IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 9, 2020, accepted July 10, 2020, date of publication July 20, 2020, date of current version July 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3010612

Counterfeit Anomaly Using Generative Adversarial Network for Anomaly Detection

HAOCHENG SHEN[1], JINGKUN CHEN[2], RUIXUAN WANG[3], AND JIANGUO ZHANG[2,4], (Senior Member, IEEE)

[1]AI Lab, Tencent, Shenzhen 518054, China
[2]Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China
[3]School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510275, China
[4]Research Institute of Trustworthy Autonomous Systems, Southern University of Science and Technology, Shenzhen 518055, China

Corresponding authors: Ruixuan Wang (wangruix5@mail.sysu.edu.cn) and Jianguo Zhang (zhangjg@sustech.edu.cn)

This work was supported in part by the Research Institute of Trustworthy Autonomous Systems, in part by the National Key Research and Development Program under Grant 2018YFC1315402, in part by the Guangdong Key Research and Development Program under Grant 2019B020228001, and in part by the Guangzhou Science and Technology Program under Grant 201904010260.

ABSTRACT Anomaly detection aims to detect anomaly with only normal data available for training. It attracts considerable attentions in the medical domain, as normal data is relatively easy to obtain but it is rather difficult to have abnormal data especially for some rare diseases, making training a standard classifier challenging or even impossible. Recently, generative adversarial networks (GANs) become prevalent for anomaly detection and most existing GAN-based methods detect outliers by the reconstruction error. In this paper, we propose a novel framework called *adGAN* for anomaly detection using GAN. Unlike existing GAN-based methods, *adGAN* is a discriminative model, which uses the fake data generated from GAN as an abnormal class, and then learns a boundary between normal data and simulated abnormal data. Thus it is able to output the anomaly scores directly similar as one-class SVM (OCSVM), without any reconstruction process. We explicitly design *adGAN* with two key elements, i.e., *fake pool generation* and *concentration loss*. The fake pool is created by incrementally collecting the fake data produced by intermediate-state GAN, which are likely surrounding the normal data distribution. The concentration loss is innovatively introduced to penalize large standard deviations of discriminator outputs for normal data, aiming to make the distribution of normal data more compact and more likely to be separated from the distribution of the potential abnormal data. The trained discriminator is finally used as an anomaly detector. We evaluated *adGAN* on three datasets, including ab-MNIST for synthetic anomaly detection, the ISIC'2016 for skin lesion detection, and the BraTS'2017 for brain lesion detection. The extensive experiments demonstrate that *adGAN* is consistently superior to its competitors on all three datasets.

INDEX TERMS Anomaly detection, concentration loss, fake pool, GAN.

I. INTRODUCTION

In this paper, we consider a specific task of anomaly detection, for which there is only *normal* data available for training. It is of great interests in medical image analysis as well as in clinical routine examinations, because healthy (or normal) data is often easy to obtain but it is rather difficult to obtain abnormal data, especially for some rare diseases. The key challenge of the task is from the lack of abnormal data for training a detector.

In the medical image analysis domain, the conventional parametric and non-parametric statistical models and one-class SVM have been widely applied to anomaly detection. *Parametric* models usually refer to Gaussian or Gaussian mixture models, which estimate the density distribution of normal data from training set to predict the abnormality of a test sample. For example, Sidibe *et al.* [1] built a Gaussian mixture model from multiple healthy optical coherence tomography (OCT) images for abnormality prediction of any new images. Parametric models often assume that the normal data distribution is a Gaussian or a mixture of Gaussian distributions, and therefore work well only under the conditions of simple data distributions. In comparison, *non*-parametric

The associate editor coordinating the review of this manuscript and approving it for publication was Ahmed Farouk.

statistical models, such as Gaussian process, are more capable of modeling complex distributions but have more computational loads. Ziegler *et al.* [2] developed a Gaussian process model for pixel-level anomaly detection, which can predict the Gaussian distribution of each pixel's intensity within the grey matter region in a healthy brain according to the age, gender, and volume of grey and white matter. Both parametric and non-parametric models are bottom-up generative approaches, and therefore are limited to modeling distributions of normal data with low dimension. In contrast, *one-class* SVM is a top-down classification-based method for anomaly detection, which constructs a hyperplane as a decision boundary that best separates normal data and the origin point in a transformed (often high-dimensional) feature space, and meanwhile maximises the distance between the origin and the hyperplane. For example, Mourão-Miranda *et al.* [3] performed anomaly detection by training a one-class SVM using fMRI images from healthy population. Seeböck *et al.* [4] also used one-class SVM to detect the abnormal regions on retinal OCT images by training on super-pixels from healthy retinal OCT images.

While the conventional approaches have been widely used in the medical domain, there is one serious drawback to restrict their performance, i.e., the feature representation of images needs to be manually designed in advance. Without the need to extract hand-crafted features, generative adversarial networks (GANs) are recently becoming popular for medical anomaly detection due to their capability of implicitly modeling more complex data distribution than the conventional approaches. The first GAN for anomaly detection, called AnoGAN, was proposed by Schlegl *et al.* [5] for retinal OCT images. The basic idea is to train a generator in the AnoGAN which can generate only normal image patches, such that any abnormal patch would not be well reconstructed by the generator. A fast version of the AnoGAN called f-AnoGAN [6] was proposed by the same authors later on, with an additional encoder included to make the generator become an auto-encoder.

In this paper, we propose an alternative novel anomaly detection method based on GAN. The existing GAN-based anomaly detection methods [5], [6] are patch *reconstruction* based, of which the main purpose is to reconstruct the corresponding healthy counterpart given a new image patch. In contrast, our method is a patch-level *discriminative* model, which directly learns the boundary of the normal data distribution and is able to output the anomaly score of a new image patch without the reconstruction process. To the best of our knowledge, such a discriminative GAN-based anomaly detection model has not been explored before. We term our approach as *adGAN*. The proposed adGAN is evaluated on three different public datasets: a modified MNIST dataset for synthetic anomaly detection (termed as ab-MNIST), the ISIC'2016 dataset for skin lesion detection, and the BraTS'2017 for brain lesion detection. The extensive experiments demonstrate that adGAN is consistently superior to its competitors, including the well-known one-class SVM and the recent GAN-based methods on all three datasets.

II. RELATED WORK

Since the proposed adGAN can be considered as a special application of the GANs, in this section, we briefly review the fundamentals of relevant GAN models and their applications for anomaly detection. The conventional approaches to anomaly detection for medical image analysis have been introduced in the previous section.

GANs have achieved great success in generating data for learning models [7]–[11]. The original GAN [7] architecture consists of two networks, a generator G and a discriminator D. The generator G maps a random vector z from a prior distribution to an image space while the discriminator D maps an input image to a probability of being real or fake. The entire GAN framework is trained by optimizing a minimax loss function in an adversarial manner: G is encouraged to generate realistic images and meanwhile D is trying to distinguish between real images and generated fake images. After convergence, D is able to reject generated images that are too fake, and G can produce realistic images whose distribution is close to the real data distribution. DCGAN [8] replaces fully connected networks in the original GAN [7] with deep convolutional networks for both G and D, and is trained with the same minimax loss function by gradient descent. GANs are also extended to the conditional setting in [11] and have been used in many image-to-image translation applications [12], [13]. Unfortunately, GAN is not easy to train: the minimax loss function of GAN or DCGAN can lead to gradient vanishing problem, especially when the discriminator is trained to be very strong.

