Few-Sample Generation of Amount in Figures for Financial Multi-Bill Scene Based on GAN

Zhi-Ri Tang, *Student Member, IEEE*, Qi-Qi Chen, Zhao-Hui Sun, *Member, IEEE*, Pengwen Xiong, *Member, IEEE*, Bao-Hua Zhang, Lu Jiang, and Edmond Q. Wu, *Senior Member, IEEE*

***Abstract*— Recognition of amount in figures in the financial multi-bill scenes is crucial for the automatic banking business. However, the diversity of banking business and the limitation of customer data privacy determine that it is difficult to collect a large number of sample datasets. Aiming at the problem of insufficient training data in multi-bill scenes and the low accuracy of the detection model, this article proposes a new generative adversarial network (GAN) to generate new samples and to expand the bill dataset, which is then adopted to train a framework for recognition of the bill amount. In the proposed WGAN-SA, a residual block is adopted as the basic structure of the generator and the discriminator, and the self-attention mechanism is also utilized to improve the generation performance. In addition, Wasserstein distance is utilized to measure the distance between real and synthetic samples. Experimental results on the benchmark dataset and comparisons with state-of-the-art works show that our proposed WGAN-SA can effectively improve the few-sample learning performance. Besides, experiments on the bill dataset verify that our method can solve the problem of model collapse and has the ability to generate images of the amount in figures with better fidelity and variety, which is also helpful to achieve better bill amount recognition performance compared with other latest works.**

***Index Terms*— Financial scenes, generative adversarial network (GAN), residual block, self-attention mechanism, Wasserstein distance.**

NOMENCLATURE

Math Symbol	*Explanation*
G	Generator.
D	Discriminator.
BN	Batch normalization.
$p_{\hat{x}}$	Uniform sampling of the pair points.
p_{real}	Real distribution.
p_z	Generated distribution.
$\mathbf{z}$	Latent variables sampling from p_z.
L	Loss function.
∇	Gradient.
$\|\|\ \|\|_2$	Euclidean norm.
λ	Gradient penalty term.

Manuscript received July 12, 2021; revised November 8, 2021; accepted December 14, 2021. This work was supported by the National Natural Science Foundation of China under Grant 72192822, Grant 72192820, Grant 62171274, Grant U1933125, Grant 62163024, and Grant 61903175. *(Corresponding author: Zhao-Hui Sun.)*

Zhi-Ri Tang is with the School of Physics and Technology, Wuhan University, Wuhan 430072, China (e-mail: gerintang@163.com).

Qi-Qi Chen and Edmond Q. Wu are with the Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: cqq.jx@sjtu.edu.cn; edmondqwu@sjtu.edu.cn).

Zhao-Hui Sun is with the School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: zh.sun@sjtu.edu.cn).

Pengwen Xiong is with the School of Information Engineering, Nanchang University, Nanchang 330031, China (e-mail: steven.xpw@ncu.edu.cn).

Bao-Hua Zhang and Lu Jiang are with the Industrial and Commercial Bank of China, Shanghai 200120, China (e-mail: zhangbh@sdc.icbc.com.cn; jiangl@sdc.icbc.com.cn).

Digital Object Identifier 10.1109/TCSS.2021.3136602

I. INTRODUCTION

BANKS always have a large number of bill businesses, which all involve the amount in figures. To construct an automatic banking business and replace manual input, some artificial intelligence models have been applied, which can accurately detect and identify the amount in figures of bills. However, the difficulty in identifying the amount in figures at present lies in the lack of training samples, especially in the overseas bill business of banks. Because the bill business of different overseas institutions is different and there are some restrictions such as customer privacy, there are not enough annotations to support the training of optical character recognition models. Therefore, some effective methods are needed to solve the problem of few samples in financial multi-bill scenes.

Currently, kinds of different solutions are available to solve the problem of few-sample learning. First, it is possible to improve the model itself to make it more suitable for few-sample situations, for example, Santoro *et al.* [1] proposed a memory enhancement method to improve the performance, which adjusts the deviation by weight updating and adjusts the output by caching the expression quickly to memory. Furthermore, Siamese network [2] was presented to train a twin network in a supervised way and then reused the features extracted from the network for few-sample learning. Data augmentation by some transform methods is also a solution. However, it usually leads to a high sample repetition rate and overfitting problems of the model in the training process because no real new data are adopted basically through data augmentation. Furthermore, due to the data private and remote authorization in the financial scene, restrictions on access to data or labels always limit the performance of these works.

In recent years, generating new samples by some generative models is a popular choice for the few-sample problem. Generative adversarial network (GAN) [3] is exactly a powerful generative model, which has the ability to generate

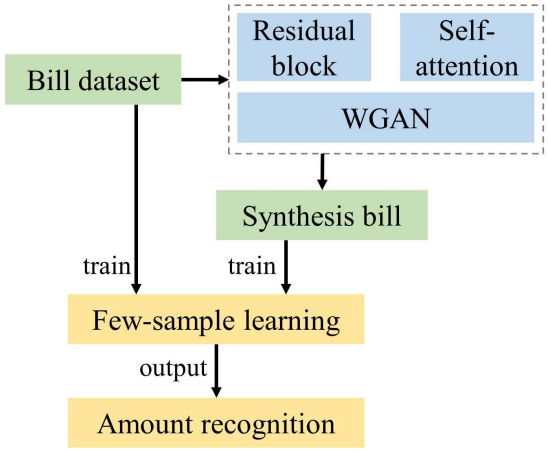

Fig. 1. Overview of the proposed framework.

realistic and high-resolution samples. The core idea of GAN is derived from the Nash equilibrium of game theory. The initial GAN model was proposed in 2014, which is composed of two models: generator G and discriminator D. The generator receives random noise and generates data samples, while the discriminator classifies the generated data samples and the real samples. The purpose of the discriminator is to accurately distinguish synthetic samples from real samples, while the purpose of the generator is to learn the distribution of real data. Hence, it can generate samples with a similar distribution with real samples and make the discriminator unable to accurately judge whether the sample is real or generated.

