resolution remote sensing scene images, including UC Merced [37], AID [76], and NWPU-RESISC45 [77], are utilized for the experiment. Figure 6 shows 10 common categories of images from the three datasets.

Figure 6. Illustrations of images of three optical high-resolution remote sensing image scene classification datasets. UC Merced, AID, and NWPU-RESISC45 datasets are displayed sequentially from top to bottom: (**a**) Baseball court; (**b**) Beach; (**c**) Storage tank; (**d**) Forest; (**e**) Harbor; (**f**) River; (**g**) Parking; (**h**) Sparse residual; (**i**) Medium residual; and (**j**) Dense residual.

The first is the UC Merced dataset, which is composed of 21 land-use scene categories that were downloaded from the U.S. Geological Survey USGS National Map Urban Area Imagery. Each category contains 100 scene images with a size of 256 × 256 and a spatial resolution of 0.3 m per pixel. The 21 scene classes include agricultural, airplane, and others. At present, the UC Merced dataset is frequently employed by most remote sensing scene classification algorithms for experimental evaluation.

The second is the AID dataset proposed by a Wuhan University research team in 2017. It includes 10,000 remote sensing images containing 30 scene categories, including airport, bare land, and so on. Each category contains 220–420 images with a size of 600 × 600 and a spatial resolution ranging from 8 meters to 0.5 m per pixel. These images show different countries and regions from the entire world, such as China, the United States, the United Kingdom, France, and Italy. Each type of image was acquired under different time and imaging conditions, which increases the intraclass diversity of the images.

The third is the NWPU-RESISC45 dataset presented by the Northwestern Polytechnical University research team in 2017. It contains 45 scene categories including airplane, airport, and so on. Each scene category contains 700 images with a size of 256 × 256. In addition to the low spatial resolution of the islands, lakes, mountains, and snow mountains, the spatial resolution of most of the test images can reach 30 m per pixel. NWPU-RESISC45 contains 31,500 remote sensing images with rich scene categories, high intraclass diversity, and high similarity among classes, which make it a challenging dataset for remote sensing image scene classification.

Dataset setup: Following the setup of the semi-supervised method [35], each dataset is split into three parts, namely the training set, validation set, and testing set. The training ratios (labeled images) of the three datasets are as follows: 10%, 50%, and 80% for the UC Merced dataset, 10% and 20% for the AID dataset, and 10% for the NWPU-RESISC45 dataset. The validation set and test set are set to 10% when the 80% data are utilized as labeled samples for the UC Merced dataset, and both the validation set and test set are set to 20% in other cases. Apart from the above training, validation, and testing set, the remainder is unlabeled images participating in the training. In addition, we also adopted the unlabeled images from the same category in the two other datasets to train each dataset. For example, the unlabeled images from the same categories of AID and NWPU-RESISC45 are employed during the training of UC Merced.

4.1.2. Evaluation Metric

The evaluation metrics used in this paper include the overall accuracy and the confusion matrix, which are commonly used for scene classification. The overall accuracy is the number of correct samples among all classifications divided by the number of samples in the population. The confusion matrix is used to quantitatively evaluate the degree of confusion between different categories. The rows and columns of the matrix represent the real and predicted scenes, respectively. Any element x_{ij} in the matrix represents the proportion of the number of images for which category i is predicted as category j to the number of test images. The value x_{ij} in the confusion matrix can be calculated as follows:

$$x_{ij} = \frac{n_{ij}}{N_i}. \tag{17}$$

The n_{ij} is the number of images for which category i is predicted as category j, and N_i stands for the total number of test images in the category i.

4.1.3. Implementation Details

In the DEGAN training, the batch size is set to 60, and the learning rates were set to 0.0006 and 0.0003 for the discriminator and generator, respectively. We set the epoch to 600 and the ADAM is adopted to minimize the total loss. For the VGGNet-16 training, the settings are similar to those in [77], in which we set the batch size to 50 and the learning rate to 0.001. The training iteration is set to 15,000. The SGD is employed as an optimizer,

and the weight decay and momentum are set to 0.0005 and 0.9, respectively. As for the IFK coding, the number of Gauss components is set to 8. All the experiments are conducted on a workstation equipped with an Intel(R) XeonE5-2650 v3@2.30 Hz × 20 CPU, an NVIDIA GTX TITAN-XP GPU, 128 GB of memory, and the Pytorch framework.