WGAN [9] alleviates the gradient vanishing and mode collapse problems by designing a new Wasserstein metric to explicitly measure the distance between two distributions. Specifically, the discriminator (also called *critic*) is trained to output an approximated Wasserstein distance between the real data distribution and the generated data distribution. And the generator is then optimized to minimize that distance to push two distributions closer. The training of WGAN can be further improved by using gradient penalty [10]: instead of using weight clipping in the original WGAN, gradient penalty [10] penalizes the norm of gradient of the critic with respect to its inputs and achieves better generated results.

It is aware that variants of GANs for anomaly detection were developed in recent years. In medical imaging domain, one representative for anomaly detection is AnoGAN [5]. AnoGAN learns a mapping between a random distribution and the image manifold of the normal class. Specifically, it first trains a DCGAN to generate fake images that look *normal*. For a test image, AnoGAN then seeks the *optimal* vector in the latent space through back-propagation such that the difference between the generated image and the test image is minimum, where anomaly detection is performed based on that difference. It is worth noting that during inference, AnoGAN relies on a heavy optimization process [5], thus

resulting in a high computational load. Schlegl *et al.* [6] later on modified the AnoGAN model by adding an encoder to the generator, such that any real image can be directly mapped to the latent space and the encoded latent feature can then generate the reconstructed healthy counterpart. In this way, optimization process is not required during anomaly detection, thus resulting in a fast AnoGAN (f-AnoGAN).

Besides applications in the medical domain, GANs have also been applied to anomaly detection in the domain of natural images or videos recently [14]–[20]. For example, similar to the f-AnoGAN, an adversarially learned one-class classifier (ALOCC) [14] was recently proposed where the auto-encoder (as the generator) aims to reconstruct the original input images while the discriminator aims to differentiate the reconstructed image from the corresponding original image. Different from the f-AnoGAN, during testing, ALOCC uses the probabilistic output of the discriminator as the abnormality score for the reconstructed input, by assuming that the discriminator would give high scores for the reconstruction of original normal images but low scores for that of originally abnormal images. Ravanbakhsh *et al.* [15] trained GANs using normal frames and corresponding optical-flow images to learn an internal representation of the normality in crowded scenes, which is then applied for abnormal events detection in videos. Deecke *et al.* [17] detected anomaly by searching for a good representation of a sample in the latent space of the generator, and the sample is treated anomalous if such a representation is not found. Pidhorskyi *et al.* [19] developed a probabilistic approach to compute how likely a sample is generated by the inlier distribution using autoencoder. Perera *et al.* [20] extended ALOCC by further explicitly constraining the latent space to exclusively represent the given class.

III. ADGAN FOR ANOMALY DETECTION

The proposed adGAN is built on the improved WGAN with gradient penalty [10]. The core idea in adGAN is to counterfeit anomaly using the fake data generated from intermediate-state GAN, which are then combined with normal data to learn a boundary between normal data and simulated abnormal data. We designed the adGAN framework with two key components: *fake pool generation* and *concentration loss*, and experimentally showed that they are crucial for the superior performance of adGAN.

A. FAKE POOL GENERATION

In medical imaging domain, although the appearances of lesion regions differ from healthy regions in an image, to some extent they share more visual similarity compared to the similarity between medical and non-medical image domains. In other words, the distribution of visual features from abnormal regions is neither heavily overlapped nor far away from the feature distribution of normal regions. Instead, it is approximately surrounding or in the boundary area of the distribution of normal regions. Since we would like to use the generated data to simulate the abnormal class, the generated data should ideally come from such surrounding area of the distribution of normal regions.

During the training phase of WGAN, the generated data distribution is gradually getting closer to the real data distribution as the generator aims to minimize the Wasserstein distance between two distributions, thus we hypothesize that the generated data from the intermediate-state WGAN are likely to locate in the boundary area of the real data distribution. Motivated by this, we propose the "*fake pool*", which is collected by incrementally saving the generated images during WGAN training phase.

The details of using fake pool are described in Algorithm 1 (line 1-13). We first train a WGAN with gradient penalty [10] using healthy images only, which aims to produce realistic health-looking images gradually (line 2-9), starting from random noises. The model parameters of the improved WGAN, i.e., the critic's parameters w and the generator's parameters θ, are updated as in line 2-6 and line 7-9, respectively. Note that we use the hyper-parameters directly from [10]. In order to optimally approximate the Wasserstein distance, the critic is updated n_{critic} times per generator's update. The WGAN is trained for N iterations, and a number of n fake images are collected from the current generator into the fake pool after every k iterations (line 10-12).

B. CONCENTRATION LOSS

After the fake pool is created, the critic D_w in the improved WGAN will be retrained using the same set of healthy images

Algorithm 1 The Proposed adGAN With a Generator G_θ and a Critic D_w Where θ and w Represents the Model Parameters. We Use Default Parameter Values of $\lambda = 10$, $n_{critic} = 5$ and Gradient Penalty (GP) as in [10]

Require: training iteration N for generating *fake pool*, training iteration M for the critic afterwards.

1: **for** $i < N$ **do**
2: **for** $t = 1, \ldots, n_{critic}$ **do**
3: Sample real data $\mathbf{x} \sim \mathbb{P}_r$ and fake data $\widehat{\mathbf{x}} \sim G_\theta(\mathbf{z})$
4: $L_c = \mathbb{E}(D_w(\widehat{\mathbf{x}})) - \mathbb{E}(D_w(\mathbf{x})) + \lambda \cdot \text{GP}$
5: $w \leftarrow \text{Adam}(\nabla_w L_c, w)$
6: **end for**
7: Sample m fake data $\widehat{\mathbf{x}} \sim G_\theta(\mathbf{z})$, $\mathbf{z} \sim p(\mathbf{z})$
8: $L_g = -\mathbb{E}(D_w(\widehat{\mathbf{x}}))$
9: $\theta \leftarrow \text{Adam}(\nabla_\theta L_g, \theta)$
10: **if** $i \,\%\, k == 0$ **then**
11: Generate and collect n fake samples into fake pool.
12: **end if**
13: **end for**
14: **for** $j < M$ **do**
15: Reset D_w with random initialization
16: Sample real data $\mathbf{x} \sim \mathbb{P}_r$, fake data $\widetilde{\mathbf{x}} \sim$ fake pool
17: $L_c' = \mathbb{E}(D_w(\widetilde{\mathbf{x}})) - \mathbb{E}(D_w(\mathbf{x})) + \alpha \cdot \text{S}(D_w(\mathbf{x})) + \lambda \cdot \text{GP}$
18: $w \leftarrow \text{Adam}(\nabla_w L_c', w)$
19: **end for**

as well as the generated images in fake pool with a new *concentration loss* function. Recall that the original loss function for the critic in WGAN is defined as

$$L_c = \mathbb{E}(D_w(\widehat{\mathbf{x}})) - \mathbb{E}(D_w(\mathbf{x})) + \lambda \cdot GP \quad (1)$$

