Latent Regularized Generative Dual Adversarial Network For Abnormal Detection

Chengwei Chen , Jing Liu , Yuan Xie* , Yin Xiao Ban , Chunyun Wu , Yiqing Tao and **Haichuan Song**

East China Normal University

{52184501028, 51174500035}@stu.ecnu.edu.cn, yxie@cs.ecnu.edu.cn, {51194501102, 51184501161,10161900112}@stu.ecnu.edu.cn, hcsong@cs.ecnu.edu.cn

Abstract

With the development of adversarial attack in deep learning, it is critical for abnormal detector to not only discover the out-of-distribution samples but also provide defence against the adversarial attacker. Since few previous universal detector is known to work well on both tasks, we consider against both scenarios by constructing a robust and effective technique, where one sample could be regarded as the abnormal sample if it exhibits a higher image reconstruction error. Due to the training instability issues existed in previous generative adversarial networks (GANs) based methods, in this paper we propose a dual auxiliary autoencoder to make a tradeoff between the capability of generator and discriminator, leading to a more stable training process and high-quality image reconstruction. Moreover, to generate discriminative and robust latent representations, the mutual information estimator regarded as latent regularizer is adopted to extract the most unique information of target class. Overall, our generative dual adversarial network simultaneously optimizes the image reconstruction space and latent space to improve the performance. Experiments show that our model has the clear superiority over cutting edge semi-supervised abnormal detectors and achieves the state-of-the-art results on the datasets.

1 Introduction

Anomaly detection attempts to detect abnormal samples that are drawn far away from the learned distribution of training samples. Heading into the deep learning era, deep neural networks are used as a general tool for anomaly detection task, especially in secure authentication scenarios (*e.g.* , intrusion detection and fraud detection). Since these techniques provide protection against abnormal activities, attackers will exploit some methods to circumvent these deep networks techniques. Therefore, abnormal detection task requires identifying not only the out of distribution samples, but also the adversarial attackers. To our best knowledge, few universal detector is known to work well on both scenarios. Hence, it is necessary for us to construct the robust and effective technique to against both of adversarial attacks and abnormalities.

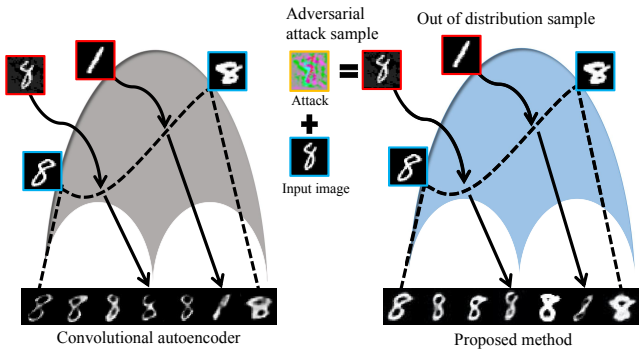

Figure 1: Due to the compactness of the proposed entire latent space corresponding to images from digit 8, all projections into the latent space in return produce images of digit 8, even for the out of distribution (digit 1) with higher reconstruction error. In attack example, compared with CAE, the proposed method could get rid of interference and restore to the original image.

A lot of GAN-style methods based on the image reconstruction error are proposed for abnormal detection task such as [Perera *et al.*, 2019]. However, using image reconstruction error to detect abnormal samples is not our original intension. Actually, our objective is to find more separable latent representation features between normalities and anomalies by optimizing the latent space. Besides, due to the imbalance of capability between generator and discriminator, the training of GAN-style method is always unstable such that mode collapse and non-convergence exist, leading to yield blurry reconstructions.

Motivated by the above limitations, we propose a novel dual adversarial autoencoder network under the semi-supervised learning framework. The auxiliary autoencoder is proposed to make a balance between generator and discriminator, leading to a more stable training process, which further fully taps the potential of GAN-style network with lower training loss and high-quality reconstruction, as verified in section 4.2. Besides, To further identify in a low-dimensional latent space, mutual information estimator regarded as a latent regularizer is expected to distinguish be-

*Contact Author

tween normal samples and abnormal ones in a discriminant way. To achieve this objective, the most unique information of the normal samples is extracted by mutual information estimator in the training process.

