Adaptive Adversarial Latent Space for Novelty Detection

BIN JIANG®, FANGQIANG XU, YUN HUANG, CHAO YANG®, WEI HUANG®, AND JUN XIA

Corresponding author: Bin Jiang (jiangbin@hnu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62072169 and Grant 61702176.

ABSTRACT Novelty detection is a challenging task of identifying whether a new sample obeys to a known class. Note that the boundary between normal and novel is not clear enough in existing works, resulting from adequately reconstructing novel samples or crudely reconstructing normal samples. To tackle the above issues, we propose a general framework named Adaptive Adversarial Latent Space (AALS), which mainly consists of two components, Adaptive Latent Space Generator (ALSG) and Constrained-based Adversarial AutoEncoder (CAAE). ALSG is established to obtain the real latent space distribution by adaptive mapper and the manner of adversarial learning. Moreover, CAAE is presented to obtain a better boundary between normal and novel by Global and Local Channel-wise Attention (GLCA), which is proposed to reweight the global and local information in different channels. Given a sample, the latent space representation $z$ of the sample is firstly obtained by encoder with GLCA. Next, the output of ALSG is applied to constrain the latent space by latent adversarial loss. Further, $z$ is transformed into the decoder to reconstruct the given sample with GLCA. Experiments on three available datasets demonstrates the effectiveness of the proposed method, which achieves state-of-the-art performance compared with other methods.

INDEX TERMS Novelty detection, adaptive latent space, latent constraint, channel attention.

I. INTRODUCTION

Novelty detection is defined as how to distinguish the normal and novel samples, which has a number of applications, such as video surveillance and medical diagnosis. Different from other tasks in computer vision, the key problem of novelty detection is how to ensure intra-class consistency and inter-class distinction given normal class only. The crucial challenge of novelty detection is how to find an appropriate boundary between normal and novel when only given normal samples.

Convolutional autoencoder (AE) structure and its variants are widely applied in novelty detection and related fields [1]–[7]. Compared with traditional methods, AE allows to extracting better features and achieves better reconstruction results. In this case, the reconstruction error of novel input will be low since they are close to the normal class. For better performance, adversarial learning methods are introduced to detect novelties [8]–[12]. In general, they not only change the training mode of the network, but also take the output of the discriminator as one of the criterions of novelty score.

Nevertheless, the above-mentioned works are all based on an assumption that may not always hold: novel samples are incapable of reconstructing well. The source of the problem is to “generalized” representation capacity of the overall network. In order to constrain the expression ability of network, OCGAN [13] presents a network structure of constraining latent space. They adopt a discriminator to constrain the latent space into a known prior distribution, and another discriminator to boost the quality of generated images. To obtain a better boundary, a sampling technique based on gradient descent is introduced. Using the samples of prior distribution that have achieved good results, OCGAN intends to generate normal class examples as much as possible. However, OCGAN relies on more convolutions and sophisticated operations, leading to inelegant implementation of the model.

We note that it is not sometimes effective to assume that the latent space obeys to a known prior distribution. In this paper, we propose a general unsupervised novelty detection model AALS by making a trade-off between the expression ability of the whole network and image quality. Specifically, AALS includes an Adaptive Latent Space Generator (ALSG) and a Constrained based Adversarial AutoEncoder (CAAE). Firstly, a number of samples are randomly selected from the
normal distribution to train ALSG and then its parameters are fixed. Notably, ALSG contains an adaptive mapper, a generator and a discriminator. Secondly, CAAE utilizes the output of adaptive mapper to generate adaptive latent space distribution of known class. To reconstruct a given input, the overall network is optimized by employing pixel adversarial loss and pixel reconstruction loss. Note that channel attention is often utilized in convolution layer to extract features better such as SENet. However, the channel attention usually takes the feature map as a whole, which is not necessarily a good decision. Therefore, GLCA is proposed to extract features better by integrating global and local information, and its effectiveness is also proved in subsequent experiments. In addition, we also introduce feature reconstruction loss intention to generate higher quality images. Finally, in the test phase, the sample is only fed into CAAE, and the novelty score is composed of pixel-feature reconstruction loss. Compare with other methods, our model requires no assumptions about data or latent space, which makes the model elegant and efficient.

The key contributions of this paper are four-fold:

- An adaptive adversarial model named AALS is proposed for novelty detection. Different from other approaches, AALS obtains a better trade-off between the expression ability of the whole network and image quality.
- Global and Local Channel Attention is proposed to obtain better features in CAAE. GLCA integrates the global and local information of features, which is more convincing and reasonable than existing channel attention mechanism.
- An adaptive latent space generator is proposed, which constructs an appropriate latent space distribution according to each class. It makes our model requires no assumptions about the latent space.
- We achieve state-of-the-art performance in three popular benchmarks available datasets [14]–[16]. As a side contribution, we will release the codes.