where GP stands for the gradient penalty term of which detailed form can be found in [10], λ is its corresponding weight, and $\mathbb{E}$ denotes the expectation. This loss aims at maximizing the critic output for real data $\mathbf{x}$ meanwhile minimizing the critic output for fake data $\widehat{\mathbf{x}}$ from generator. Thus the value of the critic output could potentially be used to measure the (ab)normality of any input image to the critic, i.e., relatively large output value indicates that the input image is more likely normal, while relatively small output indicates that the input image is more likely abnormal. However, the original loss (Eq. 1) only maximizes the between-class distance (e.g., the distance between real data and fake data) without considering any within-class distance. For anomaly detection, ideally the between-class distance should be larger (e.g., the normal data is relatively far away from the abnormal data) than the within-class distance (e.g, the normal data is relatively close to each other). Thus, we innovatively add a concentration term into the original loss to decrease the standard deviation of critic output for real data, resulting in the new loss function (termed as *concentration* loss):

$$L_c' = \mathbb{E}(D_w(\widetilde{\mathbf{x}})) - \mathbb{E}(D_w(\mathbf{x})) + \alpha \cdot S(D_w(\mathbf{x})) + \lambda \cdot GP \quad (2)$$

where the concentration term $S(\cdot)$ can be specifically represented as:

$$S(D_w(\mathbf{x})) = \sqrt{\mathbb{E}((D_w(\mathbf{x}) - \mathbb{E}(D_w(\mathbf{x})))^2)} \quad (3)$$

and α is its corresponding weight.

With this new loss (also see lines 15-18, Algorithm 1), it not only helps to maximize the critic output for real data $\mathbf{x}$ and minimizes the critic output for fake data $\widetilde{\mathbf{x}}$ from the fake pool (i.e., increase the between-class distance), but also simultaneously helps to minimize the standard deviation of the critic output for real data $\mathbf{x}$ (i.e., decrease the within-class distance). Intuitively, at the later stage of WGAN training phase, the generated data would look realistic to the real data, thus the distribution of images in the fake pool and the distribution of real images (inevitably) have some overlap in the original critic output space. Adding this concentration term would make the critic outputs for the real images more compactly distributed, which would reduce the within-class distance and in turn make these two output distributions more separable.

Note that the new loss only concentrates on reducing the within-class distance of normal data, and there is no constraint on the within-class of abnormal data (i.e., it allows large standard deviation of the critic output of fake data from the fake pool). This is because abnormal data could come from any complicated multi-mode density distribution, such that their standard deviation of critic output could be naturally large.

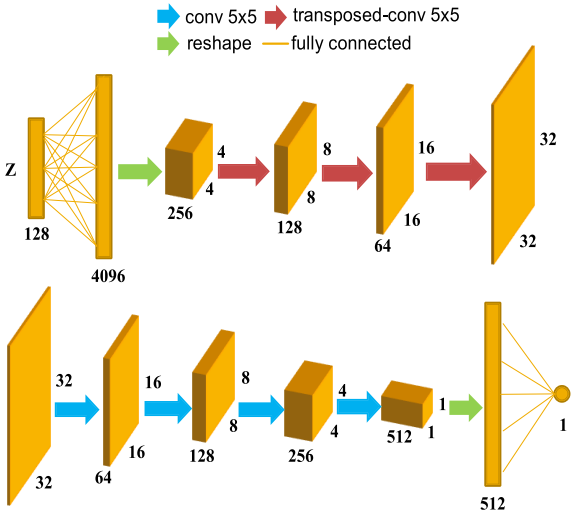


FIGURE 1. The architecture of the proposed adGAN. Top: the generator architecture; Bottom: the discriminator (critic) architecture. The numbers around each cuboid indicate the spatial size and the number of input or output channels for each convolutional layer.

C. THE adGAN ARCHITECTURE

The architecture of adGAN is shown in Figure 1. Specifically, the generator (top) takes a 128-dimensional random noise as input. Such input connects a fully connected layer with 4096 neurons, and then is reshaped to 256 feature maps with spatial size 4 × 4. These feature maps go through 3 transposed convolutional layers with 5 × 5 kernel, each of which doubles the spatial size and halves the number of feature maps except the last one, which outputs a 32 × 32 task-dependent image (e.g, RGB three-channel or grayscale one-channel image).

The critic (bottom) is made up of four convolutional layers and one fully-connected layer. It takes 32 × 32 images as input, and analogously halves the size of feature maps at every convolutional step using 5 × 5 kernel with stride 2. The number of feature maps starts at 64 at the first convolutional layer, and is doubled at every layer before the fully connected layer. Intuitively, the critic may have a symmetrical architecture with the generator (i.e., three rather than four convolutional layers followed by a fully-connected layer). However, such an architecture would make the output neuron of the last convolutional layer encode local information, thus would be unlikely to achieve the objective of anomaly detection which often needs to consider image-scale information. Thus, we add another convolutional layer in the critic using a 4 × 4 kernel, which takes the entire feature maps from the previous layer into consideration and results in feature maps of dimension 1 × 1 × 512, followed by a reshape operation and a fully-connected layer to produce the final output.

Both the generator and the critic use LeakyReLU [21] nonlinearities. The generator includes batch normalization [22] modules while the critic omits them as the batch normalization violate the penalization form of the gradient with respect to each input independently [10].

IV. EVALUATION

The proposed adGAN is evaluated on three datasets: ab-MNIST (a modified MNIST dataset), and two real medical datasets ISIC (skin lesion detection on International Skin Imaging Collaboration 2016 dataset [23]) and BraTS (brain lesion detection on Brain Tumor Segmentation benchmark 2017 dataset [24]–[26]). For all the experiments, the area under Receiver Operating Characteristic (ROC) curve (AUC) is reported as the evaluation metric. This section is organized as follows: the experimental results and analysis on the three datasets are described in Section IV-A, IV-B, IV-C, respectively. Then the comparison with the state-of-the-art anomaly detection methods and an ablation study are presented in Section IV-D and Section IV-E.

The proposed adGAN is implemented in Tensorflow. When training WGAN, the hyper-parameters λ and n_{critic} were set to 10 and 5 respectively, following the same setting as in [10]. The Adam optimizer [27] was used with 10^{-4} as the learning rate; the batch size m was set to 64. For *fake pool* (FP) generation, we set n to 64 and k to 100, which means the generator inserts 64 images into *FP* for every 100 iterations during training. For the new loss function, the weight of concentration loss term α was set to 1, and the effects of different α values are further discussed in Section IV-E.

For all three datasets, the model was trained for 200,000 iterations (N in Algorithm 1) to generate the fake pool; the critic with the proposed new loss function was then further trained for another 150,000 iterations (M in Algorithm 1). This setting is consistent across all experiments. There may be room to find an optimal early-stop criteria; it could be a potential future research direction.

A. ANOMALY DETECTION ON ab-MNIST

The proposed method is first evaluated on ab-MNIST (a modified MNIST). MNIST is a dataset of grey-scale handwritten digits from 0 to 9. To create the ab-MNIST, we use the official training set with 60,000 images for training adGAN. For testing, the official test set containing non-overlapped 10,000 images are treated as the normal class, and 10,000 abnormal images are artificially synthesized by adding some random square noises into the test set (randomness is applied to location, size and pixel intensity). Examples of the normal digit and the synthetic abnormal digits are illustrated in Fig. 2.

FIGURE 2. Examples of the synthetic abnormal digits (left) and normal digits (right) on ab-MNIST.

