Supervised Generative Adversarial Network Based Sample Generation for Scene Classification

Wei Han[12], Ruyi Feng[12], Lizhe Wang[12] and Jia Chen[12]
[1]School of Computer Science, China University of Geosciences, Wuhan, 430074, P.R.China.
[2] Hubei Key Laboratory of Intelligent Geo-Information Processing, China University of Geosciences, Wuhan 430074, P.R.China.

***Abstract*—High-resolution remote sensing (HRRS) image scene classification has been a critical task and greatly important for many applications, wherein convolutional neural network (CNN)-based methods have achieved considerable improvements. However, the CNN-based methods have countered a severe problem that massive annotation samples are required to obtain ideal model for scene classification. There is no dataset with a comparative scale to ImageNet to meet the sample requirement and labelling samples is labor-intensive and time-consuming. To solve the problem of insufficient annotation samples, a new generative adversarial network (GAN)-based sample generation method for scene classification is implemented. The proposed method is able to generate HRRS images with specific label and improve scene classification performance for the CNN-based methods.**

***Index Terms*—Deep Learning, High-Resolution Remote Sensing, Scene Classification.**

I. INTRODUCTION

With the rapid development of Earth observation technologies, massive high-resolution remote sensing (HRRS) images are obtained for military and civilian applications. HRRS scene classification, which focuses on the smart classification of land cover and land use (LCLU) according to the image content, has been a fundamental task and has attracted extensive attention. Generally, remote sensing scenes are composed of a lot kinds of diverse objects, including building, tree, roads soil, which construct a complicatedly spatial pattern and serve as a functional zone. Some objects or LCLU classes are generally shared among different scene categories. Due to a large number kinds of objects and complex patterns, scene classification becomes a fairly challenging problem.

In the past few decades, much effort has been made to promote different recognition models and feature representation method for HRRS scene classification. The bag-of-visual-word (BOW)-based method [1] are one kind of notable methods to solve the challenge of scene classification. The performance of the BOW based methods strongly relies on the handcrafted features, e.g., texture feature, color histogram and local structural. As image resolution and data volume of HRRS images increases, HRRS images show more details of objects and more complex patterns. The handcrafted features meet a severe problem that they are limited to effectively and efficiently represent massive HRRS images. Driven by the development of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [2], deep learning methods such as convolutional neural networks (CNNS), have achieved great improvements and have been applied in image classification [3], [4].

Owing to the impressive representation capability of deep learning features, different deep-learning-feature-based methods have been applied to remote sensing image scene classification [5], [6], [7], [8]. But, the deep-learning-based methods have countered a problem that they have to learn from massive annotation data to achieve ideal performance. There is no HRRS dataset with a comparative scale to ImageNet [2] that can meet the requirements of training samples for the deep learning methods. Developing an effective model working with limited annotation samples is an available solution to address the problem. Liu *et al.* [9] utilized weakly labeled images as input and triplet network to address limitation of clearly labeled datasets. Cheng and Han *et al.* [10] used metric learning to reduce the requirement of labeled samples.

High-resolution sample generation is a different approach to address the problem of insufficient annotation samples. It has been long-standing challenge in machine leaning [11], and has been greatly boosted by the development of generative adversarial networks (GANs) [12]. Defining a game between two competing networks, a GAN composes a generator and a discriminator. The generator is trained to fool the discriminator, and the discriminator acts to enhance the recognition ability. Recently, GANs have also been used in hyperspectral image classification for feature extraction [13], [14].

In the light of the excellent performance of GANs in image generation, and to address the problem of insufficient training samples, a GAN-based high-resolution remote sensing image generation method is proposed in this paper. The proposed method is able to generate high-resolution images with consistent label information. In the proposed method, Wasserstein distance [15] is a more accurate metric than the Jensen-Shannon (JS) divergence to estimate the difference between the generator distribution and the real data distribution. Supervised label information [11] is also employed to guide the proposed method to produce images with consistent distinct label, as well as to accelerate the convergence.