II. RELATED WORK

Novelty detection is a field closely related to outlier detection, anomaly detection, etc. There are many traditionally methods for novelty detection and related fields, such as OCSVM [17], PCA [18]. With the advent of deep neural networks, most of the methods utilized in novelty detection can be summarized as follows: Auto Encoder (AE) based methods, Adversarial AutoEncoder (AAE) based methods and Latent Constraint (LC) based methods.

A. AE-BASED METHODS

The end-to-end models have recently achieved promising results, attracting more and more attention in novelty detection. Sakurada and Yairi [19] first utilize and prove the effectiveness of autoencoders in anomaly detection. In [20], stacking denoising autoencoders is combined with SVM. Chong and Tay [21] then propose a spatiotemporal autoencoder for abnormal event detection in the video domain.

By splitting the input of autoencoders, Zhou and Paffenroth [22] present robot deep autoencoders to improve the detection efficiency. Thereafter, Zong et al. [23] integrate autoencoders and gaussian mixture model for anomaly detection. Li et al. [24] also propose a multi-scale method to extract spatiotemporal information with Gaussian mixture model. The latest methods based on AE is IAE [25] and MBC [26], the former introduces channel attention mechanism into novelty detection and proposes a method of minimizing cross entropy to limit the hidden space, while the latter describes the quality of reconstruction from the perspective of gradient. However, though AE adopts the reconstruction loss, it is not adept at reconstructing images, no matter whether the input is normal or not, leading to the unacceptable performance. In contrast, our method aims to generate high quality reconstruction results by utilizing pixel-feature reconstruction loss.

B. AAE-BASED METHODS

Increasingly adversarial based models [4], [27] are proposed in novelty detection to address the defect of AE structure. Schlegl et al. [8] first apply DCGAN [28] in anomaly detection, and propose a mapping from image to latent space by back propagation in inference phase. Unfortunately, there is no mapping between image and latent space in the training phase, causing a lousy mapping in inference. Therefore, Zenati et al. [12] propose ALAD, a bi-directional GAN structure ensuring the mapping between the image and the latent space, so as to improve the detection results. However, the training process of their network structure is complicated, which requires a lot of adjustment to achieve training stability. Ravanbakhsh et al. [9] adopt Flow2Pixel and Pixel2Flow generators for training. During testing, two networks take each frame as input and the outputs jointly determine whether it is novel or not. Li et al. [29] propose a new anomaly score function and a spatio-temporal framework combined by U-net and adversarial learning. Similarly, Dong et al. [30] propose a new approach with a dual discriminator-based generative adversarial network and U-net structure. ALOCC [10] combines denoising AE and adversarial AE, which is based on the assumption that the novel samples may not get a high discriminator score. Consequently, ALOCC applies the output of the discriminator as the novelty score. More recently, Salehi et al. [31] makes the network structure more effective by randomly combining the input of the network into several blocks. Both AE and AAE-based methods are based on the following assumptions: normal samples achieve favorable reconstruction and novel samples are reversed. In many cases, this assumption may not be hold. To this end, the proposed approach alleviates these problems by providing a constraint for the latent space in AE or AAE structure.

C. LC-BASED METHODS

With the development of the science and technology, researchers turn their attention to latent space with AE structure. They aim to constrain the expression ability of the whole network by constraining the latent space. GPND [32] first
FIGURE 1. The overall framework of AALS, which contains two components ALSG and CAAE. During training, the ALSG is pre-trained and fixed before CAAE. Note that when ALSG and CAAE are trained, the VGG Layer is to calculate the pixel-feature loss using the generated image and the ground truth. During testing, the sample is only transformed into CAAE to obtain pixel-feature reconstruction loss which is combined into a novelty score. In addition, the latent representation $z$ is constrained by the output of Adaptive Mapper. All details of the network structure is established in Table 5.

D. CHANNEL ATTENTION

Since Hu et al. [35] propose SENet for classification task, a large amount of channel attention has been proposed in various fields of computer vision. Woo et al. [36], propose a general attention framework which integrates spatial attention and spatial attention. Huang et al. [37], combine channel attention and UNet, which apply attention mechanism in skip-connection. Guo et al. [25], introduce channel attention to the field of novelty detection. Lee and Cho [38], propose locally adaptive channel for denoising images. However, the aforementioned works either consider global information or local information, which is unreasonable and incomplete. Therefore, we propose Global and Local Channel Attention which utilizes both global and local information simultaneously.

III. PROPOSED METHOD

The proposed AALS model consists of two major components: Adaptive Latent Space Generator (ALSG) and Constrained-based Adversarial Auto Encoder (CAAE). AALS aims to obtain a clear boundary in normal and novel classes by making a trade-off between the expression ability of the whole network and image quality. As depicted in Figure 1, ALSG is pre-trained and fixed using the data sampled from normal distribution. In CAAE, given an input, we firstly feed it into encoder to obtain the latent space $z$, and $z$ is constrained by the output of Adaptive Mapper in ALSG. The reconstruction output is obtained through decoder. In the following sections, we will elaborate the design of ALSG and CAAE in detail.