Compared with convolutional autoencoder, the proposed method could learn the most representive concept of the target class in latent manifold by using the latent regularizer and dual adversarial framework. To make an intuitive comparison, two learned latent manifolds of (digit 8) images are obtained by convolutional autoencoder and proposed method respectively, as shown in Figure 1. In our method, all projections into the latent manifold in return produce images of digit 8, even for the out-of-distribution samples (digit 1), which receives a higher reconstruction error. The main reason is that our method could capture the real concept of target class (digit 8) in the entire latent manifold under the constraint, leading to a more compact learned latent manifold, from which the abnormal samples could not be represented well. While, as for the latent space learned from the conventional convolutional autoencoder, the recovery of digital 1 is more like to itself, which is harmful to distinguish. The similar observation will be observed for adversarial attack samples.

Finally, we summarize the major contributions of this paper:

- We propose a novel latent regularized dual adversarial autoencoder network by jointly optimizing on reconstruction space and latent feature space. To generate discriminative and robust latent representations, the mutual information estimator is adopted to constrain the learned latent space in an unsupervised manner.
- To obtain more stable training process, an auxiliary autoencoder is introduced to make a balance of capability between generator and discriminator, further fully tapping the potential of GAN style network and achieving high-quality image reconstruction.
- To present the effectiveness and robustness of proposed method, multiple scenarios and several challenging detection tasks are adopted to the experiments, where the experimental results demonstrate that our method outperforms many state-of-the-art competitors.
- We construct a new stop sign dataset with more complex attacks such as BIM, DeepFool and FGSM, which will be used to measure the robustness and effectiveness of abnormal detection methods in adversarial examples scenarios.

2 Related Work

2.1 Adversarial Attack

The fast gradient sign method (FGSM) [Goodfellow *et al.*, 2014] attends to find a small perturbation by using the gradients of the neural network to create an adversarial attacks. The purpose of this method is to let the deep neural network make the wrong classification result. Instead of applying the perturbation in a single step, basic iterative method (BIM) [Kurakin *et al.*, 2016] based on the fast gradient sign method is applied multiple times with a smaller step size. In this method, the pixel values of intermediate results are clipped after each step to ensure that they are in a neighbourhood of the original image. Moosavi-Dezfooli [Moosavi-Dezfooli *et al.*, 2016] proposes DeepFool attack method, which makes a hyperplane separating each class. This method iteratively linearizes the decision boundary by using a L_2 minimization-based formulation to search for adversarial examples. To design more complex scenarios, all of these adversarial attack methods above are considered in the proposed dataset.

2.2 Out-of-distribution Samples Detection

Recently, deep learning based autoencoders [Rushe and Mac Namee, 2019] are used to learn the real distribution of normal behaviors and exploit reconstruction loss to detect anomalies. For example, variational autoencoder (VAE) tackles the problem by learning a mapping to a lower dimensional representation, where the real distribution is modeled. Out-of-distribution samples will obtain high image reconstruction error from the learned latent space. Some GAN-style work [Perera *et al.*, 2019; Sabokrou *et al.*, 2018] also rely on image reconstruction error to detect out-of-distribution samples. However, original objective of abnormal detection task is to make normal and abnormal samples more separable in the latent representation features. All of these methods ignore the importance of optimization in latent feature space. Besides, GAN-style methods suffer from training process issue, leading to the blurry reconstructions.

2.3 Adversarial Samples Detection

To detect the adversarial samples, multiple strategies are proposed to improve the robustness of deep networks. One part of methods [Tramèr *et al.*, 2017] is built upon the idea of adversary training by taking adversarial samples into the training process. Another way of strategies [Liang *et al.*, 2018] is to design and train a subsidiary model to detect adversarial samples. All of these methods need available adversarial samples. However, in the real world applications, it is impractical to take all types of adversarial samples (maybe unknown attack) into account, meanwhile collecting adversarial samples for training is costly and time-consuming in the supervised learning.

3 Proposed Method

3.1 Network Architecture

Proposed adversarial dual autoencoder network, which is shown in Figure 2, consists of four components: two autoencoder networks, a discriminator and a mutual information estimator.

The first autoencoder served as generator attempts to generate the input image to fool the discriminator. To avoid being fooled, the discriminator learns the real characteristic of input images to distinguish input images from generated images. Generator and discriminator compete with each other while collaborating to understand the underlying concept in the target class to obtain high-quality generated images.