The remainder of this paper is organized as follows: In Section 2, we introduce the details of the proposed method. The experiments and analysis are conducted in Section 3. Conclusions are drawn in Section 4.

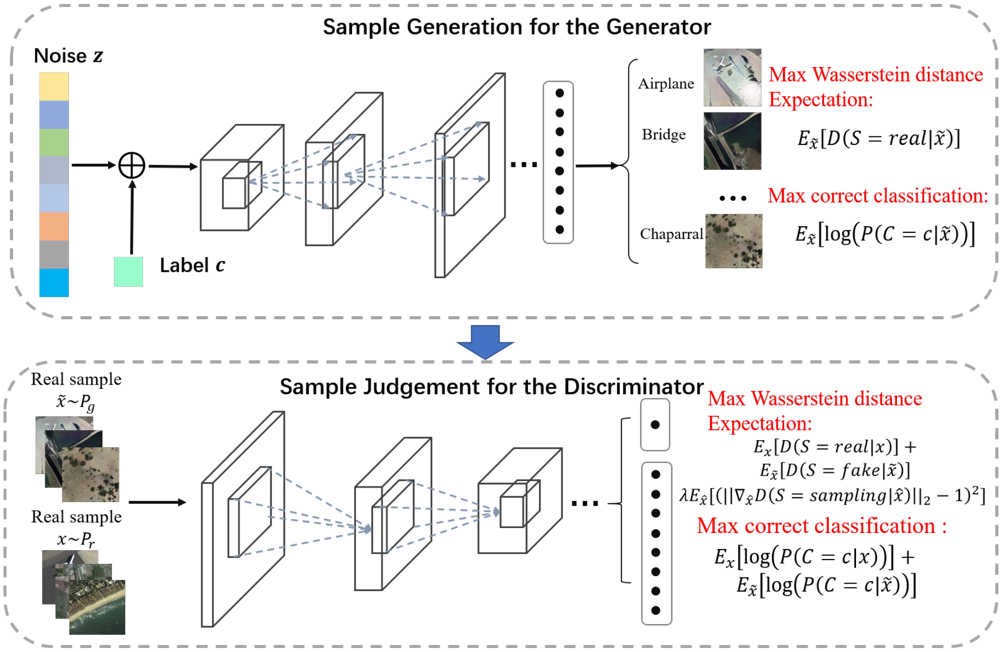

Fig. 1: Flowchart of the proposed method.

II. Methodology

To generate high-resolution and annotated remote sensing images, the Wasserstein distance is used to measure the difference between the real data distribution and the fake data distribution. Then, superivsed category information is used to guide to generated samples with class label. The main flowchart of the proposed method is shown in Fig. 1. It contains two components: sample generation for the generator and sample judgement for the discriminator. In the first component: a set of noise vectors z is sampled from a spherical Gaussian distribution. Category samples c generated from a random seed algorithm is added to z as the input of the generator. Next, the input vectors are processes by a set of deconvolutional layers to generate high-resolution samples $\tilde{x}$ with specific labels. In the second component: the real samples x, and the generated smaples $\tilde{x}$ are feed into the discriminator. The discriminator have two branches to carry out two different predictions: one for the sample source ($fake$ or $real$) and the other for the category for the input samples. The networks for the generator and the discriminator are implemented based on convolutional layers and the generated samples are with a resolution of 128 $\times$ 128. The flowchart of the proposed method can be summarized as follows:

1) z and c are fed into the generator to generate fake samples $\tilde{x}$.
2) The real data x and the fake data $\tilde{x}$ are jointly fed to the discriminator. The discriminator judges the data source, and predicts the category labels of the input samples by additionally adding an auxiliary classifier to the discriminator.
3) The source loss and category likelihood loss are calculated, respectively. The discriminator is update based on the back propagation algorithm.
4) The generator is also improved according to the loss of fake samples.