A. PROBLEM DEFINITION

The key problem of novelty detection is how to distinguish the normal class from the novel class on the test set when

...
training on the only normal class. Similar to OCGAN [13], we train a model to limit the expression ability of the network to known class to distinguish normal and novel. When a model is trained, any new sample will be reconstructed into a known class as much as possible. Therefore, the value of reconstruction loss for normal class is very small, while that of novel class is the opposite, so it is easy to distinguish between normal and novel samples. As shown in Figure 3, assume normal class is digital 8, AALS tends to generate a reconstruction image of known class “8” regardless of the digital of real samples in testing, so the value of reconstruction loss for novel class is large. In CAAE, the function of the encoder is to obtain the latent space representation of the image. Due to the limitation of ALSG, the encoder tends to generate the latent space of known(normal) class. In addition, since the input of the generator is a constrained latent space, the generator generates images that are close to the normal class.

B. ADAPTIVE LATENT SPACE GENERATOR

It is unreasonable to assume that the latent space obeys to a prior distribution, so we propose an adaptive mapper which uses the prior distribution to obtain a specific distribution. Similar to StyleGAN [39], a sample $x$, as the input is randomly calculated from the prior distribution $N(0, 1)$, and the sample $w$ of an adaptive specific distribution is obtained through the adaptive mapper. Thereafter, the sample $w$ is fed into the generator to obtain an image $\hat{x}$. The overall process is as follows:

$$\begin{align*}
    w &= f_w(s, \theta_w), \\
    \hat{x} &= f_g(w, \theta_g),
\end{align*}$$

where $\theta_w$ and $\theta_g$ denote the parameters of the Adaptive mapper $f_w(\cdot)$ and Generator $f_g(\cdot)$, respectively. After training, only $f_w(\cdot)$ is utilized to generate adaptive latent space in CAAE. Although our structure is different from the general GAN, the parameters of the two networks are still able to train in an end-to-end manner. Specifically, we aim to get the true distribution of normal class images by employing the adversarial training of discriminator and generator. Pixel adversarial learning is defined as follows:

$$l_{adv1} = E_{x \sim P_{data}}[\log D_{adv1}(x)] + E_{x \sim N(0,1)}[\log(1-D_{adv1}(f_g(f_w(s)))).$$

where $x$ is an image sample from the dataset, $s$ is a sample from the prior distribution $N(0, 1)$, $f_w(\cdot)$ and $f_g(\cdot)$ are used to generate images close to the real distribution. D aims to distinguish the generated image from real image. Lacking direct constraints on the pixels and features of the image, the whole network is time-consuming to train and the quality of the images is unsatisfied. In addition, many current generation tasks constrain images in feature space to improve performance, such as style transfer [40] and image inpainting [6]. In order to improve the image quality and training speed, we introduce the pixel and feature reconstruction loss [6], [40], which is defined as follows:

$$l_{rec} = \alpha_1 || x - \hat{x} || + \alpha_2 \sum_{i} || f_i(x) - f_i(\hat{x}) ||,$$

where $x$ is the input image, $\hat{x}$ is the output of overall network and $f_i(\cdot)$ is the $i$-th layer output of pre-trained VGG-19 [41] network. $\alpha_1$ and $\alpha_2$ are the weights of two reconstruction losses. Similar to other computer vision tasks [40], we define $\alpha_1$ and $\alpha_2$ as 1 and 0.1, respectively. Similar to the aforementioned work [6], the final loss function formula is balanced by the hyper-parameters $\lambda_{a1} = 0.1$ and $\lambda_{a2} = 10$ and is calculated as follows:

$$l_{alsg} = \lambda_{a1}l_{adv1} + \lambda_{a2}l_{rec}.$$
where $\lambda_{c1}$, $\lambda_{c2}$ and $\lambda_{c3}$ are the hyper-parameters in training and testing. For good reconstructions, a larger weight is assigned to pixel-feature reconstruction loss which is similar to OCGAN [13]. Other coefficients are chosen empirically based on the quality of reconstruction. In practice $\lambda_{c1} = 0.1$, $\lambda_{c2} = 1$ and $\lambda_{c3} = 10$ achieve desirable results in all our experiments.

D. GLOBAL AND LOCAL CHANNEL ATTENTION

1) GLOBAL CHANNEL ATTENTION

The essence of channel attention is to re-weight each feature map. Therefore, global channel attention (GCA) is to treat a feature map as a whole, and each point in the feature map is given the same weight. GCA is expressed as follows:

$$gca = \sigma(W_2 \cdot \alpha(W_1 \cdot GP(f)))$$

where $\alpha$ is ReLU function and $\sigma$ is Sigmoid function, $W_1$ and $W_2$ are the parameters of two fully connected layers and $GP(f) \in R^{H \times W \times C}$ which is obtained by global pooling feature maps. For simplicity, the combination of two FC layers and two activation functions is called calculating attention module. The global channel attention is obtained by the above formula, which is proposed in SENet [35].