The auxiliary autoencoder has the same structure as the generator without sharing parameters. To deal with the unstable training process issue of adversarial learning, the auxiliary

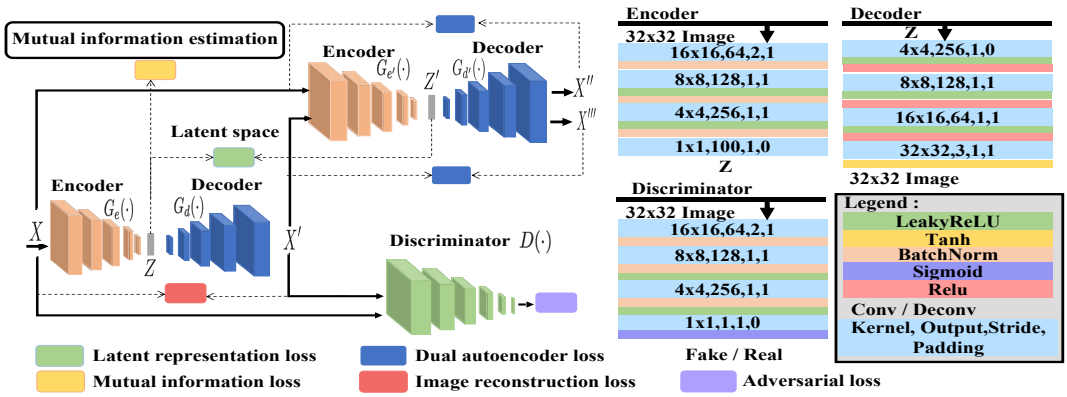

Figure 2: Our framework consists of two autoencoders, one discriminator and mutual information estimator.

autoencoder is proposed to make a balance between generator and discriminator, leading to a more stable training process and fully tapping the potential of GAN style network with lower training loss. The objective of almost all prior abnormal detection methods is to minimize the reconstruction error. They hope to learn more discriminative latent representations and use these representations directly to distinguish the normal samples from abnormal ones. However, the discriminative ability of the latent representations is not strongly related to the reconstruction error. They ignore the significance of optimization to the latent feature space. Therefore, We hope that the representation regularized by mutual information estimator helps us to identify the sample from the input images. In other words, the most unique information should be extracted from the input by mutual information estimator.

3.2 Overall Loss Function

In the training process, we define a loss function in equation (1), consisting of five terms, the adversarial loss, the image reconstruction loss, the mutual information loss, the latent representation loss and the dual autoencoder loss, which can be formulated as :

$$\mathcal{L} = w_i\mathcal{L}_{irec} + w_a\mathcal{L}_{adv} + w_z\mathcal{L}_{zrec} + w_e\mathcal{L}_e + w_d\mathcal{L}_{dual}, \quad (1)$$

where w_i, w_a, w_z, w_e, and w_d are the weighting parameters balancing the impact of individual item to the overall object function.

Adversarial Loss. Adversarial loss is adopted to train the first generator G and discriminator D. According to the previous work, this adversarial game could be formulated as:

$$\begin{aligned}\mathcal{L}_{adv} = &\min_G \max_D \left(E_{\boldsymbol{x}\sim p_{\mathbf{x}}}[\log(D(\mathbf{x}))]\right. \\ &\left.+E_{\boldsymbol{x}\sim p_{\mathbf{x}}}[\log(1-D(G(\mathbf{x})))]\right).\end{aligned} \quad (2)$$

Image Reconstruction Loss. In order to fool the discriminator, the generator tries to generate high-quality images in the process of training by minimizing the pixel-wise error between original input images x and generated images $G(x)$.

$$\mathcal{L}_{irec} = \mathbb{E}_{x\sim p_{\mathbf{x}}}\|x - G(x)\|_1 \quad (3)$$

Latent Representation Loss. Only for target class samples, the auxiliary autoencoder can reconstruct the latent representation z well from generated image x'. Besides, the latent regularizer might incur the distribution distortion in latent feature space, the feature representation z' can be regarded as the anchor to prevent z from drifting. Hence, we consider to add a constraint by minimizing the distance between latent feature of input images $G_e(x)$ from generator and encoded latent feature of generated image $G_{e'}(x')$ from auxiliary encoder as follows.

$$\mathcal{L}_{zrec} = \mathbb{E}_{x\sim p_{\mathbf{x}}}\|G_e(x) - G_{e'}(x')\|_2 \quad (4)$$

Dual Autoencoder Loss. we regard the auxiliary autoencoder as a discriminator D'. Compared with conventional adversarial loss, we match distribution between losses, $\mathcal{L}_{direc}$ and $\mathcal{L}_{girec}$, not between samples. The training objective of this discriminator is to reconstruct the realistic inputs x faithfully while fail to do so for generated input $G(x)$, as shown below:

$$\mathcal{L}_{girec} = \|x - D'(x)\|_1, \mathcal{L}_{direc} = \|G(x) - D'(G(x))\|_1, \quad (5)$$

$$\mathcal{L}_{dual} = \mathcal{L}_{girec} - k\mathcal{L}_{direc}, \quad (6)$$

where k controls how much emphasis is put on the pixel-wise error of generated input $\mathcal{L}_{direc}$ during gradient descent.