Although GAN is a good model for new sample generation, the training process is extremely unstable, and model collapse is easy to occur, resulting in the generation of meaningless samples. To solve this problem, Wasserstein distance is applied in [4] instead of Jensen–Shannon divergence to calculate the distance between the real and the generated distributions. It provides a meaningful gradient for two non-overlapping distributions in the high-dimensional space and makes model training more stable. However, there is no approach that works very well under the remote authorization scene and few-sample scenes. Hence, how to adopt a Wasserstein distance-based GAN in the financial multi-bill scene is still a challenge.

Inspired by the above, this work proposes a few-sample generation of the amount in figures for financial multi-bill scene named WGAN-SA, where an overview of the framework is shown in Fig. 1. Some corresponding improvements are also adopted to improve the generation performance and construct the WGAN-SA model. Then the proposed method is verified by both benchmark dataset and private bill dataset, where the experimental results show that our method can achieve better performance compared with other state-of-the-art works. The main contributions of this work can be summarized as follows.

1) A WGAN-SA model with residual blocks and the self-attention mechanism is adopted for both the generator and the discriminator, aiming to improve the generation performance with few training samples and solve the model collapse problem.
2) The generated cheque fragments with the amount in figures are then adopted to train an amount recognition model, aiming to improve the few-sample learning performance.
3) Generation performance on the bill dataset shows that the proposed method can synthesize images with higher resolution and diversity compared with other generation models. In addition, experimental results on both benchmark and bill datasets are presented to show the performance of few-sample learning benefiting from the proposed method. Numerous comparisons with state-of-the-art works verify the efficiency of the proposed method.

The rest of this article is arranged as follows. Section II presents a review of related works. Model descriptions and corresponding experimental results are given in Sections III and IV, respectively. Conclusions and discussions on future works are proposed in Section V.

II. Related Work

A. Few-Sample Learning

Due to the application value of few-sample learning in practical engineering, many few-sample learning methods have emerged in recent years. Besides image processing and computer vision, it has also been applied in many other areas, such as data service [5] and blockchain [6]. Prototype network [7], which is based on the prototype representation for each category, presents the mean value of the support set in the embedded space. Through that process, it turns the classification problem into the nearest neighbor problem in the embedded space to solve the few-sample problem. Sung *et al.* [8] model the metrics for measuring distance, so it trains a network to learn the metrics for distance rather than using a single and fixed distance measurement mode. Ravi and Larochelle [9] focus on reasons why the gradient-based optimization algorithm failed under a situation of a few data, while Finn *et al.* [10] propose a method that enables the model to obtain better generalization performance with a small number of iterative steps on a few samples. However, these methods still need to be further studied in the aspects of the mechanism of the fusion of prior knowledge and machine learning algorithms as well as the extraction and encapsulation of specific empirical knowledge. Adopting deep transfer learning techniques is also a solution. For example, MTL [11] is proposed to adopt a deep network to a few sample learning by adjusting it. Furthermore, there are also some other works, such as natural image identification [12], seam detection [13], spatial–temporal attention [14], semi-supervised pedestrian reidentification [15], saliency detection [16], and group recommendation [17]. However, this kind of method requires a large auxiliary dataset in the relevant fields as the target domain. For remote authorization scenes without a corresponding annotated dataset, it is still not a feasible way.

In addition, some data augmentation methods can be applied to the original dataset to expand the number of data and annotations in datasets, which is also a feasible way to handle the few-sample problem. Some works [18], [19] propose some

data augmentation methods in deep learning models, such as random cropping, translation, changing image contrast, and so on. Although these methods can solve the problem of few samples to some extent, the sample repeatability after data augmentation may be too high, leading to overfitting issues easily during the training process. Hence, the generation of new data needs to be developed.

B. Generative Adversarial Networks

Another feasible way to solve the few-sample problem is to synthesize new samples through some generative models. Kingma and Welling [20] propose a variational auto-encoder (VAE) model, which forces the encoder to generate latent variables that obey unit Gaussian distribution by adding constraints to it. After sampling the probability distribution of the latent variables, the decoder can utilize them to generate new samples. This model can effectively learn the distribution of the training set and map the original probability distribution to the probability distribution of the training set, so it can generate samples that have a similar distribution with the training samples. However, due to the lack of a discriminative network in VAE, the generated image cannot retain the original image's clarity, which tends to generate nebulous images. Goodfellow *et al.* [3] propose another deep generative model named GAN, which makes the distribution of inputs fit the distribution of real samples as much as possible through adversarial learning for the first time. This kind of model can generate more realistic and high-resolution samples.