2) LOCAL CHANNEL ATTENTION

Different from global channel attention (GCA), local channel attention (LCA) takes a small region in the feature map as a whole. Each small region has the same weight, and different regions have different weights. LCA is defined as follows:

$$lca = \sigma(W_2 \cdot \alpha(W_1 \cdot \rho(AP(f))))$$

where $\alpha$ is ReLU function and $\sigma$ is Sigmoid function, $W_1$ and $W_2$ are the parameters of two fully connected layers. In order to give different weights to each small region, the whole feature region is divided into $k \times k$ small regions by adaptive pooling (AP) operation. We define $AP(f)$ as $z$ for more precise expression and the combination of two FC layers and two activation functions is called calculating attention module. In order to make it suitable for fully connected layer, resize operation $\rho$ is proposed to convert $\rho(z) \in R^{k \times k \times C}$ to $\rho(z) \in R^{1 \times 1 \times k^2C}$. Finally, $lca \in R^{1 \times 1 \times k^2C}$ is obtained by two convolution layers and two activation functions.

3) ATTENTION FUSION MODULE

In this section, we will show the integration of global attention and local attention. Assume that the original feature maps $f \in R^{H \times W \times C}$. Firstly, global scaling operation extends $gca \in R^{1 \times 1 \times C}$ to $R^{H \times W \times C}$ by copying the values of each channel. Secondly, local scaling operation converts $lca \in R^{1 \times 1 \times k^2C}$ to $R^{H \times W \times C}$ by corresponding scaling in each region. Finally, the product of $gca$ and $lca$ is activated by sigmoid function. The above process is defined as follows:

$$ca = \sigma(\delta_1(gca) \odot \delta_2(lca))$$

where $gca$ and $lca$ represent two kinds of attention respectively and $\odot$ is the multiplication of corresponding positions. Moreover, $\delta_1$ is global scaling operation, $\delta_2$ is local scaling operation and $\sigma$ is sigmoid function. In addition, the above-mentioned three attention processes is depicted in Figure 2.

IV. EXPERIMENT

In this section, the proposed method is evaluated on three popular benchmarks available datasets, including MNIST [14], Fashion-MNIST [15], CIFAR-10 [16] datasets.

All the dataset published in novelty detection set a class from 0 to 9 as normal and the rest as novelty. All datasets are applied to novelty detection by setting a class from 0 to 9 as normal and the rest as novelty. Similar to the previous methods, we adopt two experimental protocols [13]. Finally, we present the ablation study of the proposed method.

A. PROTOCOL

1) PROTOCOL 1

The 10 classes take turns being regarded as normal class, while the other 9 classes are novel class. All the normal data in the dataset are randomly reshuffled, and the novel data are scrambled at the same time. Take 80% of normal data as training set and the rest as normal test set. In addition, the novel test set with the same size as the normal test set is randomly extracted from the novel data. A test set consists of a normal test set and an exception test set of equal size.

2) PROTOCOL 2

In this protocol, we use the training set or test set split of the original dataset. After processing, the training set contains
only normal class of the original training set, in other words, only 1/10 of the training data is used for training. For the test set, we first relabel all the data into two classes: normal class and new class. The normal class consists of the class corresponding to the training set, while the novel class contains all the remaining classes. For example, if 8 is a normal class, then the remaining classes from 0 to 9 form a novel class.

In order to better compare with previous works, Fashion-MNIST only experiments on protocol 1 and CIFAR-10 only experiments on protocol 2. Furthermore, MNIST adopts two protocols for experiments in this work.

### B. IMPLEMENTATION DETAILS

For convenience, we resize all images to $32 \times 32$ size, and then map all values to [0, 1] through standardization and normalization. The proposed method is optimizer by Adam [42] with a learning rate of 0.0002. The training progress stops when epochs are up to 100 or the loss of 10 epochs does not drop significantly and batch size is set to 64. All of the experiments are implemented on Intel® Core i5-8600K CPU@3.6GHz, 64GB, NVIDIA GeForce GTX 1080Ti 11GB GPU with Pytorch 1.1.0(python 3.6.0).

### C. EVALUATION METRICS

Due to the unreasonable distribution of the dataset, like most methods, we use receiver operating characteristic (ROC) and area under the curve (AUC) as the evaluation criteria. ROC curve is determined by true positive rate (TPR) and false positive rate (FPR):

\[
TPR = \frac{TP}{TP + FN}, \quad (12)
\]

\[
FPR = \frac{FP}{FP + TN}, \quad (13)
\]

where TP means the real label is true and the the prediction is positive, FN means the real label is true and the prediction is negative, FP means the real label is false and the prediction is positive, TN means the real label is false and the prediction is negative. AUC is defined as the area enclosed by the coordinate axis under the ROC curve. Its value is ranged from 0 to 1.