Mutual Information Loss. Mutual information [Yang *et al.*, 2019] measures the essential relevance of two instances, which can also be adopt to estimate the similarity between the input sample X and the latent feature Z. The mutual information could be formulated as:

$$I(X,Z) = \iint p(z|x)p(x)\log\frac{p(z|x)}{p(z)}dxdz, \quad (7)$$

where $p(x), p(z|x)$ and $p(z)$ are the distribution of original input data, latent feature and latent space respectively, and then $p(z) = \int p(z|x)p(x)dx$. The aim of encoder is to extract the most unique feature from the original input sample. To achieve it, the mutual information between input sample

and feature should be as large as possible, which could be formulated as:

$$p(z|x) = \max_{p(z|x)} I(X, Z). \tag{8}$$

In addition, To make latent space more regular, the hidden feature vector obeys the prior distribution of the standard normal distribution with KL divergence, which is combined with equation (7) and equation (8) with different weights. We can get the minimum objective function.

$$\begin{aligned} p(z|x) = \min_{\theta_e}\{&-\beta I(X, Z) \\ &+\gamma \mathbb{E}_{x\sim p(x)}[KL(p(z|x)\|q(z))]\}. \end{aligned} \tag{9}$$

We hope to maximize mutual information between input X and feature. However, the KL divergence is unbounded. In order to optimize more effectively and enlarge the distance between $p(z|x)p(x)$ and $p(z)p(x)$, we do not use KL divergence and change a metric with an upper bound: JS divergence.

$$\begin{aligned} p(z|x) = \min_{\theta_e}\{&-\beta JS(p(z|x)p(x), p(z)p(x)) \\ &+\gamma \mathbb{E}_{x\sim p(x)}[KL(p(z|x)\|q(z))]\}. \end{aligned} \tag{10}$$

According to the definition of variational estimation of JS divergence [Nowozin *et al.*, 2016], we substitute $p(z|x)p(x)$ and $p(z)p(x)$ into formula:

$$\begin{aligned} \mathcal{L}_{e_g} = \min_{\theta_e} \{&-\beta \left(\mathbb{E}_{(x,z)\sim p(z|x)p(x)}[\log \sigma(T(x, z))]\right. \\ &\left.+\mathbb{E}_{(x,z)\sim p(z)p(x)}[\log(1-\sigma(T(x, z)))]\right) \\ &+\gamma \mathbb{E}_{x\sim p(x)}[KL(p(z|x)\|q(z))]\}. \end{aligned} \tag{11}$$

This approach follows Mutual Information Neural Estimation[Belghazi *et al.*, 2018], which estimates mutual information by training a discriminator $\sigma(T(x, z))$ to distinguish positive sample pair, consisting the input image and corresponding latent feature z, from the negative sample pair, including the same input image and noise feature vector z_t randomly selected from the disturbed batch. equation (11) shows the global mutual information between X and Z.

We also consider local mutual information $\mathcal{L}_{e_l}$ in Figure 3, which measures the relevance between the feature map of input image and its latent feature. The discriminator distinguishs the positive pair (feature map of input samples and its latent feature) from the negative pair (feature map of random image and the same latent feature as above). Therefore, the mutual information loss includes global mutial information loss and local mutial information loss, which can be formulated as follows: $\mathcal{L}_e = \mathcal{L}_{e_g} + \mathcal{L}_{e_l}$.

3.3 Optimization

In the process of training, the generator and the auxiliary autoencoder are trained by optimizing image reconstruction loss, latent representation loss, dual autoencoder loss and mutual information loss. All of components are randomly initialized. We use adaptive moment estimation(Adam) as the optimizer and set the initialized learning rate as 0.002. For all

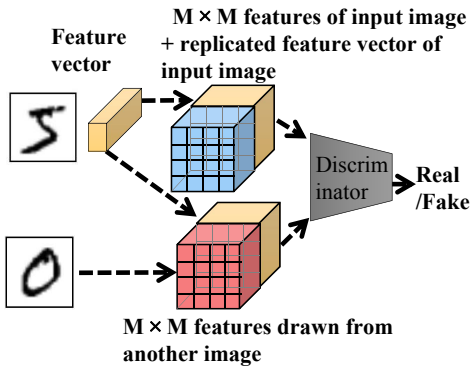

Figure 3: Local mutual information estimation. The feature map is obtained from the middle layer of convolutional network. The latent feature is extracted from the encoder.

experiments, we train on the standard training set and test on the validation set. Besides, data augmentation (random cropping and horizontal flipping) and normalization (subtracted and divided sequentially by mean and standard deviation of the training images) is applied to all the training images.