Since the initial GAN was proposed in 2014. Many recent studies have focused on improving the GAN model to improve the stability of the model training. Nowozin *et al.* [21] and Fedus *et al.* [22] have introduced their own theories to explain that GANs can study distribution by diffusing minimum angle to achieve the Nash equilibrium. Che *et al.* [23] propose two regularizers to improve the model stability and solve the problem of mode collapse. Some studies promote the convergence of GANs by improving the loss function. LSGAN [24] utilizes the least-square loss function to replace the cross-entropy loss function on the discriminator to weaken the gradient dispersion and generate images with better quality than the original GAN. [25] explains the reasons for the training instability of GAN and provides methods to avoid these problems. WGAN-GP [4] and [26] propose Wasserstein distance to calculate the distance between the distribution of generated samples and the distribution of real samples instead of Jensen–Shannon divergence. It provides a meaningful gradient for two non-overlapping distributions in the high-dimensional space, making model training more stable and solving model collapse. Bellemare *et al.* [27] propose the Cramer distance, which combines the advantages of Wasserstein distance and KL divergence and has a better performance in the training process. Both OT-GAN [28] and EBGAN [29] introduce the energy function to improve training stability.

Apart from the basis image generation, GAN has also been applied in many other areas [30], [31]. Zhang *et al.* [32] present a human face sketch model based on edge optimization and GAN. Some researchers adopt GAN into image super-resolution [33], [34], while others apply it to semantic segmentation [35], facial attribute editing [36], pedestrian detection [37], and image re-raining [38].

In addition, many studies focus on structure modification to improve the quality of the generated images. Since the original GAN is unconditionally restricted, the form of generated data is uncontrollable. Additional information is added by CGAN [39] to restrict the model and thus guide the data generation process. Infogan [40] adds the metric of mutual information to increase the degree of dependence between noises and generated data. It refines the noise for generating samples and mines some potential information to improve the generation performance. Denton *et al.* [41] propose to use the convolutional network with cascaded Laplacian pyramid structure to generate images from rough to fine. Similarly, SGAN [42], ProGAN [43], and some other networks provide a multi-level stacked network model for progressive training to generate higher-quality images. These architectural improvements make the model more complex and require a lot of computing power, which is not realistic in practical application scenarios. In this work, we take advantage of the idea that GAN can be used for data augmentation, mainly referring to the WGAN model's core idea and improving it for better performance.

III. METHOD

In order to solve the problem of insufficient data for subsequent deep learning tasks, GAN is selected for data augmentation to generate more diverse and realistic samples. Considering the stability and good performance of WGAN, we propose our improved WGAN-SA model, where an overview of the proposed WGAN-SA is shown in Fig. 2. First, the generator and discriminator with the self-attention mechanism and residual block are adopted. Then, the generated bill images and real data are fed to the discriminator to judge whether it is real or fake using a Wasserstein loss. Finally, the well-trained GAN is adopted to generate high-resolution bill images, which are then fed to train the learning system with real data. To better describe our method, a notation table for math symbols is presented as Nomenclature.

A. Wasserstein Loss Function

The Wasserstein distance [4] is proposed in WGAN, where the earth mover distance instead of the Jensen–Shannon divergence is utilized to evaluate the distance between the distribution of the real samples and the generated samples. The advantage is that Wasserstein distance can measure the distance between two non-overlapping distributions, providing a meaningful gradient for its smoothness and thus making the training process more stable. WGAN-GP [26] utilizes gradient penalty strategy to replace weight pruning to ensure that all parameters of the discriminator are bounded, and it only takes effect on the concentrated region of true or false samples and the transition zone between them, making the numerical distribution of the model parameters more reasonable. We take the advantage of its loss function to build the proposed

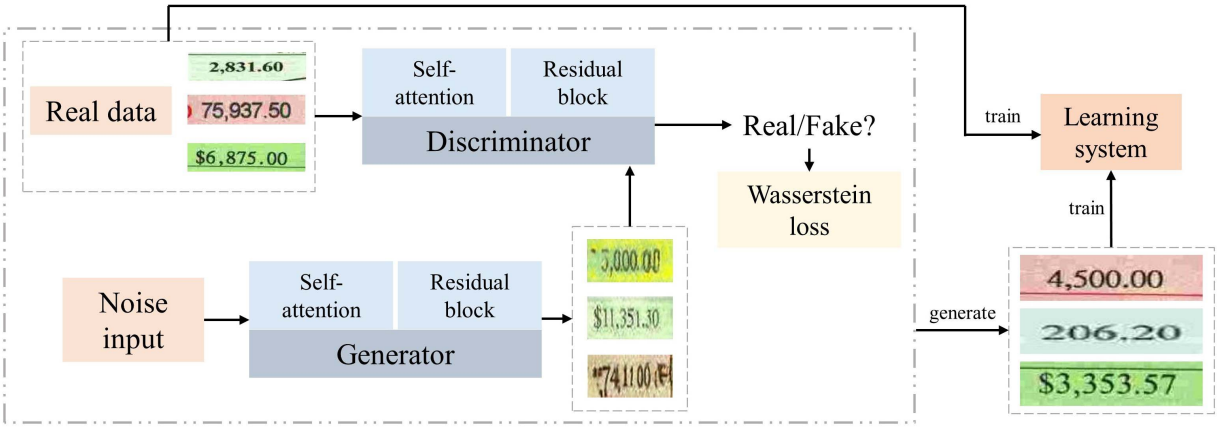

Fig. 2. Overview of the proposed WGAN-SA framework for few-sample learning.