The original GAN [12] used JS divergence to measure the difference of the real data and the generated samples. It causes the problems of gradient vanishing and mode collapse. To overcome the problems, Wasserstein distance [15], [16] is utilized to more accurately measure the difference of two distributions, which can always provide a meaningful gradient for updating the networks. The model objective function of Wasserstein GAN (WGAN) is defined as:

$$\min_G \max_D \ \mathbb{E}_{x\sim\mathbb{P}_r}\,[D(x)] - \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}\,[D(\tilde{x})] + \lambda\mathbb{E}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}[(\| \nabla_{\hat{x}}\, D(\hat{x})\|_2 - 1)^2] \tag{1}$$

where $x \sim \mathbb{P}_r$ is the real data distribution, $\tilde{x} \sim \mathbb{P}_g$ is the fake data distribution, $\lambda\mathbb{E}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}[(\| \nabla_{\hat{x}}\, D(\hat{x})\|_2 - 1)^2]$ is denoted as the gradient penalty. $\hat{x} \sim \mathbb{P}_{\hat{x}}$ is implicitly defined to sample uniformly along straight lines between pairs of points sampled from the real data distribution $\mathbb{P}_r$ and the generator distribution $\mathbb{P}_g$. On the sampling points, $\nabla_{\hat{x}} D(\hat{x})$ is used as the gradient norm, and is constrained to 1. By adding the gradient penalty, the objective function satisfies the limitation of the Lipschitz continuity, and the advantages of the Wasserstein distance can be made available. Empirically, the penalty coefficient λ is set to 10 [16].

To exploit supervised learning for GANs, some works has applied label information to promote training stability of models and generated samples with category information, the auxiliary classifier generative adversarial network (ACGAN)

[11]. An auxiliary classifier is added to the discriminator of ACGAN. Therefore, the discriminator has two branches to simultaneously predict the category and the source of input images. By doing this, supervised category loss is added into the objective function, which guides ACGAN to achieve the generated sample with class information.

We applied the auxiliary classifier to WGAN for generating high-resolution and annotated remote sensing, which combines the advantages of WGAN and ACGAN. For the discriminator, the objective function is presentaed as follows:

$$\begin{aligned} \max\ \mathbb{E}_{x\sim\mathbb{P}_r}\ [P(S=real|x)] + \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}\ [P(S=fake|\tilde{x})] + \\ \mathbb{E}_{x\sim\mathbb{P}_r}[log(P(C=c|x))] + \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}[log(P(C=c|\tilde{x}))] + \\ \lambda\mathbb{E}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}[(\|\bigtriangledown_{\hat{x}} D(S=smapling|\hat{x})\|_2 - 1)^2] \end{aligned} \tag{2}$$

where $P(S|X), P(C|X) = D(X)$ indicates a probability distribution over the sources and a probability distribution over the category labels. $(\|\bigtriangledown_{\hat{x}} D(S=sampling|\hat{x})\|_2 - 1)^2$ is the gradient penalty for the discriminator to meet the Lipschitz continuity, $\lambda = 10$. The objective for the discriminator is to maximize the expectation for correctly judging the source and category of the input samples.

For the generator, the objective function is shown as follows:

$$\max \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}[log(P(C=c|\tilde{x}))] + \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}\ [P(S=real|\tilde{x})] \tag{3}$$

The objective function of the generator is to maximize the expectation of the possibility that fake data is judged to be real, and the possibility that the categories of fake data are classified correctly.

III. EXPERIMENTS AND ANALYSIS

A. Dataset Introduction and Experimental Settings

The proposed method was evaluated on the UCM dataset [17]. It contains a total of 2,100 images with 256×256 pixels, which are manually labeled as belonging to 21 land-use classes, with 100 images for each class. The dataset were randomly split into the training set and the test set, as three settings: 1) 20% for training and 80% for testing; 2) 40% for training and 60% for testing; and 3) 60% for training and 40 for testing. To match the proposed method, all the images were resized to 128 × 128. To evaluate the scene classification results, a CNN model was implemented based on VGGNet [4], which is 10-layer network structure (seven convolutional layers and three fully connected layers), denoted as VGG-10. The training settings of the proposed method are that the batch size is set to 64 with a learning rate of 0.0002, and the number of the training iteration is 100,000.