During training, the proposed fake pool incrementally collects the generated images at every 100 iterations. There is no mode collapse happened during the training, demonstrating the superiority of WGAN [10] compared to the original GAN [7], [8]. All digits in the fake pool are treated as abnormal class, which are combined with the normal digits in the training set to retrain the critic using the proposed loss function with the concentration term (Eq. 2).

TABLE 1. AUC values of different methods for anomaly detection on ab-MNIST.

	OCSVM	adGAN	adGAN(c)
AUC	0.89	**0.99**	0.97

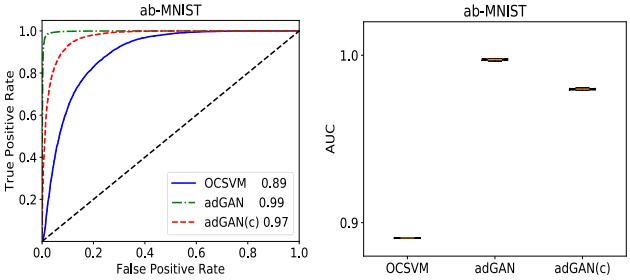


FIGURE 3. ROC curves of OCSVM, adGAN and adGAN(c) on ab-MNIST dataset (Left), and boxplots of AUC values of 10 repeated experiments (Right).

We choose the well-known one-class SVM (OCSVM) [28] as a baseline and compare our method to it first (see Table 1) since they are both discriminative models. The full comparison with the state-of-the-art methods will be described in Section IV-D. Note in Algorithm 1, after training WGAN, there are two options to train the critic with the new loss function using the real data and the fake data in *FP*: 1) reset the weights of the critic with random initialization and train it from scratch; and 2) carry the weights of the critic from WGAN and fine-tune it. We term the former as *adGAN* and the latter as *adGAN(c)* and evaluate both variants. As shown in Table 1, both adGAN and adGAN(c) give higher AUC values than OCSVM (i.e., adGAN 0.99 vs adGAN(c) 0.97 vs OCSVM 0.89), where adGAN and adGAN(c) have a 10% and a 7% improvement, respectively. The ROC curves of OCSVM, adGAN and adGAN(c) on ab-MNIST dataset are shown in Fig 3 (Left).

To further evaluate the stability of the proposed model, we repeat the experiment 10 times and visualise the variations of the AUC values using boxplots for both adGAN and adGAN(c). The corresponding boxplots are shown in Fig 3 (Right). In contrast, we also plot the boxlot of the AUC values of OCSVM for the 10 experiments, although OCSVM does not have such variations in performance due to its nature of convexity. It could be observed that the variations of the AUC values of adGAN and adGAN(c) are very small, which indicates that the proposed model is stable and its performance is

reproducible. We also perform a statistical test and the tiny p-value ($p < 10^{-5}$) indicate that the performance of the proposed model is significantly better than OCSVM.

B. SKIN LESION DETECTION ON ISIC

Skin lesion detection dataset is from the International Skin Imaging Collaboration (ISIC) challenge [23] in International Symposium on Biomedical Imaging (ISBI) 2016. It contains 900 RGB skin images with pixel-level annotations delineating lesion regions. Exemplar images along with their annotations are shown in Fig. 4. It can be seen that even for the healthy regions, there are large variations in appearance, colour, illumination and texture.

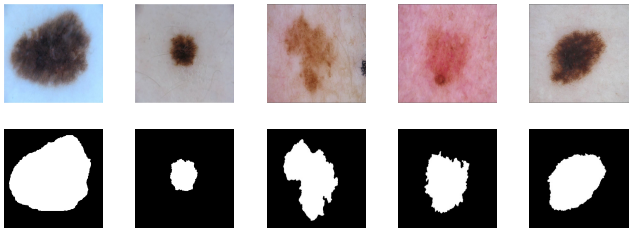


FIGURE 4. Exemplar RGB images (first row) and their annotations (second row) of skin lesions in the ISIC.

We exclude some images due to heavy hairs and artificially imposed noises, resulting in a subset of 660 images in our experiments. All images are resized to 128 × 160 and then patches of size 32 × 32 pixels with 75% overlap rate are extracted. If a patch contains no lesion area, we refer it a healthy patch; on the other hand, if more than 75% of a patch is lesion area, this patch is considered as a lesion patch. This results in 25744 healthy patches for training, and 7,332 healthy patches and 8,608 lesion patches for testing.

Examples of generated skin patches at the 1,000-th, 10,000-th and 200,000-th iterations during fake pool generation stage are shown in Fig. 5. As expected, it can be seen that the generated skin patches start from having relative single brown colour (Fig. 5, Left), to having varying colour patterns which are similar to the training set (Fig. 5, Middle). After training, there are more details formed within the patches, such as skin texture and speckle, so that the generated patches look more realistic (Fig. 5, Right).

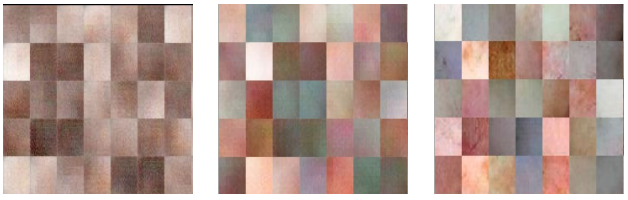


FIGURE 5. From left to right: examples of generated skin patches at the 1,000-th, 10,000-th and 200,000-th iteration in the adGAN training phase. Best viewed in colour.

Similarly as in Section IV-A, we evaluate both adGAN and adGAN(c) against the baseline OCSVM (see Table 2). Partially due to the relative high contrast between healthy and lesion patches, using OCSVM can already achieve an AUC value of 0.96. However, the proposed adGAN outperforms OCSVM. For example, adGAN and adGAN(c) obtain AUC values of 0.97 and 0.98, which have 1% and 2% improvement, respectively. In contrast to the ab-MNIST dataset, adGAN(c) is slightly better than adGAN. We attribute this phenomenon to the fact that ISIC is a more complicated real clinical dataset compared to the relative simple MNIST, thus fine-tuning the weights of critic is a better strategy than the random initialization. The ROC curves of OCSVM, adGAN and adGAN(c) on ISIC dataset and the boxplots of 10 repeated experiments are shown in Fig 6. The clear gaps in the distributions of AUC values between methods indicate that adGAN(c) indeed performs best, and the improvement is *not* from the randomness of the model. Statistically, the performances of both adGAN(c) and adGAN are significantly better than OCSVM ($p < 10^{-5}$).

To further evaluate the visual performance, we select some exemplar patches of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) (Fig. 7). Given the fact that a lower critic output value indicates a higher anomaly score and vice versa, we show those healthy patches with the lowest critic output values (aka FP shown in Fig. 7(a)); the lesion patches with the highest critic output values (FN, Fig. 7(b)); the healthy patches with the highest critic output values (TN, Fig. 7(c)) and the lesion patches with the lowest critic output values (TP, Fig. 7(d)).