WGAN-SA and solve the problem of mode collapse and gradient vanishing

$$L = -\mathbb{E}_{\mathbf{x}\sim p_{\text{real}}(\mathbf{x})}[D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z})}[D(G(\mathbf{z}))] + \lambda \cdot \mathbb{E}_{\mathbf{x}\sim p_{\hat{x}}}[(\| \nabla_{\hat{x}} D(\hat{x}) \|_2 - 1)^2] \quad (1)$$

where D is the discriminator, G is the generator, and $p_{\hat{x}}$ is the uniform sampling of the pair points between the real distribution p_{real} and the generated distribution p_z. The first two terms are the loss of the generator and the discriminator, respectively, to measure the distance between the real and synthetic images. The third term is the gradient penalty term and we set λ as 10 in experiments.

B. Self-Attention Mechanism

In order to improve the performance of the generated image, we introduce the self-attention module [44] in the proposed WGAN-SA. The advantage of the self-attention mechanism is that it enables the network to capture the correlation between not only short-range features, but also long-range features and global dependencies, which enables the model to learn more high-level semantic features to improve the generation effect. A single self-attention module is able to achieve good results in machine translation tasks. Similarly, the self-attention mechanism is helpful for GAN to capture the features of the original image more comprehensively and generate relatively reasonable images close to its distribution.

The self-attention mechanism is utilized in our method and the details are shown in Fig. 3. First of all, the feature map operates with three 1×1 convolution kernels to merge the features of each channel, and dimensional reduction is carried out commensurately to reduce unnecessary parameters in the intermediate operation and reduce the computational complexity. In addition, this operation can significantly increase the non-linearity without loss of resolution and while keeping the scale of the feature map unchanged. Thus, a meaningful gradient can always be maintained while increasing the depth of the network. The input feature map of the module has the same size as the output feature map, so it is flexible to be added to any part of the network to better capture the details and global features of the images.

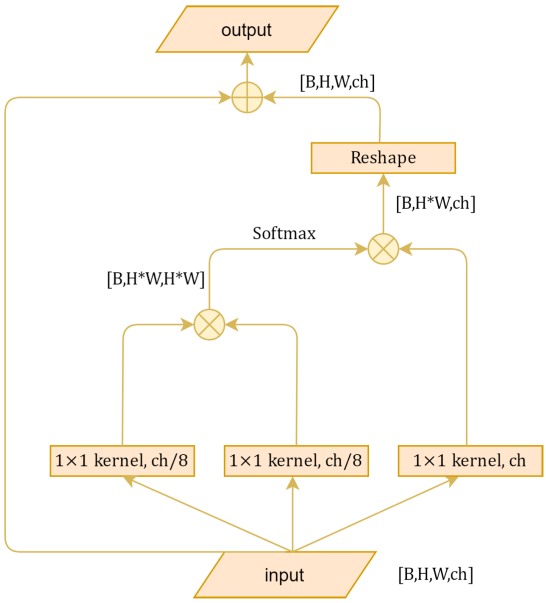

Fig. 3. Structure of the self-attention block in the proposed WGAN-SA.

C. Residual Block

Residual block is utilized as the basic structure of the proposed WGAN-SA, and the residual structure without bottleneck [45] is used in both the generator and the discriminator. It is difficult to train a deep neural network, where the phenomenon of gradient vanishing is easy to occur during backpropagation and results in performance degradation. The residual structure is connected through a shortcut to fit the residual mapping by stack layers rather than fitting the identity mapping, thus solving the problem of gradient vanishing. ResNet improves the performance of image tasks to a new level and SENet [46] also utilizes the residual structure for reference and presents excellent performance on the ImageNet dataset, which demonstrates the residual structure is suitable for feature extraction and deep model training. In addition, we added batch normalization layers to the residual block for regularization to accelerate model convergence and control overfitting. The residual block in our model is shown in Fig. 4.

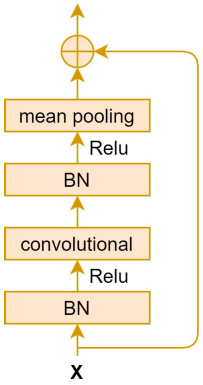

Fig. 4. Structure of the residual block in the proposed WGAN-SA.

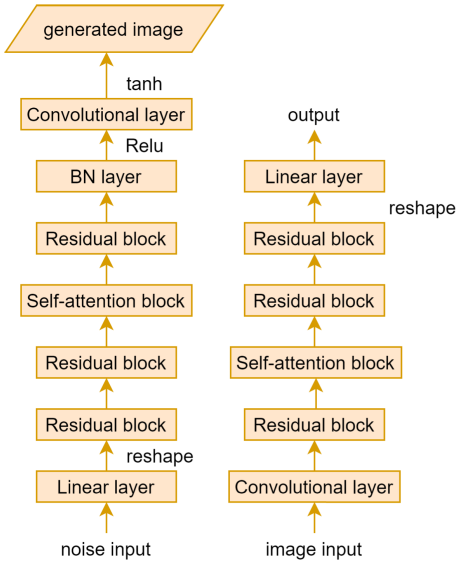

Fig. 5. Details of the structure of the generator and discriminator in the proposed WGAN-SA.

D. Architecture of the Generator and Discriminator

This is the maximization problem where the generator and the discriminator need to achieve the Nash equilibrium

$$\min_G \max_D f(G, D) = \mathbb{E}_{\mathbf{x}\sim p_{\text{real}}(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))] \quad (2)$$

where p_{real} is the distribution of real samples and p_z is the probability distribution of noise prior. $\mathbf{z}$ is the latent variables sampling from p_z, which is the input of the generator. In the image generation scene, both the generator and the discriminator are usually deep convolutional neural networks [47].