4.2. Experimental Results

4.2.1. Ablation Study

Compared to the traditional GAN, DEGAN possesses two improvements: one is to adopt a conditional entropy in the discriminator loss such that the unlabeled images can participate in the model training, and the other is to design a DEN to enhance the generator and further promote the discriminator. Therefore, the effectiveness of the unlabeled images and DEN needs to be verified, respectively. In addition, since the prior knowledge of natural images is introduced to the proposed method by fine-tuning the VGGNet-16 for strengthening the classification performance, we also validated the effect of VGGNet-16 for the improvement of classification accuracy.

To this end, we conduct an experimental comparison using the NWPU-RESISC45 dataset at a training ratio of 20%. The baseline is the DEGAN without DEN, and the unlabeled images do not participate in the training. We gradually add the unlabeled images and DEN to the baseline to investigate their influence on the overall classification accuracy. To calculate the classification accuracy, we select the one-dimensional features derived from the fully connected layer of the discriminator, which are then imported into the SVM for classification. Table 2 provides the overall accuracy comparison of different models. It can be seen that the classification accuracy of the discriminator is improved by approximately 6% and 10% by using unlabeled images and adding DEN, which indicates that the DEN further improves the ability of the discriminator while enhancing the generator. DEGAN and the proposed scene classification method based on DEGAN achieve 91.21% and 94.81%, respectively, indicating that the classification performance is improved with the introduction of prior knowledge and the coding of two-dimensional convolution features.

Table 2. Overall accuracy and standard deviations (%) of the GAN without DEN and DEGAN on the NWPU-RESISC45 dataset under the training ratio of 20%.

Methods	Overall Accuracy
Baseline	74.78 ± 0.31
Baseline + unlabeled images	81.41 ± 0.20
Baseline + DEN	83.16 ± 0.17
Baseline + unlabeled images + DEN (DEGAN)	91.21 ± 0.15
Baseline + unlabeled images + DEN + VGGNet-16 (ours)	94.81 ± 0.13

4.2.2. Comparisons with State-of-the-Art Semi-Supervised Methods in Terms of Overall Accuracy

We compare the proposed method with several state-of-the-art semi-supervised methods, including Attention-GAN [59], Self-training [35], Co-training [78], SSGA-E [35], Fixmatch [52], and Mixmatch [48], for which their overall accuracies on three datasets are provided in Tables 3–5, respectively.

From Tables 3–5, we can see that the proposed method achieves the best overall classification accuracy compared to the other semi-supervised comparison methods on three datasets with different training ratios. It is indicated that the proposed scene classification method based on DEGAN is suitable for both small-scale (UC Merced) and large-scale (AID and NWPU-RESISC45) datasets, which significantly improves the classification performance. The classification results on the NWPU-RESISC45 dataset, which are characterized by high intraclass diversity and high similarity among classes, strongly demonstrate the effectiveness of our method. All the comparison results show that the proposed semi-

supervised framework can enhance the ability of scene classification by effectively utilizing labeled and unlabeled training images.

Table 3. Overall accuracy and standard deviations (%) of the proposed method and comparison methods on the UC Merced dataset.

Methods	10% Training Set	50% Training Set	80% Training Set
Attention-GAN [59]	-	89.06 ± 0.50	97.69 ± 0.69
Self-training (ResNet) [35]	-	91.57 ± 2.00	-
Co-training [78]	93.75 ± 1.42	-	-
SSGA-E [35]	94.52 ± 1.38	-	-
Fixmatch [52]	96.22 ± 0.21	-	-
Mixmatch [48]	95.45 ± 0.43	-	-
Our Method	97.89 ± 0.21	98.57 ± 0.24	99.15 ± 0.18

Table 4. Overall accuracy and standard deviations (%) of the proposed method and comparison methods on the AID dataset.

Methods	10% Training Set	20% Training Set
Attention-GAN [59]	-	78.95 ± 0.23
Self-training (ResNet) [35]	-	89.38 ± 0.87
Co-training [78]	90.87 ± 1.08	-
SSGA-E [35]	91.35 ± 0.83	-
Fixmatch [52]	93.63 ± 0.60	-
Mixmatch [48]	92.52 ± 0.48	-
Our Method	94.93±0.21	95.88 ± 0.19

Table 5. Overall accuracy and standard deviations (%) of the proposed method and comparison methods on NWPU-RESISC45 dataset.