### D. BASELINE METHODS

- **DCAE [19]**: DCAE applies Denoising-Auto Encoder structure for novelty detection.
- **VAE [21]**: As a generation model comparable like GAN, VAE combines the advantages of Bayesian method and deep learning.
- **Pixel CNN [5]**: An image density generation model based on pixel CNN, which can generate images better. PixelCNN can be applied to AE or VAE as a powerful decoder.
- **GAN [8]**: Generator and Discriminator constitute a dynamic game process. This paper uses the same way to operate GAN as used in AnoGAN [8].
- **AnoGAN [8]**: AnoGAN is based on DCGAN structure, trying to find the mapping from image to latent space through gradient propagation during testing. The anomaly scores are weighted by reconstruction error and the output of discriminator.
- **ALOCC [10]**: ALOCC utilizes both reconstruction loss and adversarial loss to train the network. During inference, the output of the discriminator is adopted to represent the novelty scores.
- **GPND [32]**: The proposed method constrains latent space using a prior distribution. GPND computes the probability distribution as novelty scores in testing.
- **LSA [33]**: Similarly, LSA limits the expression ability of network through latent space autoregression. Novelty scores are carried out by summing the reconstruction loss and the log-likelihood loss.
- **OCGAN [13]**: OCGAN aims to improve the reconstruction loss of novelty samples by generating normal class images as much as possible. Therefore, they apply reconstruction loss as novelty scores.
- **IAE [25]**: IAE is based on Autoencoder structure, which is introduced channel attention and entropy minimization in novelty detection.
- **MBC [26]**: MBC is a model-based characterization of neural networks which characterizes novelty from the model perspective using gradients.
- **PAE [31]**: PAE utilizes a new AAE framework that is trained based on solving puzzles on randomly permuted image patches.

### TABLE 1. Novelty detection results for MNIST and Fashion-MNIST using Protocol 1.

| Method  | MINST  | Fashion-MNIST |
|---------|--------|---------------|
| DCAE    | 0.899  | 0.908         |
| ALOCC D | 0.820  | 0.601         |
| ALOCC DR| 0.880  | 0.753         |
| GPND    | 0.932  | 0.901         |
| OCGAN   | 0.977  | 0.924         |
| **Ours**| **0.984** | **0.945**     |

1) **MNIST**

MNIST contains 70000 digital samples from 0-9 with $28 \times 28$ resolution. Each pixel is represented by a gray value. We demonstrate the performance on MNIST using two protocols. For MNIST dataset, the proposed model AALS obtain state-of-the-art performance, which is compared with ALOCC, GPND, LSA, OCGAN, etc. AALS achieves the mean AUC of 0.984 and 0.9794 in two protocols. In protocol 2, the proposed method has achieved the state-of-the-art performance in six classes: 1, 2, 3, 6, 8, 9. It is also close to the most advanced performance in other classes. The proposed method is similar to OCGAN, and the generated novel class data are also close to known(normal) class as shown in Figure 3.
2) FASHION-MNIST
Fashion-MNIST dataset is a 28 × 28 pixels clothing image data, which is more complex than the scene of MNIST dataset. For Fashion-MNIST dataset, Table 1 lists the corresponding mean AUC values, which demonstrate the optimal performance of the proposed method. Compared with GPND and OCGAN, the DCAE and ALOCC methods performed poorly on this dataset. Moreover, the average AUC score of the proposed method is 0.945, which improves the latest method OCGAN by 0.021 and proves the effectiveness of the proposed method.

3) CIFAR-10
CIFAR-10 is a color image dataset that is closer to the regular object. There are 10 classes of 32 × 32 pixels RGB color images. Compared with MNIST dataset, CIFAR-10 dataset contains real world objects with a lot of noises and different characteristics. As a result, novelty detection results are comparatively weaker for all methods on CIFAR-10 dataset. Figure 3 shows our visual inspection results and Table 3 lists the corresponding mean AUC scores, such as VAE, Pixel CNN, GAN, AnoGAN, LSA, OCGAN and our method. Specifically, our proposal model outperforms other methods in AUC scores of seven classes and average AUC scores. Moreover, the difference between the remaining classes and the optimal performance is only about 2% AUC. In Figure 3 we visualize the normal images have a better reconstruction results, and the novel images tend to be close to the normal images. In summary, our method shows superior performance in complex datasets, which improves mean AUC scores 7.18% compared with the latest method OCGAN.

E. ABLATION STUDY
In this section, in order to prove the effectiveness of each components, we design a set of ablation experiments on CIFAR-10 dataset as shown in Table 4. Our first experiment considers only AAE structure which is the backbone structure of our method. Secondly, we use AAE with Latent Constraint(LC).
FIGURE 3. The graphic results of MNIST and CIFAR-10 datasets using AALS. The first 4 rows are denoted the normal samples and the rest is denoted the novel samples. The first 4 columns refer to the generated results of AALS and the rest refer to the ground truth.