The details of the proposed WGAN-SA are represented in Fig. 5. The generator and the discriminator are made of the same basic blocks, which are composed of the three residual blocks described above and a self-attention block. Layers in the residual blocks of the generator and the discriminator are different. Each residual block of the generator contains one upsampling, while that of the discriminator contains three downsampling processes correspondingly. The generator generates synthetic images from the input of random noise vectors. The generated images and real images in the training set are sent into the discriminator as input synchronously. The generator and the discriminator are alternately trained as an unsupervised learning process, and the overall loss function consists of the generator loss, the discriminator loss, and the gradient penalty.

IV. Experiments

A. Datasets and Pre-Processing

First, two benchmark datasets for few-sample learning, including Omniglot [48] and miniImageNet [49], are adopted in this work. Omniglot dataset includes 1623 handwritten characters from over 50 different languages. The number of samples for each character is only 20 and the size of the samples is 28×28. The miniImageNet dataset is extracted from the ImageNet dataset, which includes 60 000 images and 100 categories. The size of the samples in the miniImageNet is 84×84. In addition, a classic few-sample learning framework named MAML [10] is selected as the learning system backbone in this work.

Besides, to test the performance of the proposed WGAN-SA, the dataset of the amount in figures from the Macau cheque scene is used to test the performance of the model. It is a private dataset owned by the Industrial and Commercial Bank of China. The number of the training set data is about 14 000, and the validation set is about 1000. The size of both the original images and the generated images is 288×72. Some of the images of the training set are shown in Fig. 6.

B. Implementation Details

All experiments are implemented in the deep learning framework Tensorflow on a single GPU. Images of the training set directly input to the discriminator as a tensor form with the original image size. The Adam optimizer is used to optimize the whole network iteratively, with an initial learning rate of 0.0001, a moment estimation of exponential decay rate of 0.6, and a second moment estimation of exponential decay rate of 0.9, to accelerate the convergence of the model. The maximum iteration time is set to 200 000. In addition, as described above, the BN layer is added to both the generator and discriminator parts of our model to avoid the problem of gradient disappearance. The settings of the MAML framework follow work [10].

C. Few-Sample Learning Performance on Benchmark Datasets

Following the experimental protocol presented in MAML [10], N-way classification results on two benchmark datasets, including Omniglot [48] and miniImageNet [49], with one- and five-shot are presented in this section. Specifically, the N-way classification denotes that it selects K different instances of N classes and test the model performance within the N classes. 1200 characters in the

Fig. 6. Samples of cheque fragments with amount in figures in the training set.

TABLE I
FEW-SAMPLE LEARNING PERFORMANCE ON THE OMNIGLOT DATASET

Method	5-way		20-way	
	1-shot	5-shot	1-shot	5-shot
MANN [50]	82.8%	94.9%	-	-
Siamese nets [2]	97.3%	98.4%	88.2%	97.0%
matching nets [49]	98.1%	98.9%	93.8%	98.5%
neural statistician [51]	98.1%	99.5%	93.2%	98.1%
memory mod [52]	98.4%	99.6%	95.0%	98.6%
MAML [10]	98.7%	**99.9%**	95.8%	98.9%
MAML+WGAN-SA (Ours)	**98.8%**	**99.9%**	**96.0%**	**99.0%**

TABLE II
FEW-SAMPLE LEARNING PERFORMANCE ON THE MINIIMAGENET DATASET

Method	5-way	
	1-shot	5-shot
nearest neighbor baseline [49]	41.08%	51.04%
matching nets [49]	43.56%	55.31%
meta-learner LSTM [9]	43.44%	60.60%
MAML [10]	48.70%	63.11%
MAML+WGAN-SA (Ours)	**49.91%**	**64.54%**

Omniglot dataset are chosen for training and others are for test, where the dataset is augmented with rotations.

Apart from the few-sample learning performance using MAML and the proposed WGAN-SA (marked as MAML+WGAN-SA), six other state-of-the-art works are included for comparisons, including MAMN [50], Siamese nets [2], matching nets [49], neural statistician [51], memory mod [52], and MAML [10].

The performance comparisons on the Omniglot dataset are shown in Table I. For the five-way setting, our proposed WGAN-SA framework can effectively improve the few-sample learning performance based on the MAML method, where it achieves 98.8% and 99.9% accuracy for one- and five-shot, respectively. In addition, the MAML+WGAN-SA can achieve 0.2%–7.8% higher accuracy for 20-way one-shot setting. Also, it can also achieve 0.1%–2.2% higher accuracy for five-shot, which shows the efficiency of the proposed method for improving the few-sample learning performance.

The performance comparisons on the miniImageNet dataset are shown in Table II. Besides matching nets [49] and MAML [10], baseline models based on nearest neighbor [49] and meta-learner LSTM [9] are also adopted for comparisons. From the table, our MAML+WGAN-SA can achieve the best performance for both one- and five-shot settings, which verifies the merits of our WGAN-SA for improving the few-sample learning performance.

D. Generation Performance on the Bill Dataset

We conduct experiments on both our WGAN-SA framework and the original WGAN-GP model to compare the generation performance using the bill dataset. Partly generated images of the two models are shown in Figs. 7 and 8, respectively. By comparing images generated from the two models, it is clear that the cheque images generated from the WGAN-GP are not good enough. Synthesized images are much more blurry compared with the original dataset. Some digits are difficult to distinguish because they are incomplete, and even some twisted and meaningless images are generated. In Fig. 8, images generated by our WGAN-SA framework are clearer and more realistic, where the generated digits are complete and meaningful. Specifically, the generated images from our WGAN-SA have a higher similarity with the real images and are more reasonable. In addition, the background of generated images by our method has a variety of colors and the digit also has a variety of styles to meet the requirements of diversity. This indicates that our model is suitable for the generation of the amount in figures in this scenario.