4 Experiments

4.1 Experimental Setting

Datasets. We evaluate proposed method on the well-known COIL100, MNIST, CIFAR10 and fMNIST datasets in out-of-distribution samples detection experiments. In addition, DCASE dataset is considered in the experiment, which is a public available acoustic novelty detection dataset.

To evaluate adversarial attacks detection task, we use the GTSRB dataset [Stallkamp *et al.*, 2011] including adversarial boundary attack on stop signs boards. Moreover, for physical adversarial examples detection, Replay-Attack [Chingovska *et al.*, 2012] dataset and CASIA-MFSD dataset consider different attacks in face anti-spoofing detection task.

Evaluation Methodology. Two protocols in the literature are proposed for abnormal detection [Perera *et al.*, 2019].

Protocol 1 : 80% of in-class samples are regarded as normal class, the rest of 20% of in-class samples are adopt in testing process. Out-of-class samples are serviced as abnormal class, which are randomly selected from testing dataset, constituting half of the test set.

Protocol 2 : All of in-class samples from the training part of dataset is only used to train in the proposed method. Testing data of all classes are used for testing.

Evaluation Measures. The performance metrics we employed are Area Under Curve (AUC) and Half Total Error Rate (HTER) [Bengio and Mariéthoz, 2004].

Implementation Details. For a given test point x , we can naturally define an anomaly score s for proposed method by calculating the distance between input sample and corresponding generated image $G'_d(G'_e(G(x)))$, as follows $s = \|x - G'_d(G'_e(G(x)))\|_1$. We implement our approach in PyTorch by optimizing the weighted loss (defined in equation (1)) with the weight values $w_i = 1$, $w_a = 5$, $w_z = 1$,

	MNIST	STOP SIGN
Single autoencoder	0.510	0.194
Single autoencoder + Discriminator	0.531	0.403
Dual autoencoder + Discriminator	0.972	0.894
Dual autoencoder + Discriminator + Mutual information	0.985	0.921

Table 1: Ablation study for proposed method performed on MNIST and STOP SIGN (AUC).

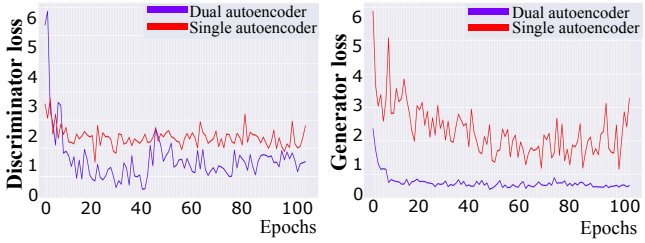

Figure 4: Training loss comparison between single autoencoder and dual autoencoder.

$w_e = 0.05$ and $w_d = 1$, which are empirically chosen to yield optimum results. For dual autoencoder loss, the parameter k is set to 0.4. The experiments are carried out on a standard PC with a NVIDIA-1080 GPU and a multi-core 2.1 GHz CPU. The detailed structures of the autoencoder, the auxiliary autoencoder, mutual information estimator and discriminator are described in Figure 2.

4.2 Ablation Study

Setup. To investigate the effectiveness of each additional component, as illustrated in Table 1, we conduct experiments based on the these four scenarios by using Protocol 2 in out of distribution samples (MNIST) and adversarial samples (STOP SIGN) dataset.

Results. Mean AUC for each class of MNIST dataset and STOP SIGN dataset is presented in Table 1. The potential of GAN style network is fully tapped by dual autoencoder and makes a great improvement for the performance. Besides, Full proposed model generates discriminative and robust latent representations by using the mutual information estimator to regularize the learned latent space. The performance improves further by 1.3% and 2.7% respectively in MNIST and STOP SIGN.

To further prove the stabilization effect of dual autoencoder component, we show the plot of training loss for discriminator and generator, as illustrated in Figure 4. To sum up, thanks to the dual autoencoder, we can easily observe that the dual autoencoder structure could make a balance between the capability of generator and discriminator and further tap the potential of GAN style network with lower training loss, leading to a more robust and accurate detector.