TABLE 5. AALS architecture and parameters in novelty detection. The padding of all convolutional is 1 but † means padding = 0. All L-ReLU slope is 0.2 unless specified otherwise.

| Module       | Name                      | Operation | Kernel | Strides | Filters Units | Non Linearity | Batch Norm |
|--------------|---------------------------|-----------|--------|---------|---------------|---------------|------------|
| Encoder      | Conv2D+GLCA               | 4 x 4     | 2 x 2  | 64      | L-ReLU        | ✓             |            |
|              | Conv2D                    | 4 x 4     | 2 x 2  | 128     | L-ReLU        | ✓             |            |
|              | Conv2D                    | 4 x 4     | 2 x 2  | 256     | L-ReLU        | ✓             |            |
|              | DecConv2D†                | 4 x 4     | 1 x 1  | 100     | L-ReLU        | ❌             |            |
|              | DecConv2D†                | 4 x 4     | 2 x 2  | 128     | ReLU          | ✓             |            |
|              | DecConv2D†                | 4 x 4     | 2 x 2  | 64      | ReLU          | ✓             |            |
|              | GLCA+DecConv2D            | 4 x 4     | 2 x 2  | 3       | Tanh          | ×             |            |

CAAE

| Module       | Name                      | Operation | Kernel | Strides | Filters Units | Non Linearity | Batch Norm |
|--------------|---------------------------|-----------|--------|---------|---------------|---------------|------------|
| Global Pooling| FC                        | -         | -      | -       | -             | -             |            |
|              | PC                        | -         | -      | 64      | ReLU          | ×             |            |
|              | PC                        | -         | -      | 8       | Sigmoid       | ×             |            |
|              | Upsample                  | -         | -      | -       | -             | -             | ×          |
| Local Attention| FC                        | -         | -      | 1024    | ReLU          | ×             |            |
|              | PC                        | -         | -      | 128     | Sigmoid       | ×             |            |

Discriminator2

| Module       | Name                      | Operation | Kernel | Strides | Filters Units | Non Linearity | Batch Norm |
|--------------|---------------------------|-----------|--------|---------|---------------|---------------|------------|
| Conv2D       | Conv2D                    | 4 x 4     | 2 x 2  | 64      | L-ReLU        | ✓             |            |
|              | Conv2D                    | 4 x 4     | 2 x 2  | 128     | L-ReLU        | ✓             |            |
|              | Conv2D                    | 4 x 4     | 2 x 2  | 256     | L-ReLU        | ✓             |            |
| Latent Discriminator| Conv2D†                | 4 x 4     | 1 x 1  | 100     | L-ReLU        | ×             |            |
|              | PC                        | -         | -      | 100     | L-ReLU        | ×             |            |
|              | PC                        | -         | -      | 100     | L-ReLU        | ×             |            |
|              | PC                        | -         | -      | Sigmoid | ×             | ✓             |            |

Adaptive Mapper

| Module       | Name                      | Operation | Kernel | Strides | Filters Units | Non Linearity | Batch Norm |
|--------------|---------------------------|-----------|--------|---------|---------------|---------------|------------|
| Conv2D       | Conv2D                    | 4 x 4     | 2 x 2  | 64      | L-ReLU        | ✓             |            |
|              | Conv2D                    | 4 x 4     | 2 x 2  | 128     | L-ReLU        | ✓             |            |
|              | Conv2D                    | 4 x 4     | 2 x 2  | 256     | L-ReLU        | ✓             |            |

ALSG

| Module       | Name                      | Operation | Kernel | Strides | Filters Units | Non Linearity | Batch Norm |
|--------------|---------------------------|-----------|--------|---------|---------------|---------------|------------|
| Generator    | DeConv2D†                | 4 x 4     | 1 x 1  | 256     | ReLU          | ✓             |            |
| Discriminator1| Conv2D                    | 4 x 4     | 2 x 2  | 64      | L-ReLU        | ✓             |            |
|              | Conv2D                    | 4 x 4     | 2 x 2  | 128     | L-ReLU        | ✓             |            |
|              | Conv2D                    | 4 x 4     | 2 x 2  | 256     | L-ReLU        | ✓             |            |

Thirdly, we combine feature reconstruction loss (FL) in AAE. Thirdly, LC and FL is combined with AAE. To better test the role of GLCA, we next combined SENet [35] with our model. In the final scenario, we adopt full proposed model AALS.

Table 4 shows the mean AUC of all classes of CIFAR-10 dataset. Compared with the baseline method, each part has corresponding performance improvement. More concretely, the model without feature loss can’t get high-quality images, while the model without implicit space constraints
can’t distinguish normal and novel well which is depicted in Figure 4. Meanwhile, it proves that our model make a trade-off the expression ability of the network and the image quality. Moreover, GLCA also has obvious effect improvement compared with SENet [35]. Finally, we conduct experiments on the efficiency of the model in CIFAR-10 dataset. AALS consumes 2.9ms and 0.31ms (per image) in the training and testing stage respectively, which proves that our model has good real-time performance.