To better show the performance of the proposed WGAN-SA on solving the collapse problem, ablation experiments of generation are shown in Fig. 9. Specifically, Fig. 9(a) presents the generation results by the original WGAN-GP. Fig. 9(b) and (c) illustrates the generated images with self-attention mechanism and residual block, respectively. From the results, the self-attention mechanism can solve the collapse problem to some degree and help to obtain more reasonable images compared with the original WGAN-GP. Similar to it, the residual block in the WGAN can also help to generate images with few distortions. However, there are still some twisted images existed in Fig. 9(b) and (c). The generation results of the proposed WGAN-SA are shown in Fig. 9(d), which show clearer and more realistic images. Based on the ablation experiments of generation, our proposed WGAN-SA framework shows good ability to avoid the collapse problem in the bill dataset.

Fig. 7. Samples of cheque fragments with amount in figures generated by WGAN-GP.

Fig. 8. Samples of cheque fragments with amount in figures generated by our WGAN-SA.

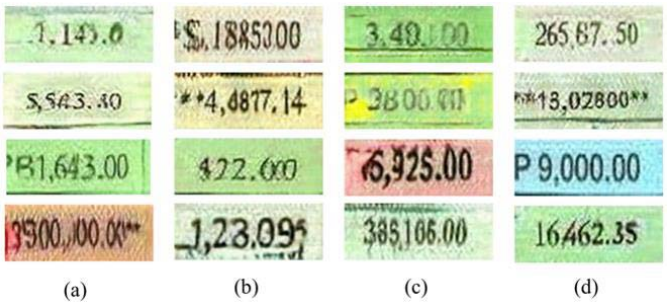

Fig. 9. Samples of cheque fragments with amount in figures generated by (a) WGAN-GP, (b) WGAN with self-attention mechanism, (c) WGAN with residual block, and (d) WGAN-SA (Ours).

TABLE III
FEW-SAMPLE LEARNING PERFORMANCE ON THE BILL DATASET

Method	10-way	
	1-shot	5-shot
MAML [10]	96.1%	98.5%
MAML+WGAN-GP	88.7%	95.5%
MAML+WGAN+SAM	94.3%	98.6%
MAML+WGAN+RB	92.8%	97.4%
MAML+WGAN-SA	**96.5%**	**99.0%**

E. Recognition Performance on the Bill Dataset

To perform the recognition experiments on the bill dataset, the MAML [10] is also applied to the backbone network. Because the bill dataset setting is different from the Omniglot and miniImageNet datasets, only the recognition performance of the ten-way case is presented in this work. Besides, 100 clear cheque fragment samples are adopted to construct the training and test sets, where each category roughly shares the same number to avoid class imbalance problems. The recognition performance on one- and five-shot is given in Table III, which includes the recognition performance from the MAML backbone, the MAML with the original WGAN-GP framework (MAML+WGAN-GP, for short), the MAML with the WGAN and self-attention mechanism (MAML+WGAN+SAM), the MAML with the WGAN and residual block (MAML+WGAN+RB), and the MAML with the proposed WGAN-SA framework (MAML+WGAN-SA).

From Table III, the MAML has achieved a good performance on both one- and five-shot cases. However, due to the model collapse problem of the original WGAN-GP, the MAML+WGAN-GP shows worse recognition performance. Specifically, for the one-shot case whose performance relies heavily on accurate and clear annotations, the generated images with more distortions will lead to worse recognition performance. Although the MAML+WGAN+SAM and MAML+WGAN+RB solve this issue to some degree and help to improve the performance on the five-shot case,

TABLE IV
FEW-SAMPLE LEARNING PERFORMANCE ON THE BILL DATASET WITH NOISE

Method	10-way	
	1-shot	5-shot
MAML [10]	94.6%	97.5%
MAML+WGAN-GP	85.5%	93.2%
MAML+WGAN+SAM	93.1%	96.8%
MAML+WGAN+RB	92.2%	97.8%
MAML+WGAN-SA	**94.8%**	**97.9%**

they still result in worse performance on the one-shot case. For comparisons, our proposed MAML+WGAN-SA shows performance improvement on both one- and five-shot cases, which verifies the efficiency of the proposed framework.

Apart from the above few-sample recognition performance, random Gaussian noise is added to the images in the bill dataset to verify the robustness of the proposed WGAN-SA framework. The Gaussian noise follow Noise $= N(0, 10^4)$, where the $\mu = 0$ and $\sigma = 100$. The recognition setting is the same as the above, and the performance is shown in Table IV. Although the noise leads to performance degradation to some degree from the table, our proposed WGAN-SA framework can also improve the few-sample learning performance based on the MAML backbone and shows the efficiency on the bill dataset in the real-world applications.