4.3 Out-of-distribution Samples Detection

Setup. In this subsection, we present that the proposed method has the clear superiority over cutting-edge semi-supervised abnormal detectors(*i.e.,* ALOCC [Sabokrou *et al.*, 2018],OCGAN [Perera *et al.*, 2019]). We consider to use both of protocols in the experiments. MNIST, COIL and fMNIST datasets are applied in the protocol 1. MNIST and CIFAR10 dataset are evaluated by the protocol 2. We also evaluate the DCASE dataset in acoustic novelty detection task by protocol 2. DCASE dataset includes three different abnormal event (*i.e.,* gunshot, babycry and glassbreak). All of these abnormal event audios are artificially mixed with background audios respectively which includes 15 different kinds of environmental settings (*i.e.,* home, bus, and train).

Methods	MNIST	COIL	fMNIST
ALOCC DR ('18)	0.88	0.809	0.753
ALOCC D ('18)	0.82	0.686	0.601
DCAE ('14)	0.899	0.949	0.908
GPND ('18)	0.932	0.968	0.901
OCGAN ('19)	0.977	0.995	0.924
Proposed method	**0.985**	**1.0**	**0.995**

Table 2: Mean One-class abnormal detection using Protocol 1.

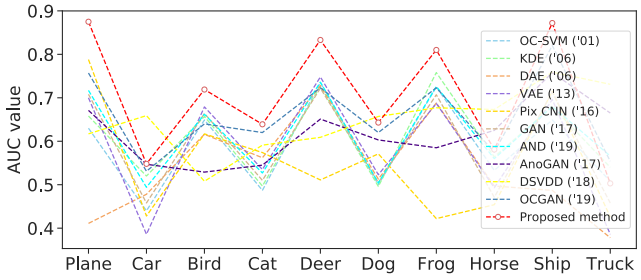

Figure 5: Class designated as normal class

Results. When protocol 1 is used in MNIST, COIL and fMNIST, proposed model yields an improvement in Table 2. For CIFAR10 dataset in protocol 2, as it is shown in Figure 5. Our method achieves the best performance compared with other compatitors. In term of AUC value, it is illustrated by using red curves in Figure 5. For MNIST dataset in protocol 2, proposed method is on par with other state-of-art approaches, which is presented in Figure 6.

In acoustic anomaly detection task, proposed method aims at distinguishing abnormal acoustic signals from the normal ones. The performance of three models across the 15 datasets is shown in Table 3. We find that the proposed model consistently outperforms WaveNet [Rushe and Mac Namee, 2019] in almost all datasets, with a tie in a residential area scenario.

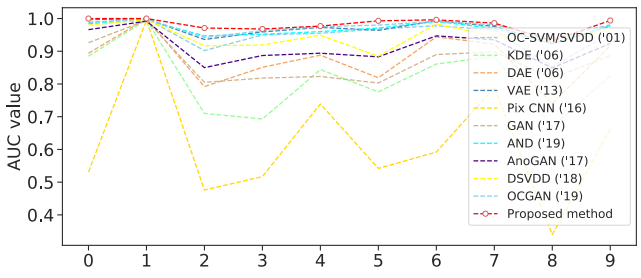

Figure 6: Digit designated as normal class

Loss composition	CAE ('16)	WaveNet ('19)	Proposed method
Beach	0.69	0.72	**0.82**
Bus	0.79	0.83	**0.96**
Cafe/restaurant	0.69	0.76	**0.80**
Car	0.79	0.82	**0.99**
City center	0.75	0.82	**0.89**
Forest path	0.65	0.72	**0.78**
Grocery store	0.71	0.77	**0.90**
Home	0.69	0.69	**0.90**
Library	0.59	0.67	**0.89**
Metro station	0.74	0.79	**0.89**
Office	0.78	0.78	**0.87**
Park	0.70	0.80	**0.95**
Residential area	0.73	**0.78**	**0.78**
Train	0.82	0.84	**0.92**
Tram	0.80	0.87	**0.97**

Table 3: AUC scores for all methods on each dataset for acoustic anomaly detection.