V. CONCLUSION

In this paper, we propose a general adaptive adversarial latent space (AALS) network for novelty detection. Specifically, the propose method consists of two core components: Adaptive Latent Space Generator (ALSG) and Constrained-based Adversarial Auto Encoder (CAAE). The ALSG is designed to generate adaptive latent space from a prior distribution. The CAAE utilizes adaptive latent space to constrain the latent space in AAE and adopt GLCA to obtain better feature maps. From the above analysis and experiments, we prove the effectiveness of GLCA. Meanwhile, we adapt pixel-latent adversarial loss and pixel-feature reconstruction loss for novelty detection. Moreover, pixel-feature loss is combined as adversarial loss and pixel-feature reconstruction loss for the effectiveness of GLCA. Meanwhile, we adapt pixel-latent adversarial loss and pixel-feature reconstruction loss for novelty detection. Moreover, pixel-feature loss is combined as adversarial loss and pixel-feature reconstruction loss for novelty detection. Furthermore, pixel-feature loss is combined as adversarial loss and pixel-feature reconstruction loss for novelty detection.

REFERENCES

[1] G. Hinton and R. Zemel, “Autoencoders, minimum description length and Helmholtz free energy,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), Denver, CO, USA, 1993, pp. 3–10.

[2] P. Vincent, H. Bengio, and P. Manzagol, “Extracting and composing robust features with denoising autoencoders,” in Proc. Int. Conf. Mach. Learn. (ICML), Helsinki, Finland, Jun. 2008, pp. 1096–1103.

[3] D. Kingma and M. Welling, “Auto-encoding variational Bayes,” in Proc. Int. Conf. Learn. Represent. (ICLR), Banff, AB, Canada, Apr. 2014, pp. 1–14.

[4] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, “Adversarial autoencoders,” 2015, arXiv:1511.05644. [Online]. Available: http://arxiv.org/abs/1511.05644

[5] A. Oord, N. Kalchbrenner, L. Espeholt, K. Kavukcuoglu, O. Vinyals, and A. Graves, “Conditional image generation with PixelCNN decoders,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), Barcelona, Spain, Dec. 2016, pp. 4790–4798.

[6] H. Liu, B. Jiang, Y. Xiao, and C. Yang, “Coherent semantic attention for image inpainting,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Seoul, South Korea, Oct. 2019, pp. 4169–4178.

[7] B. Jiang, W. Tu, C. Yang, and J. Yuan, “Context-integrated and feature-refined network for lightweight object parsing,” IEEE Trans. Image Process., vol. 29, pp. 5079–5093, 2020.

[8] T. Schlegl, P. Seebock, S. Waldstein, U. Schmidt-Erfurth, and G. Langs, “Unsupervised anomaly detection with generative adversarial networks to guide marker discovery,” in Proc. Inf. Process. Med. Imag. (IPMI), Boone, NC, USA, Jun. 2017, pp. 146–157.

[9] M. Ravanchkhsh, 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), Beijing, China, Sep. 2017, pp. 1577–1581.

[10] M. Sabokrou, M. Khaloei, M. Fathy, and E. Adeli, “Adversarially learned one-class classifier for novelty detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA, Jun. 2018, pp. 3379–3388.

[11] H.-G. Wang, X. Li, and T. Zhang, “Generative adversarial network based novelty detection using minimized reconstruction error,” Frontiers Inf. Technol. Electron. Eng., vol. 19, no. 1, pp. 116–125, Jan. 2018.

[12] H. Zenati, M. Romain, C.-S. Foo, B. Lecouat, and V. Chandrasekhar, “Adversarially learned anomaly detection,” in Proc. IEEE Int. Conf. Data Mining (ICDM), Singapore, Nov. 2018, pp. 727–736.

[13] P. Perera, R. Nathapati, and B. Xiang, “OCGAN: One-class novelty detection using GANs with constrained latent representations,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 2898–2906.

[14] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
B. Jiang et al.: AALS for Novelty Detection

IEEE Access

VOLUME 8, 2020

205097

[15] H. Xiao, K. Rasul, and R. Vollgraf, “Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms,” 2017, arXiv:1708.07747. [Online]. Available: http://arxiv.org/abs/1708.07747

[16] A. Krizhevsky and G. Hinton, “Learning multiple layers of features from tiny images,” Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, Tech. Rep., 2009.

[17] B. Scholkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, “Support vector method for novelty detection,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), Denver, CO, USA, Nov. 1999, pp. 582–588.

[18] H. Hoffmann, “Kernel PCA for novelty detection,” Pattern Recognit., vol. 40, no. 3, pp. 863–874, Mar. 2007.

[19] M. Sakurada and T. Yairi, “Anomaly detection using autoencoders with nonlinear dimensionality reduction,” in Proc. Mach. Learn. Sensor Data Anal. Workshops (MLSDAW), Gold Coast, QLD, Australia, Dec. 2014, p. 4.

[20] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe, “Learning deep representations of appearance and motion for anomalous event detection,” in Proc. Brit. Mach. Vis. Conf. (BMVC), Swansea, UK, Sep., 2015, pp. 8.1–8.12.

[21] Y. S. Chong and Y. H. Tay, “Abnormal event detection in videos using spatiotemporal autoencoder,” in Proc. Int. Symp. Neural Netw. (ISNN), Muroran, Japan, Jun. 2017, pp. 189–196.