Methods	10% Training Set
Attention-GAN [59]	72.21 ± 0.21
Self-training (VGG-S) [35]	81.46 ± 0.68
Self-training (ResNet) [35]	85.82 ± 1.30
Co-training [78]	87.25 ± 0.95
SSGA-E [35]	88.60 ± 0.31
Fixmatch [52]	90.45 ± 0.43
Mixmatch [48]	89.22 ± 0.29
Our Method	**92.23 ± 0.16**

4.2.3. Confusion Matrices

The confusion matrices of the proposed method on three datasets under the training ratios of 10% are given in Figures 7–9. The value on the diagonal of the matrix indicates the proportion of each class classified correctly, and the sum of each row number should be equal to 1. However, since the decimals are rounded when calculating the confusion matrix, the sum of each row is approximately 1. We can make the following observations. From Figure 7, we can see that most categories have high accuracy on the UC Merced data. However, since the medium density residential and density residential scenes have a similar building distribution, they are often confused during classification, which leads to relatively low accuracy. The same phenomenon appears in the AID data. In addition to the above scenes, there are a few other confusing categories due to the similar shapes and structures from Figures 8 and 9, such as palace and church scenes, terrace and rectangular farmland scenes, square and park scenes, desert and bare land scenes, and so on.

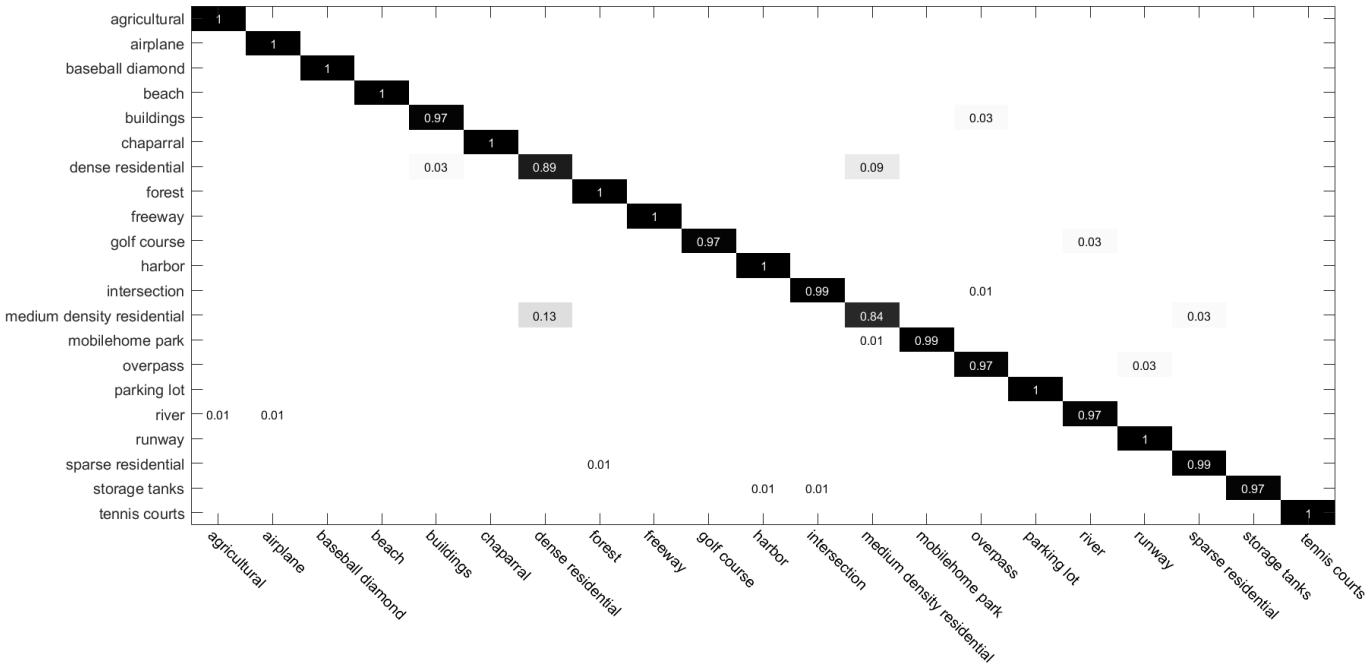

Figure 7. Confusion matrix of the proposed method on the UC Merced dataset under the training ratio of 10%.