It is shown that some healthy patches with high anomaly scores (Fig. 7(a)) contain some "abnormal" regions, such as red or dark spots, hair and scale marks, which is reasonable for adGAN to consider them as abnormal with low scores. Note that some healthy patches in blue colour are misclassified as abnormal. This is because the fact that the number of blue healthy patches in the training set is limited, thus the distribution of this kind of health patches cannot be well represented. For the lesion patches with low anomaly scores (Fig. 7(b)), it can be observed that their appearance is relatively homogeneous and smooth with low intensity contrast, which makes them look normal and it is difficult to identify the lesion regions within the patches compared with the easy cases (Fig. 7(d)). For TN and TP (Fig. 7(c) & (d)), adGAN works as expected: clear and homogeneous patches are classified as normal with the highest critic output values, while the patches containing large dark lesion regions are classified as abnormal with the lowest critic output values.

C. BRAIN LESION DETECTION

Brain lesion detection dataset is from BraTS 2017 dataset, which is a benchmark for magnetic resonance (MR) brain

TABLE 2. AUC values of different methods for anomaly detection on ISIC.

	OCSVM	adGAN	adGAN(c)
AUC	0.96	0.97	**0.98**

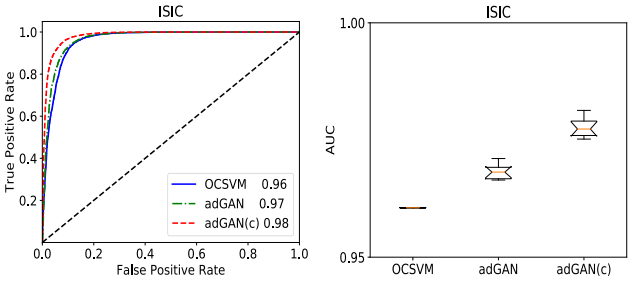


FIGURE 6. ROC curves of OCSVM, adGAN and adGAN(c) on ISIC dataset (Left), and boxplots of AUC values of 10 repeated experiments (Right).

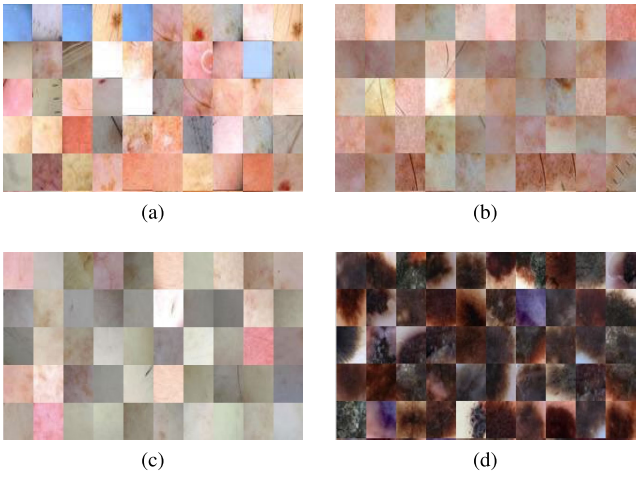


FIGURE 7. Exemplar patches of (a) false positives, (b) false negatives, (c) true negatives, and (d) true positives for skin lesion detection.

tumor segmentation [24]–[26] released in the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2017 BraTS challenge. In this dataset, each subject contains 4 MR modalities (T1, T1 contrast, T2 and FLAIR) as well as pixel-level annotations indicating tumor regions. In our experiments, we use the FLAIR modality only from each subject to detect complete tumor regions as the lesion regions, which include all tumor sub-regions, such as edema, necrosis, enhancing and non-enhancing tumor. Exemplar FLAIR images are shown in the first column of Fig. 10.

We randomly split 210 high grade gliomas subjects into 146 training and 64 testing cases. All images are normalized to have zero mean and unit standard deviation as in [29]. We extracted patches of size 32×32 pixels within brain regions (ignoring the background regions) from each slice of 3D MR volumes, resulting in 68,098 healthy patches for training, and 28,878 healthy patches and 3,523 lesion patches for testing.

Examples of generated brain patches at the 1,000-th, 10,000-th and 200,000-th iterations are shown in Fig. 8. The generated brain patches started from very noisy patches (Fig. 8, Left)) to the patches with less noise, and parts of brain structures such as cerebrospinal fluid (CSF) and sulci (Fig. 8, Middle)) are formed. And realistic-looking patches showing detailed brain texture and structures are generated when training is finished (Fig. 8, Right)).

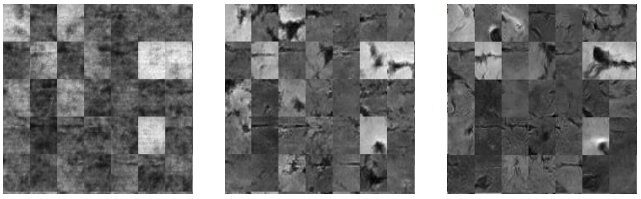


FIGURE 8. From left to right: examples of generated brain patches at the 1,000-th, 10,000-th and 200,000-th iteration of adGAN training phase.

The performance comparison between OCSVM and the proposed adGAN is summarized in Table 3. Compared to previous anomaly detection tasks, brain lesion detection is more challenging due to the larger variations in appearance and structure between subjects as well as noises and artifacts in MR images. In this case, the baseline OCSVM gives a AUC value of 0.88 while adGAN(c) achieves 0.92, which has a 4% improvement. It is noted that adGAN gives worse result (0.84) compared to OCSVM, which confirms that training the critic from scratch is not a good option for complex data. The ROC curves of OCSVM, adGAN and adGAN(c) on BraTS dataset and the boxplots of the AUC values of the 10 repeated experiments are shown in Fig 9. From the boxplots, it can also be observed that adGAN(c) not only produces higher mean AUC value of 10 repeated experiments when compared to adGAN, but also gives much smaller variance, indicating the stronger stability of the fine-tuned model. Furthermore, statistical testing shows that adGAN(c) is significantly better than both OCSVM and adGAN ($p < 10^{-5}$).

Fig. 10 shows examples of anomaly detection of brain lesion patches using adGAN(c) and OCSVM. AdGAN(c) localizes the lesion regions more accurately while OCSVM produces some false positives or false negatives. Note that adGAN only requires healthy patches and no patch labels are needed during the training phase.

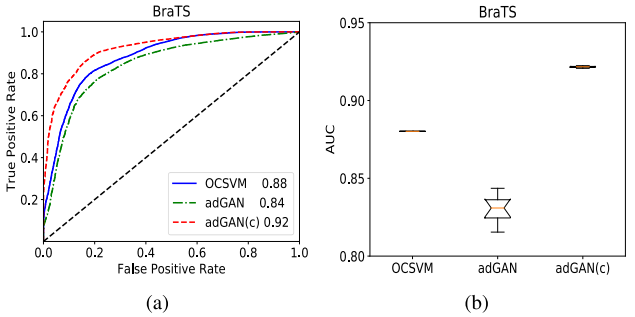
Based on the results presented in Tables 1-3, except for the ab-MNIST dataset, adGAN(c) outperforms adGAN for both ISIC and BraTS datasets which contain more complex real clinical data. On the other hand, although adGAN(c) is slightly worse than adGAN on ab-MNIST, it still outperforms OCSVM by a large margin. Thus, we apply the fine-tuned strategy for the critic in the remaining of this paper.

TABLE 3. AUC values of different methods for anomaly detection on BraTS.