Methods	Boundary	FGSM	BIM	Deepfool
OC-SVM/SVDD ('01)	67.5±1.2	19.9±0.1	19.9±0.9	18.6±0.1
KDE ('06)	60.5±1.7	69.3±0.4	53.3±0.4	51.8±0.5
IF ('11)	73.8±0.9	44.4±0.8	53.9±0.6	51.1±0.9
DCAE ('15)	79.1±3.0	36.7±0.5	45.1±0.0	36.8±0.5
MAHALANOBIS ('17)	56.0±0.5	58.1±0.3	56.1±0.5	56.7±0.9
SOFT-BOUND ('18)	77.8±4.9	82.7±0.5	74.9±0.8	78.4±0.4
ONE-CLASS ('18)	80.3±2.8	82.8±0.1	81.8±0.5	65.8±0.7
RCAE ('17)	87.4±2.7	57.6±0.7	64.5±0.8	46.3±0.3
Ours	**90.5±1.6**	**84.4±0.2**	**85.9±0.9**	**84.6±0.5**

Table 4: Average AUCs in % with StdDevs (over 10 seeds) per method on GTSRB stop signs with adversarial attacks.

4.4 Adversarial Attack Detection

Setup. To evaluate the robustness and anti-noise ability of the proposed method, we present the proposed method in two attack scenarios. Stop sign adversarial samples are generated from randomly drawn stop sign images of the test set by using Boundary Attack. Besides, to further evaluate the proposed method on more complex adversarial attack scenario, three adversarial attack datasets based on the stop sign images are generated by using FGSM, BIM and DeepFool respectively in our proposed dataset. The main competitors include RCAE [Chalapathy *et al.*, 2017], ONE-CLASS [Chalapathy *et al.*, 2018] and SOFT-BOUND[Ruff *et al.*, 2018].

Another scenario belongs to physical adversarial attack, which is face anti-spoofing detection task. Spoof face data poses a great threat to face recognition systems [Patel *et al.*, 2016]. Presentation attacks (abbreviated as PA), including printed paper face, replaying a video, and wearing a mask, are one of the most prevalent face spoofs. To present robustness and generalization of approaches, we consider training on the training set of the CASIA-MFSD dataset and testing on the testing set of the Replay-Attack dataset. We then conduct the opposite experiment. The main competitors include GFA-CNN[Tu *et al.*, 2019], De-Spoof [Jourabloo *et al.*, 2018], Auxiliary [Liu *et al.*, 2018], Color Tex [Boulkenafet *et al.*, 2016] and Color LBP [Boulkenafet *et al.*, 2015].

Figure 7: Most anomalous stop signs detected by proposed method. Adversarial examples are highlighted in green.

Methods	Train CASIA MFSD	Test Replay Attack	Train Replay Attack	Test CASIA MFSD	Average
LBP ('13)	47.0%		39.6%		43.3%
LBP-TOP ('13)	49.7%		60.6%		55.2%
Motion ('13)	50.2%		47.9%		49.1%
CNN ('14)	48.5%		45.5%		47.0%
Color LBP ('15)	37.9%		35.4%		36.7%
Color Tex ('16)	30.3%		37.7%		34.0%
Color SURF ('18)	26.9%		**23.2%**		25.1%
Auxiliary ('18)	27.6%		28.4%		28.0%
De-Spoof ('18)	28.5%		41.1%		34.8%
GFA-CNN ('19)	**21.4%**		34.3%		28.0%
Proposed method	22.3%		24.6%		**23.4%**

Table 5: Classification performance of the proposed approach in terms of HTER (%). The algorithm are trained using the CASIA-MFSD dataset and tested on the Replay-Attack dataset, and vice versa.

Results. For stop sign adversarial samples detection, Table 4 presents proposed method exactly detect the adversarial attack and achieves the best performance. In addition, Figure 7 shows the most anomalous samples detected by the proposed method which consists of adversarial attack samples and incorrectly cropped samples. For face anti-spoofing detection task, Table 5 confirms that the proposed method achieves comparable performance ($HTER = 0.223$) on the Replay-Attack testing set which includes different types of spoofing attacks. In the opposite experiment, our method achieves competitive performance ($HTER = 0.246$) for the cross testing on the testing set of the CASIA-MFSD dataset. We achieve the best average value in crossing datasets setting.

5 Conclusion

To detect the out of distribution samples and adversarial attacks in the abnormal detection task, we proposed a novel dual adversarial autoencoder framework. To deal with the instability issue of GAN methods, the auxiliary autoencoder makes a balance of capacity between generator and discriminator, leading to a more stable training process. In addition, to further identify the latent space, mutual information estimator is adopted to extract the unique characteristics of the target class and regularize the latent representation. Extensive experiments have been conducted on public available datasets and more complex proposed datasets, showing high generalization capability of trained models.