The experiments were organized into two parts: 1) Evaluation of the degree of training convergence by Wasserstein distance. 2) Evaluation of the effect of the generative samples on the classification accuracy by overall accuracy (OA).

B. Experimental Results And Analysis

1) Evaluation of the degree of training convergence: Fig.2 shows the relationship between the image quality, the Wasserstein distance, and the iteration number. In initially, the Wasserstein distance between the generator distribution and the real data distribution is large more than 200. As the iteration number increases, the Wasserstein distance dramatically drops. After 25% of the training iterations, the curves of Wasserstein distance becomes stable. The value converges to about 45 over 20% training set, 17 over 40% training set and 7 over 60% of training set (see Fig. 2 (1-3)). With regard to the generative samples, they are at fist confusing and do not contain class information. As the training iterations increases, the image quality stably improve and the class information is more specific. The experimental results prove that the proposed method is able to generate HRRS images with consistent category information, and the model achieves a high degree of convergence.

2) Evaluation the effect of the generative samples on classification accuracy: By fixing the original training set and increasing the number of generative smaples as an extended training set, the change of the test accuracy of VGG-10 was obseved. The variable is the ratio of generative samples to original samples. For instance, when the ratio is set to 2, there are 20 real samples per class and there are 40 generative samples per class. It should be noted that the proposed sample generation method was trained on the same training set as VGG-10. Therefore, the generative samples were not affected by the test samples.

As shown in Table I, we can see that the generative samples help to improve the classification accuracy of VGG-10 with the UCM dataset, by about 6% with the 20% training set, 1.9% with the 40% training set, and 1.7% with the 60% training set. The positive effect on the classification improvement of the generative samples is clearly more evident in the case of a smaller training set. One possible reason for this is that, in the case of sufficient training samples, the generative samples cannot provide more beneficial information to promote the generalization ability. Meanwhile, increasing the number of generative samples does not always improve the accuracy of the model.

IV. CONCLUSIONS

In this paper, focusing on the problems of the insufficient annotation datasets in remote sensing community, a new GAN-based sample generation method for scene classification has been presented. The proposed method is able to generate high-resolution remote sensing images with specific label and improve performance for the CNN-based method scene classification.

V. ACKNOWLEDGMENT

We gratefully acknowledged acknowledge the support of the National Natural Science Foundation of China (No.41571413, No. 41701429 and No. U1711266); the Fundamental Research Funds for the Central Universities, China University of Geosciences (Wuhan) (No. CUG170625); the Open Research