	OCSVM	adGAN	adGAN(c)
AUC	0.88	0.84	**0.92**

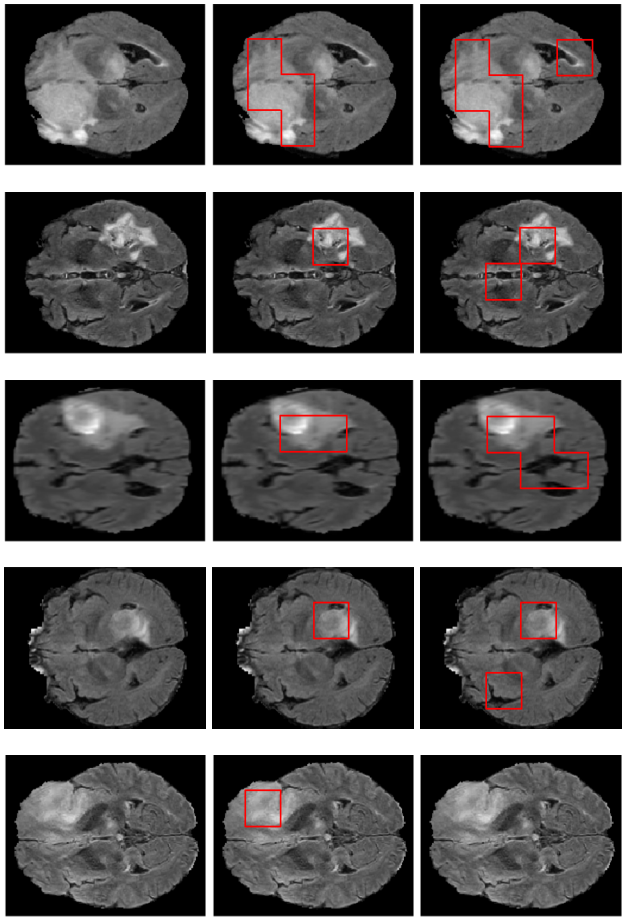
D. COMPARISON WITH THE STATE-OF-THE-ARTS

In this section, we further compare adGAN with other state-of-the-art methods on all three datasets, including a



(a) (b)

FIGURE 9. ROC curves of OCSVM, adGAN and adGAN(c) on BraTS dataset (Left), and boxplots of AUC values of 10 repeated experiments (Right).



(a) (b) (c)

FIGURE 10. Examples of anomaly detection on BraTS. (a) FLAIR images; (b) adGAN results; (c) OCSVM results. Regions within red rectangles are detected abnormal.

traditional generative approach kernel density estimation (KDE) [30], [31] as well as the most recent GAN-variant methods [5], [6], [14]. The comparison results are summarized in Table 4. It clearly shows that the proposed adGAN outperforms KDE by a large margin, indicating the superiority of discriminative learning and strong representation capability of neural networks.

In more details, here we compare adGAN to the original WGAN [10], and DCGAN [8], which were designed to model the input image distribution and aim to produce realistic-looking samples of input images. When WGAN or DCGAN is well trained, we simply take the critic as the anomaly classifier to perform anomaly detection, while the generator is not used. From Table 4, it can be observed that both WGAN and DCGAN do not perform well. For example, WGAN obtains about 0.85, 0.81, 0.86 AUCs for ab-MNIST, ISIC and BraTS respectively, which are about 12%, 17% and 6% lower compared to adGAN; DCGAN gives worse results. The main reason that WGAN or DCGAN do not perform well is that after training, the generator can generate realistic-looking fake data, thus the discriminator (or critic) is trained to differentiate the real data and the fake data mainly based on subtle features such as the checkerboard-like artifacts induced by the generator architecture [17], [32]. Thus, such a discriminator is not optimal for anomaly detection: the fake data is too realistic to mimic abnormal class. On the contrary, the fake pool of adGAN collects the fake data from the intermediate-state GAN from which the fake data is more suitable to represent abnormal class.

TABLE 4. Comparison (AUC value) with the state-of-the-art methods on three datasets.

	ab-MNIST	ISIC	BraTS
KDE [31]	0.58	0.71	0.73
DCGAN [8]	0.72	0.57	0.82
WGAN [10]	0.85	0.81	0.86
AnoGAN [5]	0.75	0.93	0.75
AnoGAN-mean	0.82	0.94	0.75
f-AnoGAN [6]	**0.98**	0.92	0.84
ALOCC [14]	0.91	0.97	0.87
adGAN	0.97	**0.98**	**0.92**

In the following, we compare adGAN to some state-of-the-art GAN-based anomaly detection methods including AnoGAN [5], [6] and ALOCC [14] in more details.

As shown in Table 4, AnoGAN obtains the AUCs of 0.75, 0.93, 0.75, respectively for ab-MNIST, ISIC and BraTS, which perform worse than the proposed adGAN. Additionally, in our experiments, we found that for a test image, the anomaly score computed by AnoGAN is sensitive to the initialization of the iterative process to find the optimal latent space point. In other words, there exists a large variety of the anomaly scores for the same image depending on different initializations. Therefore, we conduct another experiment: for each test image, we test it five times but using different initializations in testing, then the mean value of the five anomaly scores is computed as the final abnormal score for that image (denote as *AnoGAN-mean* in Table 4). This operation increases the performance of AnoGAN notably with AUCs increased to 0.82, 0.94 for ab-MNIST, ISIC while

remains the same 0.75 for BraTS. However, testing a single image for five times is obviously computational expensive.

It is worth noting that even for a single test, AnoGAN has very high computational cost during inference, as each test image has to go through an optimization process to find its corresponding place in the latent space (which requires approximately 200 iterations per test). In contrast, the proposed adGAN has no optimization process during inference and is fast for testing. We also experiment with a fast version of AnoGAN [6] (*f-AnoGAN* in Table 4). It improves the performance as well as the inference speed. Specifically, it achieves a slightly better AUC value (0.98) on ab-MNIST than adGAN (0.97). However, adGAN outperforms f-AnoGAN on both ISIC and BraTS, which are real clinical datasets.

ALOCC gives higher AUCs than AnoGAN, but still performs worse than adGAN. It obtains the AUCs of 0.91, 0.97, 0.87 for ab-MNIST, ISIC and BraTS, which are 6%, 1% and 5% lower than adGAN, respectively. Note that ALOCC gives higher AUCs on ISIC than that on ab-MNIST and BraTS. One possible reason is that skin patches are relatively homogeneous and easier to be reconstructed by the auto-encoder of ALOCC while brain patches contain some structural information, making the reconstruction more challenging.

In addition, in our experiments we find that ALOCC is difficult to train. It is important to train the auto-encoder and the discriminator with different learning rates: if the discriminator learns too fast, the auto-encoder would fail to reconstruct any images but only produce noise images; if the auto-encoder learns too fast and produces high-quality reconstructed image quickly, the discriminator would hardly distinguish the differences and degenerate, which always gives a probability predictions $p(x) \approx 0.5$ for all the inputs. Moreover, as stated in [14], it is also crucial to decide when to stop the training procedure of ALOCC. In contrast, adGAN originates from WGAN [10] and is easy to train without the mode collapse issue.

We plot anomaly scores of normal (green) and abnormal (red) test data from different methods on three datasets in Fig. 11. AdGAN (bottom row) gives more distinguishable class-specific distributions of the anomaly scores, while the other methods produce heavily overlapping distributions of the decision scores. Also, it can be observed that the effect of concentration loss in adGAN is distinctive: the score distribution of normal class is more compact and peak-shaped, making them more separable from the wide-spread score distribution of abnormal class.