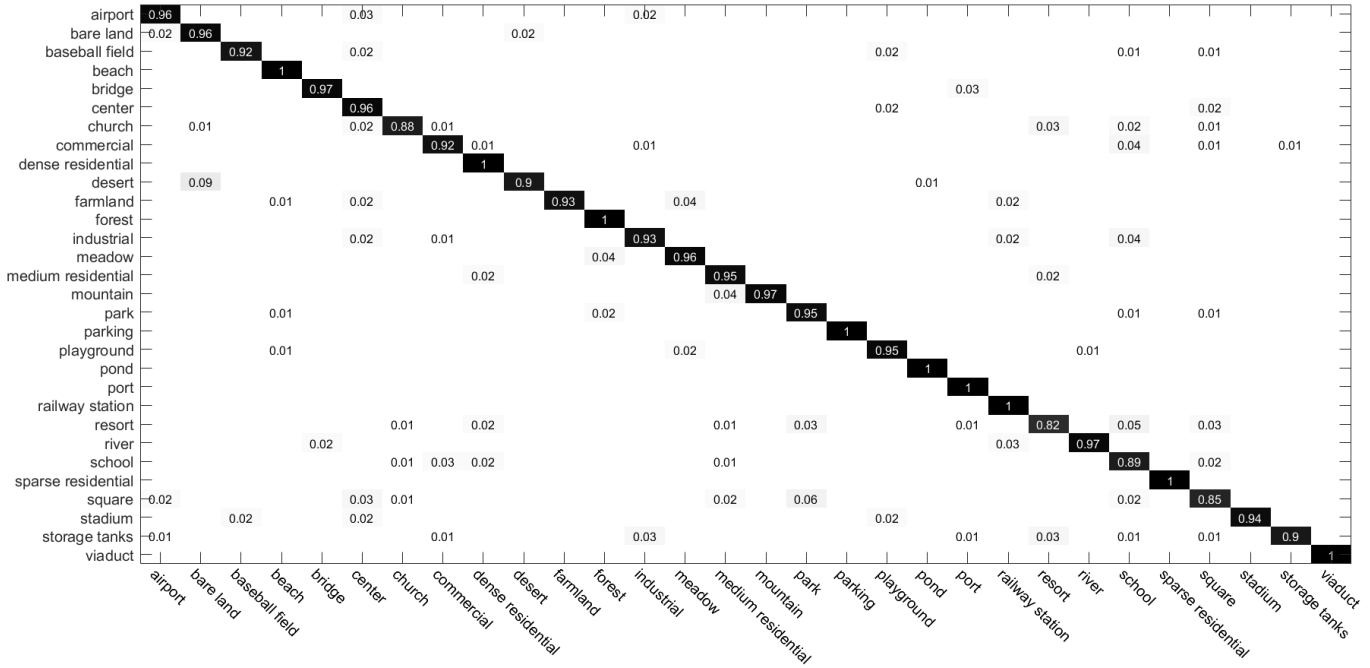

Figure 8. Confusion matrix of the proposed method on the AID dataset under the training ratio of 10%.