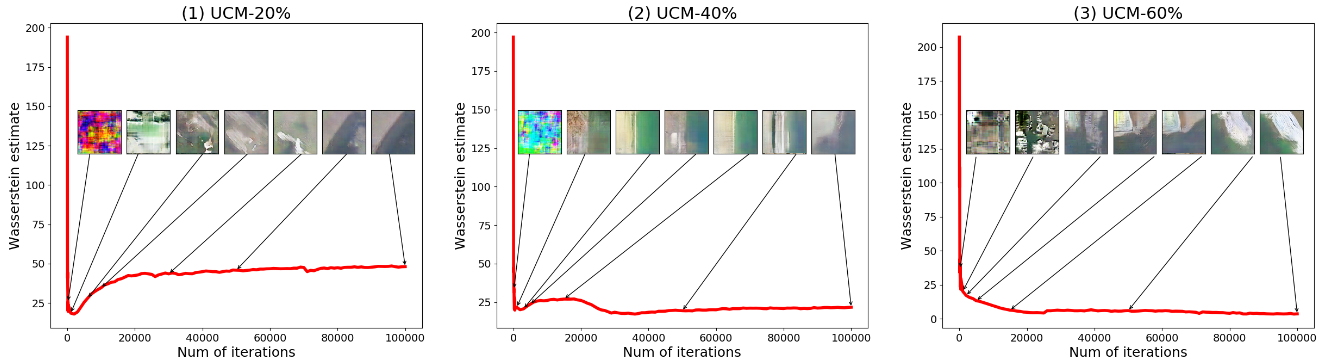

Fig. 2: Wasserstein estimation for the proposed method on the UCM datasets. The x-axis is the number of iterations and the y-axis is the value of Wasserstein distance.

TABLE I: **Evaluation of the effect of the generative samples on the classification accuracy**

Ratio of Lableled Set	**Classification Ratio (%)**					
	0	1	2	3	4	5
UCM-20%	66.15±1.45	68.95±1.85	73.59±0.35	72.5±0.89	71.37±0.98	72.22±0.59
UCM-40%	76.93±0.54	79.84±1.27	77.62±0.95	77.58±0.28	77.94±0.09	76.18±0.62
UCM-60%	82.75±0.28	83.55±0.70	83.56±0.23	83.87±0.07	84.52±0.72	84.46±0.6

Project of The Hubei Key Laboratory of Intelligent Geo-Information Processing (KLIGIP-2017B08); the Open Research Fund of Key Laboratory of Spectral Imaging Technology, Chinese Academy of Sciences (Grant No. LSIT201716D); the Fundamental Research Founds for National University, China University of Geosciences (Wuhan).

References

[1] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *9th IEEE International Conference on Computer Vision (ICCV 2003), 14-17 October 2003, Nice, France*, 2003, pp. 1470–1477.

[2] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, 2009, pp. 248–255.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, 2012, pp. 1106–1114.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," vol. abs/1409.1556, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[5] Y. Zhong, X. Han, and L. Zhang, "Multi-class geospatial object detection based on a position-sensitive balancing framework for high spatial resolution remote sensing imagery," *ISPRS J. Photogram. Remote Sens.*, vol. 138, pp. 281–294, 2018. [Online]. Available: https://doi.org/10.1016/j.isprsjprs.2018.02.014

[6] Y. Zhong, A. Ma, Y. Ong, Z. Zhu, and L. Zhang, "Computational intelligence in optical remote sensing image processing," *Appl. Soft Comput.*, vol. 64, pp. 75–93, 2018. [Online]. Available: https://doi.org/10.1016/j.asoc.2017.11.045

[7] G. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017. [Online]. Available: https://doi.org/10.1109/TGRS.2017.2685945

[8] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017. [Online]. Available: https://doi.org/10.1109/JPROC.2017.2675998

[9] Y. Liu and C. Huang, "Scene classification via triplet networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 1, pp. 220–237, 2018.

[10] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative cnns," *IEEE Trans. Geoscience and Remote Sensing*, vol. 56, no. 5, pp. 2811–2821, 2018. [Online]. Available: https://doi.org/10.1109/TGRS.2017.2783902

[11] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, 2017, pp. 2642–2651.

[12] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 2014, pp. 2672–2680.

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

[14] D. Lin, K. Fu, Y. Wang, G. Xu, and X. Sun, "MARTA gans: Unsupervised representation learning for remote sensing image classification," *IEEE Geosci. Remote Sensing Lett.*, vol. 14, no. 11, pp. 2092–2096, Oct. 2017.

[15] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," *CoRR*, vol. abs/1701.07875, 2017. [Online]. Available: http://arxiv.org/abs/1701.07875

[16] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, 2017, pp. 5769–5779.

[17] Y. Yang and S. D. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *18th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems, ACM-GIS 2010, November 3-5, 2010, San Jose, CA, USA, Proceedings*, 2010, pp. 270–279.