Acknowledgments

We thank for the support from National Natural Science Foundation of China (61772524, 61902129, 61972157, 61876161, 61701235, 61373077), National Key Technologies R&D Program (SQ2019YFC150159), Shanghai Pujiang Talent Program (19PJ1403100), Beijing Municipal Natural Science Foundation (4182067), the Science and Technology Commission of Pudong (NO. PKJ2018-Y46) and Shanghai Jiaotong University Translational Medicine Cross Foundation (ZH2018ZDA25).

References

[Belghazi *et al.*, 2018] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

[Bengio and Mariéthoz, 2004] Samy Bengio and Johnny Mariéthoz. A statistical significance test for person authentication. In *Proceedings of Odyssey 2004: The Speaker and Language Recognition Workshop*, number CONF, 2004.

[Boulkenafet *et al.*, 2015] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face anti-spoofing based on color texture analysis. In *2015 IEEE international conference on image processing (ICIP)*, pages 2636–2640. IEEE, 2015.

[Boulkenafet *et al.*, 2016] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face spoofing detection using colour texture analysis. *IEEE Transactions on Information Forensics and Security*, 11(8):1818–1830, 2016.

[Chalapathy *et al.*, 2017] Raghavendra Chalapathy, Aditya Krishna Menon, and Sanjay Chawla. Robust, deep and inductive anomaly detection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 36–51. Springer, 2017.

[Chalapathy *et al.*, 2018] Raghavendra Chalapathy, Aditya Krishna Menon, and Sanjay Chawla. Anomaly detection using one-class neural networks. *arXiv preprint arXiv:1802.06360*, 2018.

[Chingovska *et al.*, 2012] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG)*, pages 1–7. IEEE, 2012.

[Goodfellow *et al.*, 2014] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[Jourabloo *et al.*, 2018] Amin Jourabloo, Yaojie Liu, and Xiaoming Liu. Face de-spoofing: Anti-spoofing via noise modeling. In *Proceedings of the ECCV*, pages 290–306, 2018.

[Kurakin *et al.*, 2016] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[Liang *et al.*, 2018] Bin Liang, Hongcheng Li, Miaoqiang Su, Xirong Li, Wenchang Shi, and XiaoFeng Wang. Detecting adversarial image examples in deep neural networks with adaptive noise reduction. *IEEE Transactions on Dependable and Secure Computing*, 2018.

[Liu *et al.*, 2018] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *Proceedings of the IEEE Conference on CVPR*, pages 389–398, 2018.

[Moosavi-Dezfooli *et al.*, 2016] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on CVPR*, pages 2574–2582, 2016.

[Nowozin *et al.*, 2016] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, pages 271–279, 2016.

[Patel *et al.*, 2016] Keyurkumar Patel, Hu Han, and Anil K Jain. Secure face unlock: Spoof detection on smartphones. *IEEE Transactions on Information Forensics and Security*, 11(10):2268–2283, 2016.

[Perera *et al.*, 2019] Pramuditha Perera, Ramesh Nallapati, and Bing Xiang. Ocgan: One-class novelty detection using gans with constrained latent representations. In *Proceedings of the IEEE Conference on CVPR*, pages 2898–2906, 2019.

[Ruff *et al.*, 2018] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402, 2018.

[Rushe and Mac Namee, 2019] Ellen Rushe and Brian Mac Namee. Anomaly detection in raw audio using deep autoregressive networks. In *ICASSP 2019-2019 IEEE*, pages 3597–3601. IEEE, 2019.

[Sabokrou *et al.*, 2018] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. Adversarially learned one-class classifier for novelty detection. In *Proceedings of the IEEE Conference on CVPR*, pages 3379–3388, 2018.

[Stallkamp *et al.*, 2011] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: A multi-class classification competition. In *IJCNN*, volume 6, page 7, 2011.

[Tramèr *et al.*, 2017] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

[Tu *et al.*, 2019] Xiaoguang Tu, Jian Zhao, Mei Xie, Guodong Du, Hengsheng Zhang, Jianshu Li, Zheng Ma, and Jiashi Feng. Learning generalizable and identity-discriminative representations for face anti-spoofing. *arXiv preprint arXiv:1901.05602*, 2019.

[Yang *et al.*, 2019] Xu Yang, Cheng Deng, Feng Zheng, Junchi Yan, and Wei Liu. Deep spectral clustering using dual autoencoder network. In *Proceedings of the IEEE Conference on CVPR*, pages 4066–4075, 2019.