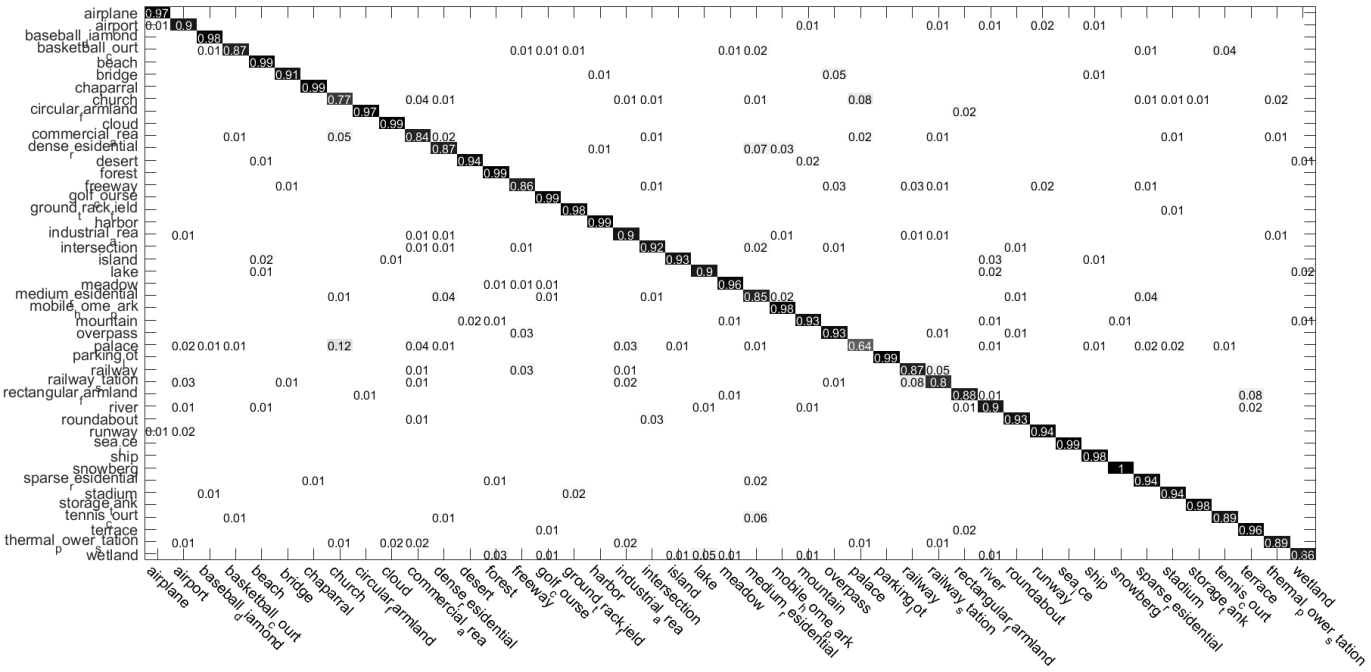

Figure 9. Confusion matrix of the proposed method on the NWPU-RESISC45 dataset under the training ratio of 10%.

4.3. Calculation Time

To analyze the computational efficiency of the proposed scene classification method, we calculate the average training time and inference time on the UCM dataset with other semi-supervised methods. Table 6 shows the comparison results. From Table 6, although the average training time of the proposed method is not minimal, our method processes the fastest inference time compared to other semi-supervised methods.

Table 6. The comparison results of calculation time on the UCM dataset.

	Method	Average Training Time (min)	Inference Time (ms)
Semi-supervised learning	Fixmatch [52]	198	34
	Mixmatch [48]	163	31
	Our Method	173	28

5. Conclusions

In this paper, we propose a novel semi-supervised Diversity Enhanced Generative Adversarial Network (DEGAN) for optical high-resolution remote sensing image scene classification. In the DEGAN, unlabeled and labeled images are jointly utilized to train models by the conditional entropy loss during the feature extraction and classifier learning, in which the experiment results demonstrate that the classification performance of our method outperforms those of other semi-supervised methods. Moreover, the DEN enhances the generator by maximizing the information entropy perspective, which further promotes the discriminative ability of features derived from the discriminator. The employment of the prior knowledge of natural images improves the final classification accuracy by fine-tuning the VGGNet-16 with remote sensing images. In the ablation study, the classification accuracy on the NWPU-RESISC45 dataset is improved by approximately 6%, 10%, and 3% with the utilization of unlabeled data, DEN, and VGGNet-16, respectively. Although the proposed method achieves advanced classification performance compared to other semi-supervised scene classification methods, the unlabeled samples are selected from the public optical remote sensing dataset and the images originating from other sources are

ignored. In a future study, the sampling range of unlabeled scene images must be expanded for improving the classification accuracy.

Author Contributions: Formal analysis, Y.L. and D.Z.; Funding acquisition, X.Q.; Methodology, J.L. and X.Q.; Software, J.L.; Supervision, X.Q.; Writing—original draft, J.L.; Writing—review & editing, J.Z., D.Z. and X.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (62076223).