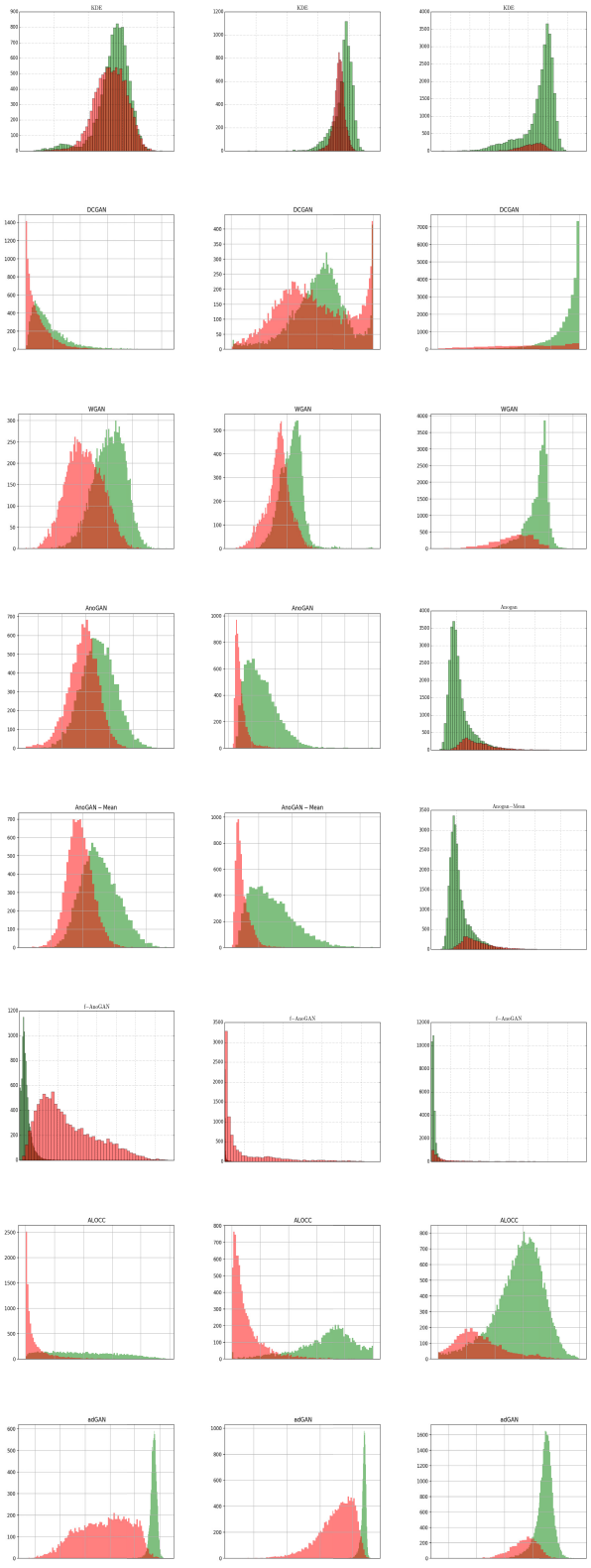


FIGURE 11. The plots of anomaly scores of normal (green) and abnormal (red) test data using different methods. From left to right: ab-MNIST, ISIC, and BraTS. Best viewed in colour.

E. ABLATION STUDY

Lastly, we conduct an ablation study to investigate the effects of two components in adGAN: the concentration loss and the fake pool generation.

1) EFFECT OF THE CONCENTRATION LOSS

To evaluate the effects of the concentration loss in Eq. 2, we vary the weight values α on all three datasets systematically, ranging from 0 to 5. Note that setting $\alpha = 0$ is equivalent to the original loss function without the concentration

term (Eq. 1). The curves of AUC values using different α are plotted in Fig 12.

From these curves, we have three observations. Firstly, setting $\alpha = 1$ achieves the highest AUC value on ISIC and BraTS, while it can get competitive results on ab-MNIST. Secondly, adGAN is not sensitive to α on ab-MNIST and BraTS as their AUC curves become stable and flat when $\alpha > 1$. However, it is more sensitive to α on ISIC. For example, the AUC value drops dramatically when α increases. This may be partially because for three-channel RGB image patches from the ISIC dataset, adGAN is more likely to overfit the training set with a large α value. Thirdly, setting $\alpha = 0$ (the original loss function in Eq. 1) does not perform well on all three datasets. Specifically, as shown in Table 5, adGAN($\alpha = 0$) gives AUC values of 0.85, 0.90, 0.77 for ab-MNIST, ISIC and BraTS, respectively, which are even lower than the baseline OCSVM. This is a solid evidence showing that the proposed concentration loss is a crucial part of adGAN for anomaly detection.

TABLE 5. The effects of concentration loss term and fake pool on ab-MNIST, ISIC and BraTS datasets.

	ab-MNIST	ISIC	BraTS
stdWGAN	0.86	0.64	0.68
2-class SVM	0.88	0.69	0.45
2-class CNN	0.80	0.72	0.54
adGAN($\alpha = 0$)	0.85	0.90	0.77
adGAN	**0.97**	**0.98**	**0.92**

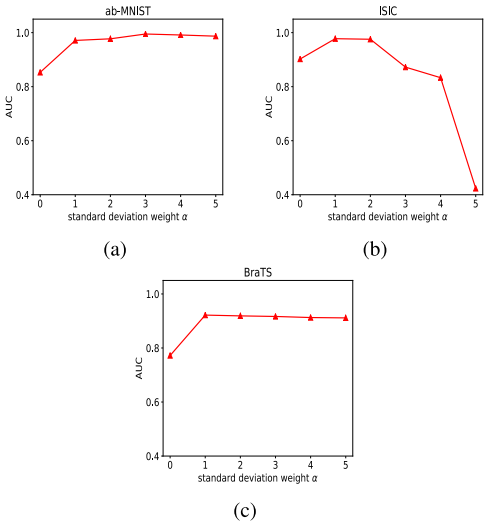


FIGURE 12. AUC values using different weights α in the concentration loss.

To further show the effectiveness of the concentration loss, we compare adGAN with two other baselines *after* the fake pool is generated: two-class SVM and two-class CNN with an architecture similar to the critic shown in Fig 1. We train these two powerful classifiers to separate the real data and the generated data in the fake pool and then use the trained models as the anomaly detectors. As shown in Table 5, these two baseline models perform much worse than adGAN and low AUC values indicate that sometimes they do not work at all. One possible reason is that the distribution of generated images in the fake pool and the distribution of real images have some overlap inevitably. Thus, maximizing the between-class distance only is not able to well separate these two distributions. The concentration loss additionally minimizes the within-class distance simultaneously, making the real images distributed more compactly, which is easier to be separated from other distributions (see Fig 11).

2) EFFECT OF FAKE POOL

To evaluate the effect of fake pool, we experiment by directly training a WGAN with Eq. 2 (termed as *stdWGAN* in Table 5) in the training set, without fake pool generation phase. And after training, the critic of *stdWGAN* is then used for anomaly detection. It is shown in Table 5 that *stdWGAN* fails and gives low AUC values, indicating that the fake pool generation is also an important component in adGAN for anomaly detection. Simply penalizing large standard deviation of the WGAN critic output is more likely to introduce overfitting during training and cannot generalize well for unseen data.