V. CONCLUSION

The purpose of our work is to utilize the generative model GAN to enhance the target data, aiming to improve the few-sample recognition performance in financial multi-bill scene. We draw on the experience of the WGAN-GP model, the Wasserstein Distance, and Gradient Penalization and optimize it. The residual structure is utilized as the model's basic structure, and a self-attention mechanism module is added to both the generator and the discriminator to improve the model performance. Numerous experiments show that our proposed WGAN-SA can first improve the few-sample learning performance on two benchmark datasets. Then, the generated images from our WGAN-SA show better generation performance than the original WGAN-GP model, where the synthetic images not only are more realistic, but also meet the requirement of diversity. In the process of training, the model is very stable without training collapse or pattern collapse, which indicates that our method can be used as a good solution for small sample data augmentation and can be applied to the generation of other types of images. Recognition performance on the bill dataset also proves the efficiency of the proposed WGAN-SA framework. In the future, we will also combine the GAN and VAE models for further experiments to see whether there will be a better performance and more helpful for real-world financial applications.

REFERENCES

[1] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "One-shot learning with memory-augmented neural networks," 2016, *arXiv:1605.06065*.

[2] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. ICML Deep Learn. Workshop*, Lille, France, vol. 2, 2015, pp. 1–30.

[3] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[4] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[5] J. Wang, Y. Yang, T. Wang, R. S. Sherratt, and J. Zhang, "Big data service architecture: A survey," *J. Internet Technol.*, vol. 21, no. 2, pp. 393–405, 2020.

[6] J. Zhang, S. Zhong, T. Wang, H. Chao, and J. Wang, "Blockchain-based systems and applications: A survey," *J. Internet Technol.*, vol. 21, no. 1, pp. 1–14, 2020.

[7] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," 2017, *arXiv:1703.05175*.

[8] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1199–1208.

[9] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *Proc. 5th Int. Conf. Learn. Represent. (ICLR)*, Toulon, France, Apr. 2017, pp. 1–11.

[10] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1126–1135.

[11] Q. Sun, Y. Liu, T.-S. Chua, and B. Schiele, "Meta-transfer learning for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 403–412.

[12] M. Long, F. Peng, and Y. Zhu, "Identifying natural images and computer generated graphics based on binary similarity measures of PRNU," *Multimedia Tools Appl.*, vol. 78, no. 1, pp. 489–506, Jan. 2017.

[13] D. Zhang, Q. Li, G. Yang, L. Li, and X. Sun, "Detection of image seam carving by using Weber local descriptor and local binary patterns," *J. Inf. Secur. Appl.*, vol. 36, pp. 135–144, Oct. 2017.

[14] X. Zhao, Y. Chen, J. Guo, and D. Zhao, "A spatial-temporal attention model for human trajectory prediction," *IEEE/CAA J. Autom. Sinica*, vol. 7, no. 4, pp. 965–974, Jul. 2020.

[15] H. Han, W. Ma, M. Zhou, Q. Guo, and A. Abusorrah, "A novel semi-supervised learning approach to pedestrian reidentification," *IEEE Internet Things J.*, vol. 8, no. 4, pp. 3042–3052, Feb. 2021.

[16] X. Wang and H. Duan, "Hierarchical visual attention model for saliency detection inspired by avian visual pathways," *IEEE/CAA J. Autom. Sinica*, vol. 6, no. 2, pp. 540–552, Mar. 2019.

[17] Z. Huang, X. Xu, H. Zhu, and M. Zhou, "An efficient group recommendation model with multiattention-based neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 11, pp. 4461–4474, Nov. 2020.

[18] A. Mikolajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," in *Proc. Int. Interdiscipl. PhD Workshop (IIPhDW)*, May 2018, pp. 117–122.

[19] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019.

[20] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[21] S. Nowozin, B. Cseke, and R. Tomioka, "f-GAN: Training generative neural samplers using variational divergence minimization," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 271–279.

[22] W. Fedus, M. Rosca, B. Lakshminarayanan, A. M. Dai, S. Mohamed, and I. Goodfellow, "Many paths to equilibrium: GANs do not need to decrease a divergence at every step," 2017, *arXiv:1710.08446*.

[23] T. Che, Y. Li, A. P. Jacob, Y. Bengio, and W. Li, "Mode regularized generative adversarial networks," 2016, *arXiv:1612.02136*.

[24] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2794–2802.

[25] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," 2017, *arXiv:1701.04862*.

[26] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of Wasserstein GANs," 2017, *arXiv:1704.00028*.

[27] M. G. Bellemare *et al.*, "The Cramér distance as a solution to biased Wasserstein gradients," 2017, *arXiv:1705.10743*.

[28] T. Salimans, H. Zhang, A. Radford, and D. Metaxas, "Improving GANs using optimal transport," 2018, *arXiv:1803.05573*.

[29] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial network," 2016, *arXiv:1609.03126*.

[30] X. Li, Y. Liang, M. Zhao, C. Wang, and Y. Jiang, "Few-shot learning with generative adversarial networks based on WOA13 data," *Comput., Mater. Continua*, vol. 60, no. 3, pp. 1073–1085, 2019.

[31] N. Yang, M. Zhou, B. Xia, X. Guo, and L. Qi, "Inversion based on a detached dual-channel domain method for StyleGAN2 embedding," *IEEE Signal Process. Lett.*, vol. 28, pp. 553–557, 2021.

[32] F. Zhang, H. Zhao, W. Ying, Q. Liu, A. N. J. Raj, and B. Fu, "Human face sketch to RGB image with edge optimization and generative adversarial networks," *Intell. Autom. Soft Comput.*, vol. 26, no. 4, pp. 1391–1401, 2020.

[33] M. Zhao, X. Liu, X. Yao, and K. He, "Better visual image super-resolution with Laplacian pyramid of generative adversarial networks," *Comput., Mater. Continua*, vol. 64, no. 3, pp. 1601–1614, 2020.