Data Availability Statement: Public datasets are available at these links: http://weegee.vision.ucmerced.edu/datasets/landuse.html, accessed on 29 July 2022; http://www.lmars.whu.edu.cn/xia/AID-project.html, accessed on 29 July 2022; http://www.escience.cn/people/JunweiHan/NWPU-RESISC45, accessed on 29 July 2022.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Feng, X.; Han, J.; Yao, X.; Cheng, G. TCANet: Triple context-aware network for weakly supervised object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 6946–6955. [CrossRef]
2. Qian, X.; Lin, S.; Cheng, G.; Yao, X.; Ren, H.; Wang, W. Object detection in remote sensing images based on improved bounding box regression and multi-level features fusion. *Remote Sens.* **2020**, *12*, 143. [CrossRef]
3. Yao, X.; Cao, Q.; Feng, X.; Cheng, G.; Han, J. Scale-aware detailed matching for few-shot aerial image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5611711. [CrossRef]
4. Zheng, X.; Wang, B.; Du, X.; Lu, X. Mutual Attention Inception Network for Remote Sensing Visual Question Answering. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5606514. [CrossRef]
5. Li, L.; Yao, X.; Cheng, G.; Han, J. AIFS-DATASET for Few-Shot Aerial Image Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5618211. [CrossRef]
6. Shafique, A.; Cao, G.; Khan, Z.; Asad, M.; Aslam, M. Deep learning-based change detection in remote sensing images: A review. *Remote Sens.* **2022**, *14*, 871. [CrossRef]
7. Zheng, X.; Chen, X.; Lu, X.; Sun, B. Unsupervised Change Detection by Cross-Resolution Difference Learning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5606616. [CrossRef]
8. Manfreda, S.; McCabe, M.F.; Miller, P.E.; Lucas, R.; Pajuelo Madrigal, V.; Mallinis, G.; Ben Dor, E.; Helman, D.; Estes, L.; Ciraolo, G.; et al. On the use of unmanned aerial systems for environmental monitoring. *Remote Sens.* **2018**, *10*, 641. [CrossRef]
9. Hu, T.; Yang, J.; Li, X.; Gong, P. Mapping urban land use by using landsat images and open social data. *Remote Sens.* **2016**, *8*, 151. [CrossRef]
10. Giustarini, L.; Hostache, R.; Matgen, P.; Schumann, G.J.P.; Bates, P.D.; Mason, D.C. A Change Detection Approach to Flood Mapping in Urban Areas Using TerraSAR-X. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 2417–2430. [CrossRef]
11. Gong, M.; Cao, Y.; Wu, Q. A Neighborhood-Based Ratio Approach for Change Detection in SAR Images. *IEEE Geosci. Remote Sens. Lett.* **2012**, *9*, 307–311. [CrossRef]
12. Zhang, Y.; Wang, S.; Wang, C.; Li, J.; Zhang, H. SAR Image Change Detection Using Saliency Extraction and Shearlet Transform. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 4701–4710. [CrossRef]
13. Lee, M.J.; Kang, M.S.; Ryu, B.H.; Lee, S.J.; Lim, B.G.; Oh, T.B.; Kim, K.T. Improved moving target detector using sequential combination of DPCA and ATI. *J. Eng.* **2019**, *2019*, 7834–7837. [CrossRef]
14. Kang, M.S.; Kim, K.T. Compressive Sensing Based SAR Imaging and Autofocus Using Improved Tikhonov Regularization. *IEEE Sens. J.* **2019**, *19*, 5529–5540. [CrossRef]
15. Qian, X.; Zeng, Y.; Wang, W.; Zhang, Q. Co-saliency Detection Guided by Group Weakly Supervised Learning. *IEEE Trans. Multimed.* **2022**. [CrossRef]
16. Qian, X.; Li, J.; Cao, J.; Wu, Y.; Wang, W. Micro-cracks detection of solar cells surface via combining short-term and long-term deep features. *Neural Netw.* **2020**, *127*, 132–140. [CrossRef]
17. Zhang, W.; Tang, P.; Zhao, L. Remote sensing image scene classification using CNN-CapsNet. *Remote Sens.* **2019**, *11*, 494. [CrossRef]
18. Zhang, J.; Zhang, M.; Shi, L.; Yan, W.; Pan, B. A multi-scale approach for remote sensing scene classification based on feature maps selection and region representation. *Remote Sens.* **2019**, *11*, 2504. [CrossRef]
19. Swain, M.J.; Ballard, D.H. Color indexing. *Int. J. Comput. Vis.* **1991**, *7*, 11–32. [CrossRef]
20. Bhagavathy, S.; Manjunath, B.S. Modeling and detection of geospatial objects using texture motifs. *IEEE Trans. Geosci. Remote Sens.* **2006**, *44*, 3706–3715. [CrossRef]
21. Li, F.-F.; Perona, P. A bayesian hierarchical model for learning natural scene categories. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 2; pp. 524–531.