Based on all the ablation experiments, it can be concluded that *both* fake pool and concentration loss term are curial components of adGAN. Removing any one of them would make adGAN fail for anomaly detection.

V. DISCUSSION AND CONCLUSIONS

Obtaining data (e.g., images) from rare disease is rather difficult. Therefore, developing approaches that are capable of detecting anomaly in the presence of normal data is a very important area for biomedical research. Different from the reconstruction-based methods using GANs in the existing biomedical abnormal detection literature, in this paper we propose an alternative novel framework adGAN for anomaly detection with two key elements: fake pool generation and a new concentration loss. The fake pool counterfeits abnormal data by collecting the generated samples from the intermediate-state GAN, and the concentration loss penalizes large standard deviations of the critic outputs for real data, which helps to reduce the within-class distance and make two output distributions more separable when learning the decision boundary between normal data and simulated abnormal data. We experimentally show that both components are crucial for the framework and removing either of them would make the model fail to work. The proposed adGAN is evaluated on three datasets, including the modified MNIST dataset for synthetic anomaly detection, the ISIC'2016 for skin lesion detection, the BraTS'2017 for brain lesion detection and achieved AUC values of 0.97, 0.98, 0.92, respectively. The extensive experiments demonstrate that adGAN is consistently superior to the state-of-the-art anomaly detectors on all three datasets. Moreover, since adGAN is a discriminative

model and is able to give anomaly scores directly without the process of reconstruction, it enables real-time inference and can be deployed for clinical usage.

The significance of our proposed method to the anomaly detection community is three-fold: firstly, we propose an approach outperforming the state-of-the-arts; secondly, we present insights of how to develop GAN for anomaly detection by concentrating on the within class variance; and lastly, we provide a wide application in the medical domain where normal data is relatively easy to obtain but rather difficult to have abnormal data, especially for rare diseases. Future work includes the exploration of optimal early-stop criteria and the extension of adGAN to the condition where a small portion of lesion data is also available.

REFERENCES

[1] D. Sidibé, S. Sankar, G. Lemaitre, M. Rastgoo, J. Massich, C. Y. Cheung, G. S. W. Tan, D. Milea, E. Lamoureux, T. Y. Wong, and F. Mériaudeau, "An anomaly detection approach for the identification of DME patients using spectral domain optical coherence tomography images," *Comput. Methods Programs Biomed.*, vol. 139, pp. 109–117, Feb. 2017.

[2] G. Ziegler, G. R. Ridgway, R. Dahnke, and C. Gaser, "Individualized Gaussian process-based prediction and detection of local and global gray matter abnormalities in elderly subjects," *NeuroImage*, vol. 97, pp. 333–348, Aug. 2014.

[3] J. Mourão-Miranda, D. R. Hardoon, T. Hahn, A. F. Marquand, S. C. R. Williams, J. Shawe-Taylor, and M. Brammer, "Patient classification as an outlier detection problem: An application of the one-class support vector machine," *NeuroImage*, vol. 58, no. 3, pp. 793–804, Oct. 2011.

[4] P. Seeböck, S. Waldstein, S. Klimscha, B. S. Gerendas, R. Donner, T. Schlegl, U. Schmidt-Erfurth, and G. Langs, "Identifying and categorizing anomalies in retinal imaging data," 2016, *arXiv:1612.00686*. [Online]. Available: http://arxiv.org/abs/1612.00686

[5] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *Proc. Int. Conf. Inf. Process. Med. Imag.*, 2017, pp. 146–157.

[6] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "F-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks," *Med. Image Anal.*, vol. 54, pp. 30–44, May 2019.

[7] I. P.-A. J. Goodfellow, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, 2672–2680.

[8] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: http://arxiv.org/abs/1511.06434

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

[10] I. A. F. Gulrajani, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5767–5777.

[11] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[12] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[13] L. Chongxuan, T. Xu, J. Zhu, and B. Zhang, "Triple generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4088–4098.

[14] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3379–3388.

[15] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe, "Abnormal event detection in videos using generative adversarial nets," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 1577–1581.

[16] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep one-class classification," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4393–4402.

[17] L. Deecke, R. Vandermeulen, L. Ruff, S. Mandt, and M. Kloft, "Image anomaly detection with generative adversarial networks," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2018, pp. 3–17.

[18] I. Golan and R. El-Yaniv, "Deep anomaly detection using geometric transformations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 9758–9769.

[19] S. Pidhorskyi, R. Almohsen, and G. Doretto, "Generative probabilistic novelty detection with adversarial autoencoders," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 6822–6833.

[20] P. Perera, R. Nallapati, and B. Xiang, "OCGAN: One-class novelty detection using GANs with constrained latent representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2898–2906.

[21] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 3–9.

[22] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167

[23] D. Gutman, N. C. F. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC)," 2016, *arXiv:1605.01397*. [Online]. Available: http://arxiv.org/abs/1605.01397

[24] B. H. Menze *et al.*, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015.

[25] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features," *Sci. Data*, vol. 4, no. 1, Dec. 2017, Art. no. 170117.

[26] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, J. Freymann, K. Farahani, and C. Davatzikos, "Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection," *Cancer Imag. Arch.*, vol. 286, 2017.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

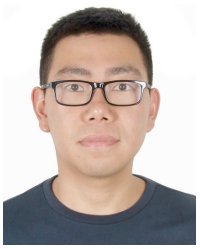
[28] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, Jul. 2001.

[29] H. Shen, R. Wang, J. Zhang, and S. J. McKenna, "Boundary-aware fully convolutional network for brain tumor segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2017, pp. 433–441.

[30] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. Evanston, IL, USA: Routledge, 2018.

[31] L. Zhang, J. Lin, and R. Karim, "Adaptive kernel density-based anomaly detection for nonlinear systems," *Knowl.-Based Syst.*, vol. 139, pp. 50–63, Jan. 2018.

[32] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, p. e3, Oct. 2016.



HAOCHENG SHEN received the Ph.D. degree in medical image analysis from the University of Dundee, U.K., in 2018. He is currently a Senior Research Engineer with the AI Lab, Tencent, Shenzhen, China. His research interests include medical image analysis, computer vision, and machine learning.

JINGKUN CHEN received the M.Sc. degree from the University of Dundee, U.K., in 2019. He is currently a Research Assistant with the Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China. His research interests include image processing, medical image analysis, and machine learning.

RUIXUAN WANG received the Ph.D. degree in computer vision from the National University of Singapore, in 2007. He was a Postdoctoral Researcher with the University of Dundee, U.K. He is currently an Associate Professor with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. His research interests include computer vision, medical image analysis, and machine learning.

JIANGUO ZHANG (Senior Member, IEEE) received the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2002. He was a Reader in computing with the University of Dundee, U.K. He is currently a Professor with the Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China. His research interests include visual surveillance, object recognition, image processing, medical image analysis, and machine learning. He currently serves as an Associate Editor for the IEEE TRANSACTIONS ON MULTIMEDIA, *IET Computer Vision*, and the *EURASIP Journal on Advances in Signal Processing*.