[22] C. Zhou and R. C. Paffenroth, “Anomaly detection with robust deep autoencoders,” in Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Halifax, NS, Canada, Aug. 2017, pp. 665–674.

[23] B. Zong, Q. Song, M. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, “Deep autoencoding Gaussian mixture model for unsupervised anomaly detection,” in Proc. Int. Conf. Learn. Represent. (ICLR), Vancouver, BC, Canada, Apr. 2018, pp. 1–19.

[24] Z. Li, Y. Li, and Z. Gao, “Spatiotemporal representation learning for video anomaly detection,” IEEE Access, vol. 8, pp. 25531–25542, 2020.

[25] D. Guo, M. Tian, Y. Cui, X. Pan, and S. Chen, “Improving autoencoder novelty detection using channel attention and entropy minimization,” 2020, arXiv:2007.01682. [Online]. Available: http://arxiv.org/abs/2007.01682

[26] K. Woon, M. Prabhushankar, D. Temel, and G. AlRegib, “Novelty detection through model-based characterization of neural networks,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Abu Dhabi, United Arab Emirates, Oct. 2020, pp. 3179–3183.

[27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Proc. Advances in Neural Information Processing Systems (NIPS), 2014, pp. 2672–2680.

[28] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” in Proc. Int. Conf. Learn. Represent. (ICLR), San Juan, Puerto Rico, May 2016, pp. 1–16.

[29] Y. Li, Y. Cai, J. Liu, S. Lang, and X. Zhang, “Spatio-temporal unity networking for video anomaly detection,” IEEE Access, vol. 7, pp. 172425–172432, 2019.

[30] F. Dong, Y. Zhang, and X. Nie, “Dual discriminator generative adversarial network for video anomaly detection,” IEEE Access, vol. 8, pp. 88170–88176, 2020.

[31] M. Salehi, A. Eftekhar, N. Sadjadi, M. H. Rohban, and H. R. Rabiee, “Puzzle-AE: Novelty detection in images through solving puzzles,” 2020, arXiv:2008.12959. [Online]. Available: http://arxiv.org/abs/2008.12959

[32] S. Pilhorsky, R. Almohsen, and G. Doretto, “Generative probabilistic novelty detection with adversarial autoencoders,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), Montreal, QC, Canada, Dec. 2018, pp. 6823–6834.

[33] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, “Latent space autoregression for novelty detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 481–490.

[34] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. Van Den Hengel, “Memorynormality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Seoul, South Korea, Oct. 2019, pp. 1705–1714.

[35] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 7132–7141.

[36] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 3–19.

[37] G. Huang, J. Zhu, J. Li, Z. Wang, L. Cheng, L. Liu, H. Li, and J. Zhou, “Channel-attention U-Net: Channel attention mechanism for semantic segmentation of esophagus and esophageal cancer,” IEEE Access, vol. 8, pp. 122798–122810, 2020.

[38] H. Lee and S. Cho, “Locally adaptive channel attention-based network for denoising images,” IEEE Access, vol. 8, pp. 34686–34695, 2020.

[39] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 4401–4410.

[40] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Amsterdam, The Netherlands, Oct. 2016, pp. 694–711.

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

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

FANGQIANG XU received the bachelor’s degree in computer science and technology from the Qingdao University of Technology, in 2019. He is currently pursuing the master’s degree with the College of Computer Science and Electronic Engineering, Hunan University, China. His research interests include novelty detection, deep learning, and machine learning.

YUN HUANG received the bachelor’s degree in computer science and technology from Northeast Forestry University, in 2019. She is currently pursuing the master’s degree with the College of Computer Science and Electronic Engineering, Hunan University, China. Her research interests include text2image, deep learning, and machine learning.

B. Jiang received the B.S., B.E., and M.E. degrees from Hunan University (HNU), China, and the Ph.D. degree from the Tokyo Institute of Technology (Tokyo Tech), Japan. From 2002 to 2003, he was with KISTEM Company Ltd. (Key Information System, Gate-Group), Japan, as a Trainee. His research interests include big data technology, artificial intelligence, machine vision, machine learning, data mining, intelligent computing, recommender systems, and social computing.
CHAO YANG received the B.E. and M.E. degrees from Hunan University (HNU), China, and the Ph.D. degree from the Tokyo Institute of Technology (Tokyo Tech), Japan. Her research interests include intelligent recommendation systems, multimedia content understanding and information retrieval, and social network computing. She is a member of the Collaborative Computing Committee, Chinese Computer Society, and a member of the Computer Application Committee.

WEI HUANG received the bachelor’s degree in computer science and technology from the Xi’an University of Technology, in 2019. He is currently pursuing the master’s degree with the College of Computer Science and Electronic Engineering, Hunan University, China. His research interests include image inpainting, deep learning, and machine learning.

JUN XIA received the bachelor’s degree in computer science and technology from the Qingdao University of Technology, in 2019. She is currently pursuing the master’s degree with the College of Computer Science and Electronic Engineering, Hunan University, China. Her research interests include image colorization and deep learning.

* * *