22. Cheng, G.; Yang, C.; Yao, X.; Guo, L.; Han, J. When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2811–2821. [CrossRef]
23. 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. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 3735–3756. [CrossRef]
24. Cheng, G.; Cai, L.; Lang, C.; Yao, X.; Chen, J.; Guo, L.; Han, J. SPNet: Siamese-prototype network for few-shot remote sensing image scene classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5608011. [CrossRef]
25. Cheng, G.; Sun, X.; Li, K.; Guo, L.; Han, J. Perturbation-seeking generative adversarial networks: A defense framework for remote sensing image scene classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5605111. [CrossRef]
26. Ghadi, Y.Y.; Rafique, A.A.; al Shloul, T.; Alsuhibany, S.A.; Jalal, A.; Park, J. Robust Object Categorization and Scene Classification over Remote Sensing Images via Features Fusion and Fully Convolutional Network. *Remote Sens.* **2022**, *14*, 1550. [CrossRef]
27. An, W.; Zhang, X.; Wu, H.; Zhang, W.; Du, Y.; Sun, J. LPIN: A Lightweight Progressive Inpainting Network for Improving the Robustness of Remote Sensing Images Scene Classification. *Remote Sens.* **2021**, *14*, 53. [CrossRef]
28. Lei, T.; Li, L.; Lv, Z.; Zhu, M.; Du, X.; Nandi, A.K. Multi-modality and multi-scale attention fusion network for land cover classification from VHR remote sensing images. *Remote Sens.* **2021**, *13*, 3771. [CrossRef]
29. Cheng, G.; Li, Z.; Yao, X.; Guo, L.; Wei, Z. Remote sensing image scene classification using bag of convolutional features. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1735–1739. [CrossRef]
30. Liu, Q.; Hang, R.; Song, H.; Li, Z. Learning multi-scale deep features for high-resolution satellite image classification. *arXiv* **2016**, arXiv:1611.03591.
31. Tong, W.; Chen, W.; Han, W.; Li, X.; Wang, L. Channel-attention-based DenseNet network for remote sensing image scene classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 4121–4132. [CrossRef]
32. Cao, R.; Fang, L.; Lu, T.; He, N. Self-attention-based deep feature fusion for remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 43–47. [CrossRef]
33. Hong, D.; Gao, L.; Yao, J.; Zhang, B.; Plaza, A.; Chanussot, J. Graph convolutional networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 5966–5978. [CrossRef]
34. Hong, D.; Gao, L.; Yokoya, N.; Yao, J.; Chanussot, J.; Du, Q.; Zhang, B. More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4340–4354. [CrossRef]
35. Han, W.; Feng, R.; Wang, L.; Cheng, Y. A semi-supervised generative framework with deep learning features for high-resolution remote sensing image scene classification. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 23–43. [CrossRef]
36. Dai, X.; Wu, X.; Wang, B.; Zhang, L. Semisupervised scene classification for remote sensing images: A method based on convolutional neural networks and ensemble learning. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 869–873. [CrossRef]
37. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.
38. Bosch, A.; Zisserman, A.; Munoz, X. Scene classification via pLSA. In Proceedings of the European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; Springer: Berlin/Heidelberg, Germany, 2006; pp. 517–530.
39. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
40. Perronnin, F.; Sánchez, J.; Mensink, T. Improving the fisher kernel for large-scale image classification. In Proceedings of the European Conference on Computer Vision, Crete, Greece, 5–11 September 2010; Springer: Berlin/Heidelberg, Germany, 2010; pp. 143–156.
41. Zheng, X.; Sun, H.; Lu, X.; Xie, W. Rotation-Invariant Attention Network for Hyperspectral Image Classification. *IEEE Trans. Image Process.* **2022**, *31*, 4251–4265. [CrossRef]
42. Zheng, X.; Gong, T.; Li, X.; Lu, X. Generalized Scene Classification From Small-Scale Datasets With Multitask Learning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5609311. [CrossRef]
43. Hu, F.; Xia, G.S.; Hu, J.; Zhang, L. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. *Remote Sens.* **2015**, *7*, 14680–14707. [CrossRef]
44. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
45. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*. Available online: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf (accessed on 29 July 2022). [CrossRef]
46. Wang, Q.; Liu, S.; Chanussot, J.; Li, X. Scene classification with recurrent attention of VHR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 1155–1167. [CrossRef]
47. Chaib, S.; Liu, H.; Gu, Y.; Yao, H. Deep feature fusion for VHR remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 4775–4784. [CrossRef]
48. Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; Raffel, C.A. Mixmatch: A holistic approach to semi-supervised learning. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 5049–5059.
49. 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.* **2018**, *41*, 1979–1993. [CrossRef]

50. Grandvalet, Y.; Bengio, Y. Semi-supervised learning by entropy minimization. *Adv. Neural Inf. Process. Syst.* **2004**, *17*. Available online: https://proceedings.neurips.cc/paper/2004/file/96f2b50b5d3613adf9c27049b2a888c7-Paper.pdf (accessed on 29 July 2022).
51. Lee, D.H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In Proceedings of the Workshop on Challenges in Representation Learning, ICML, Daegu, Korea, 3–7 November 2013; Volume 3, p. 896.
52. Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C.A.; Cubuk, E.D.; Kurakin, A.; Li, C.L. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 596–608.
53. Tian, Y.; Dong, Y.; Yin, G. Early Labeled and Small Loss Selection Semi-Supervised Learning Method for Remote Sensing Image Scene Classification. *Remote Sens.* **2021**, *13*, 4039. [CrossRef]
54. Cheng, G.; Han, J.; Guo, L.; Liu, T. Learning coarse-to-fine sparselets for efficient object detection and scene classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1173–1181.
55. Han, X.; Zhong, Y.; Zhao, B.; Zhang, L. Scene classification based on a hierarchical convolutional sparse auto-encoder for high spatial resolution imagery. *Int. J. Remote Sens.* **2017**, *38*, 514–536. [CrossRef]
56. Cheng, G.; Zhou, P.; Han, J.; Guo, L.; Han, J. Auto-encoder-based shared mid-level visual dictionary learning for scene classification using very high resolution remote sensing images. *Comput. Vis. IET* **2015**, *9*, 639–647. [CrossRef]
57. Yao, X.; Han, J.; Cheng, G.; Qian, X.; Guo, L. Semantic Annotation of High-Resolution Satellite Images via Weakly Supervised Learning. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 3660–3671. [CrossRef]
58. Lin, D.; Fu, K.; Yang, W.; Xu, G.; Xian, S. MARTA GANs: Unsupervised Representation Learning for Remote Sensing Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 2092–2096. [CrossRef]
59. Yu, Y.; Li, X.; Liu, F. Attention GANs: Unsupervised Deep Feature Learning for Aerial Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 519–531. [CrossRef]
60. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014.
61. Radford, A.; Metz, L.; Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv* **2015**, arXiv:1511.06434.
62. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein GAN. *arXiv* **2017**, arXiv:1701.07875.
63. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A.C. Improved Training of Wasserstein GANs. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
64. Mao, X.; Li, Q.; Xie, H.; Lau, R.; Smolley, S.P. Least Squares Generative Adversarial Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
65. Qi, G.J. Loss-sensitive generative adversarial networks on lipschitz densities. *Int. J. Comput. Vis.* **2020**, *128*, 1118–1140. [CrossRef]
66. Zhao, J.; Mathieu, M.; LeCun, Y. Energy-based generative adversarial network. *arXiv* **2016**, arXiv:1609.03126.
67. Berthelot, D.; Schumm, T.; Metz, L. Began: Boundary equilibrium generative adversarial networks. *arXiv* **2017**, arXiv:1703.10717.
68. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784.
69. Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2016**, *29*. Available online: https://proceedings.neurips.cc/paper/2016/file/7c9d0b1f96aebd7b5eca8c3edaa19ebb-Paper.pdf (accessed on 29 July 2022).
70. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of theIEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
71. Choi, Y.; Choi, M.; Kim, M.; Ha, J.W.; Choo, J. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. *arXiv* **2018**, arXiv:1711.09020.
72. Dai, Z.; Almahairi, A.; Bachman, P.; Hovy, E.; Courville, A. Calibrating energy-based generative adversarial networks. *arXiv* **2017**, arXiv:1702.01691.
73. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved techniques for training gans. *Adv. Neural Inf. Process. Syst.* **2016**, *29*. Available online: https://proceedings.neurips.cc/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf (accessed on 29 July 2022).
74. Springenberg, J.T. Unsupervised and semi-supervised learning with categorical generative adversarial networks. *arXiv* **2015**, arXiv:1511.06390.
75. Zhu, Q.; Zhong, Y.; Zhang, L.; Li, D. Adaptive deep sparse semantic modeling framework for high spatial resolution image scene classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 6180–6195. [CrossRef]
76. Xia, G.S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [CrossRef]
77. Gong, C.; Han, J.; Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proc. IEEE* **2017**, *105*, 1865–1883.
78. Ning, X.; Wang, X.; Xu, S.; Cai, W.; Zhang, L.; Yu, L.; Li, W. A review of research on co-training. *Concurr. Comput. Pract. Exp.* **2021**, e6276. [CrossRef]