[34] K. Fu, J. Peng, H. Zhang, X. Wang, and F. Jiang, "Image super-resolution based on generative adversarial networks: A brief review," *Comput., Mater. Continua*, vol. 64, no. 3, pp. 1977–1997, 2020.

[35] K. Liu, Z. Ye, H. Guo, D. Cao, L. Chen, and F.-Y. Wang, "FISS GAN: A generative adversarial network for foggy image semantic segmentation," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 8, pp. 1428–1439, Aug. 2021.

[36] K. Zhang, Y. Su, X. Guo, L. Qi, and Z. Zhao, "MU-GAN: Facial attribute editing based on multi-attention mechanism," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 9, pp. 1614–1626, Sep. 2021.

[37] G. Shen, Z.-R. Tang, P. Shen, and Y. Yu, "HQ-Trans: A high-quality screening based image translation framework for unsupervised cross-domain pedestrian detection," in *Proc. Int. Conf. Image Graph.* Cham, Switzerland: Springer, 2021, pp. 16–27.

[38] P. Xiang, L. Wang, F. Wu, J. Cheng, and M. Zhou, "Single-image de-raining with feature-supervised generative adversarial network," *IEEE Signal Process. Lett.*, vol. 26, no. 5, pp. 650–654, May 2019.

[39] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[40] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 2180–2188.

[41] E. Denton, S. Chintala, A. Szlam, and R. Fergus, "Deep generative image models using a Laplacian pyramid of adversarial networks," 2015, *arXiv:1506.05751*.

[42] X. Huang, Y. Li, O. Poursaeed, J. Hopcroft, and S. Belongie, "Stacked generative adversarial networks," in *Proc. CVPR*, vol. 2, Jul. 2017, pp. 15077–15086.

[43] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2017, *arXiv:1710.10196*.

[44] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

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

[46] J. Hu, L. Shen, and G. Sun, "Squeeze- and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[47] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.

[48] B. Lake, R. Salakhutdinov, J. Gross, and J. Tenenbaum, "One shot learning of simple visual concepts," in *Proc. Annu. Meeting Cogn. Sci. Soc.*, 2011, vol. 33, no. 33, pp. 2568–2573.

[49] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 3630–3638.

[50] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1842–1850.

[51] H. Edwards and A. Storkey, "Towards a neural statistician," 2016, *arXiv:1606.02185*.

[52] Ł. Kaiser, O. Nachum, A. Roy, and S. Bengio, "Learning to remember rare events," 2017, *arXiv:1703.03129*.

Zhi-Ri Tang (Student Member, IEEE) received the B.Sc. and M.Eng. degrees from Wuhan University, Wuhan, China, in 2017 and 2019, respectively, where he is currently pursuing the Ph.D. degree.

His main current research interests include machine learning, cognitive computing, brain–computer interface, intelligent healthcare, and biomedical informatics.

Qi-Qi Chen received the bachelor's degree in automation from Shanghai Jiao Tong University, Shanghai, China, in 2019, where she is currently pursuing the master's degree with the Department of Automation.

Her research interests include computer vision and medical image analysis.

Zhao-Hui Sun (Member, IEEE) is currently pursuing the Ph.D. degree in industrial intelligent system with the Department of Industrial Engineering, School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai, China.

He has authored/coauthored more than 15 research papers in top-tier refereed international journals and conferences, such as IEEE TRANSACTIONS ON ENGINEERING MANAGEMENT and IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS. His current research interests include blockchain, evolutionary optimization, neuroergonomics, brain-inspired intelligence, cognitive computing, complex networks, non-parametric machine learning, operations research, knowledge automation, and their applications in industrial or medical problems.

Mr. Sun is a member of the IEEE Computational Intelligence Society and Association for Computing Machinery. He has been severing as a reviewer for several top-tier international journals and conferences in his research field.

Pengwen Xiong (Member, IEEE) received the B.S. degree from the North University of China, Taiyuan, China, in 2009, and the Ph.D. degree in instrument science and technology from Southeast University, Nanjing, China, in 2015.

He visited the Laboratory for Computational Sensing and Robotics, Johns Hopkins University, Baltimore, MD, USA, from 2013 to 2014. He is currently an Associate Professor with the School of Information Engineering, Nanchang University, Nanchang, China, and a Post-Doctoral Fellow with the School of Instrument Science and Engineering, Southeast University. His research interests include human–robot interaction and robotic sensing and controlling.

Bao-Hua Zhang received the Ph.D. degree in computer software and theory from Fudan University, Shanghai, China, in 2009.

She joined the Software Development Center, Industrial and Commercial Bank of China (ICBC), Shanghai, in 2009, where she is currently working with the Big Data and AI Laboratory as the Technical Manager. She is also dedicated to studying the application of big data and AI technology in the financial field.

Lu Jiang is currently with the Industrial and Commercial Bank of China, Shanghai, China. His main research interests include intelligent finance and computer vision.

Edmond Q. Wu (Senior Member, IEEE) received the Ph.D. degree in controlling theory and engineering from Southeast University, Nanjing, China, in 2010.

He is currently an Associate Professor with the Key Laboratory of System Control and Information Processing, Ministry of Education, Shanghai Jiao Tong University, Shanghai, China. His research interests include deep learning, fatigue recognition, and human–machine interaction.

Dr. Wu is an Associate Editor of IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS.