Received: 20 January 2024 | Revised: 15 May 2024 | Accepted: 14 June 2024

IET Computer Vision

DOI: 10.1049/cvi2.12297

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

DualAD: Dual adversarial network for image anomaly detection★

Yonghao Wan | Aimin Feng

Nanjing University of Aeronautics and Astronautics, Nanjing, China

Correspondence
Aimin Feng.
Email: amfeng@nuaa.edu.cn

Abstract

Anomaly Detection, also known as outlier detection, is critical in domains such as network security, intrusion detection, and fraud detection. One popular approach to anomaly detection is using autoencoders, which are trained to reconstruct input by minimising reconstruction error with the neural network. However, these methods usually suffer from the trade-off between normal reconstruction fidelity and abnormal reconstruction distinguishability, which damages the performance. The authors find that the above trade-off can be better mitigated by imposing constraints on the latent space of images. To this end, the authors propose a new Dual Adversarial Network (DualAD) that consists of a Feature Constraint (FC) module and a reconstruction module. The method incorporates the FC module during the reconstruction training process to impose constraints on the latent space of images, thereby yielding feature representations more conducive to anomaly detection. Additionally, the authors employ dual adversarial learning to model the distribution of normal data. On the one hand, adversarial learning was implemented during the reconstruction process to obtain higher-quality reconstruction samples, thereby preventing the effects of blurred image reconstructions on model performance. On the other hand, the authors utilise adversarial training of the FC module and the reconstruction module to achieve superior feature representation, making anomalies more distinguishable at the feature level. During the inference phase, the authors perform anomaly detection simultaneously in the pixel and latent spaces to identify abnormal patterns more comprehensively. Experiments on three data sets CIFAR10, MNIST, and FashionMNIST demonstrate the validity of the authors' work. Results show that constraints on the latent space and adversarial learning can improve detection performance.

KEYWORDS

computer vision, feature extraction, image recognition, image reconstruction, vision defects

1 | INTRODUCTION

Anomaly detection refers to identifying patterns that do not conform to expected normal behaviour [1]. It is a critical subject with broad applications across various industries, playing a key role in diverse domains such as network security [2], intrusion detection [3] and fraud detection [4]. The task of anomaly detection is challenging due to the rarity of abnormal samples and deviations from normality being continuous and sporadic by nature.

So far, many visual anomaly detection methods are based on image reconstruction [5, 6]. These methods assume that the model trained on normal samples can only well reconstruct normal regions but fail in abnormal regions, and therefore the reconstruction errors can be utilised for identifying anomalies. With the advent of generative adversarial networks (GANs) [7], many works have been done to combine autoencoders with generative adversarial networks to mitigate the impact of low-quality image reconstructions on model performance [8–12]. However, in practice, the above assumption may not always be

valid since it is difficult to find a definitive generalisation boundary when no real defective sample is served as supervision. In this case, suppressing the model's generalisation to anomalies usually requires more compact latent representations.

Fortunately, there are many methods available to obtain more compact representations, such as One-Class Classification (OCC). The OCC has also received considerable attention in the field of anomaly detection, where the behaviour of the known class in the latent space is modelled. Earlier work of this nature used one-class modelling tools such as One-class Support Vector Machine (SVM) [13] and Support Vector Data Descriptor (SVDD) [14] to analyse data based on representations derived from latent features. SVDD is a classic one-class classification algorithm derived from the SVM. It maps all normal training data into a kernel space and seeks the smallest hypersphere that encloses the data in the space. Anomalies are expected to be located outside the learnt hypersphere. Deep SVDD [15]first incorporated this idea in a deep neural network for image anomaly detection. They used a neural network to mimic the kernel function and trained it with the radius of the hypersphere. This modification allows an encoder to learn a data-dependent transformation, thus enhancing detection performance on high-dimensional and structured data. Although DeepSVDD is a promising OCC algorithm and can effectively mitigate the model's generalisation to anomalies, it suffers from a vital problem during training called "hypersphere collapse", which means the network converges to the trivial solution of all-zero weights.

To remedy the above issues, we propose a novel anomaly detection algorithm, named Dual Adversarial Network (DualAD). It incorporates the Feature Constraint (FC) module (section 3.2, also known as the SVDD module) into the autoencoder model to impose constraints on the latent space of images, which can alleviate the model's generalisation to anomalies. Furthermore, due to the influence of the reconstruction loss, training the FC module can avoid the hypersphere collapse. With the powerful capability of GANs to generate realistic images, we employ an adversarial autoencoder (AAE) [16] as our reconstruction model to prevent the impact of low-quality reconstructions on model performance. The proposed work is shown in Figure 1. We utilise a dual adversarial network to model the normal distribution. In addition to the adversarial learning within the reconstruction module, we cleverly employ adversarial learning between the reconstruction module and the FC module to constrain the latent space of the reconstruction module more accurately and enhance the FC module's ability to detect anomalies in the latent space. Specifically, we treat the hypersphere centre c as a trainable parameter and iteratively train it with the AAE network so that we can achieve a more optimised, constrained latent space. During the inference phase, we perform anomaly detection simultaneously in both the pixel and latent spaces to identify abnormal patterns more comprehensively.

In summary, our major contributions are as follows:

1. We have proposed a new image anomaly detection framework named DualAD, which is composed of a reconstruction module and an FC module. The FC module can be substituted with other one-class classification algorithms or clustering algorithms.
2. We use dual adversarial networks to model normal data distribution. The intra-modular adversarial learning in the reconstruction module enables high-quality reconstruction of samples. The inter-modular adversarial learning between the reconstruction module and the FC module allows us to obtain a latent space that is more favourable for anomaly detection tasks.
3. Our method conducts anomaly detection both in the latent space and pixel space, enhancing the model's ability to recognise anomalies. Experiments on different datasets prove the effectiveness of our proposed method.

2 | RELATED WORKS

The existing anomaly detection methods can be divided into four categories: One-class Classification-based method, Clustering-based method, End-to-end anomaly score learning-based method, and Reconstruction-based method.

2.1 | One-class classification-based method

Before deep learning became popular, probably the most prominent method of One-Class classification (OCC) was a kernel-based method named One-Class SVM (OC-SVM) [13] and SVDD. The key idea is to learn a hyperplane by a maximum margin for OCSVM and a minimal hypersphere for SVDD in the feature space. Another kernel-based One-Class classification method is the SVDD [17]. Subsequently, DeepSVDD leverages neural networks to map data instances into the hypersphere of minimum volume and determines whether the sample is abnormal based on the relative position of the test sample's features and the hypersphere centre. After DeepSVDD, more new algorithms are proposed for image Anomaly Detection. PatchSVDD [18] divides the image into uniform patches and sends them to the model for training, which significantly enhances the model's ability to detect anomalies. DSPSVDD [19] designs an improved comprehensive optimisation objective for the deep SVDD model that considers minimising hypersphere and network reconstruction error simultaneously to extract deep data features more effectively. SE-SVDD [20] proposes a Semantic Correlation Block to improve the representation of abnormal semantics and the accuracy of anomaly localisation by extracting multi-level features. MOCCA [21] also employs the same multi-layer features for anomaly detection but uses an autoencoder to extract features and locate the boundary position of normal

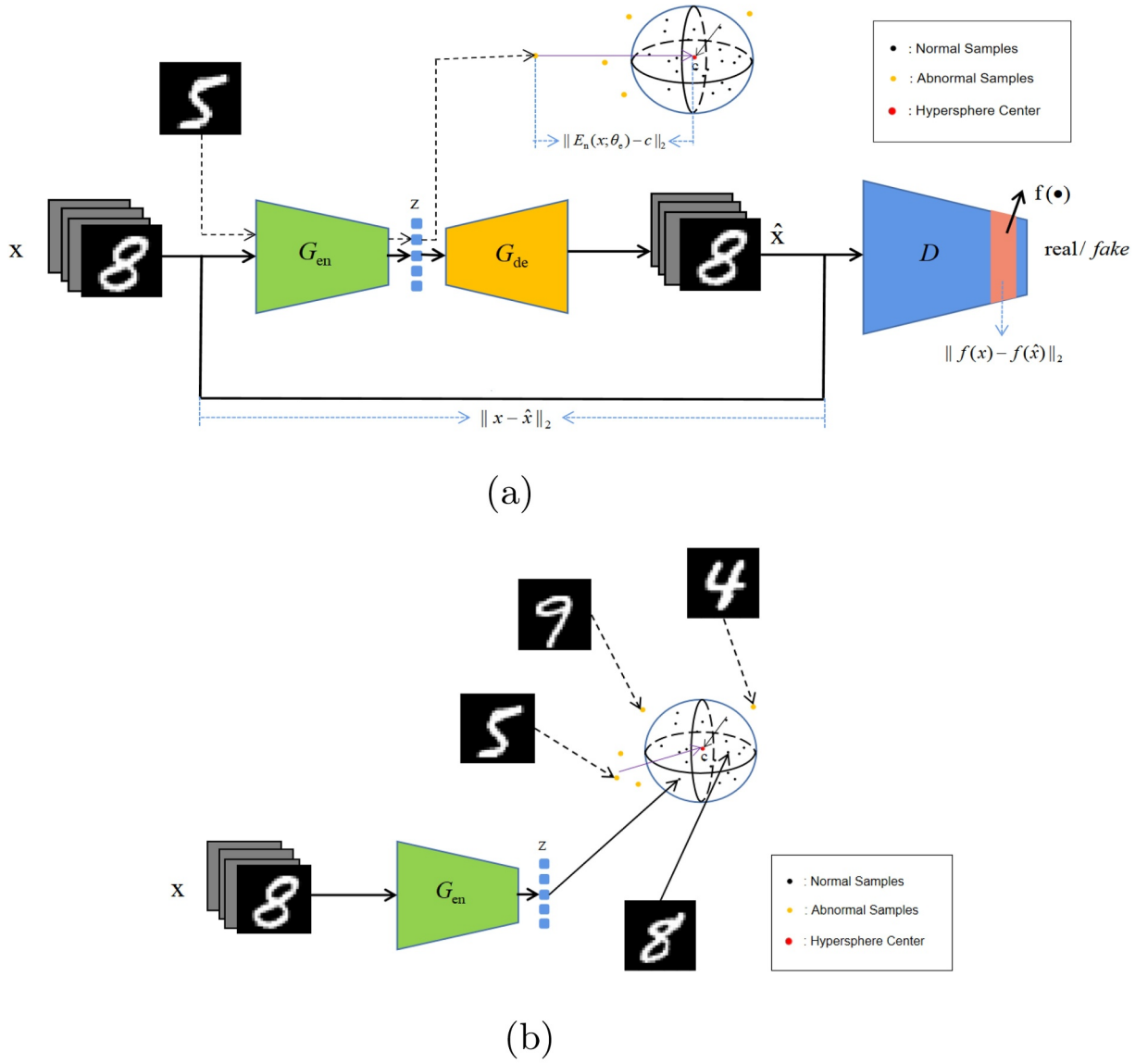


FIGURE 1 (a) Overview of the proposed DualAD. In the training phase, nominal samples are fed into the encoder G_{en} to get nominal features. Then, the decoder G_{de} is utilised to reconstruct the inputs. The discriminator D is used to adversarially train with the generator to improve the quality of reconstructed samples. The figure above uses '8' in the MNIST dataset as the target class. The black dashed line represents the behaviour of outlier samples in the FC module. (b) The schematic diagram of the FC module shows the ability of the FC module to distinguish between normal and abnormal samples after training.

features at each layer. And Sauter et al. [22] attempted to use the Xception network for classification and obtained results comparable to SVDD. Hadi Hojjati [23] combines the autoencoder and SVDD to make the model more robust, and the hypersphere is difficult to collapse. FCDD [24] employs a fully convolutional neural network for OCC. Since the relative positions of the features for each image layer do not change during the convolution process, FCDD yields more interpretable results than alternative methods. BQP [25] employs convolutional neural networks for feature extraction and utilises the BQP layer to calculate and iteratively optimise the centre and radius of the hypersphere for each batch. Furthermore, it employs a maximum entropy loss function to balance the uncertainty of boundary samples. Mathematically, this approach represents an optimal solution, avoiding trivial solutions caused by the infinite compression of the hypersphere. Subsequently, more works are all based on One-Class Classification [26–29].

2.2 | Clustering-based method

Clustering is an unsupervised learning method used to organise similar data points into distinct groups or clusters. It assigns each data point to a cluster by evaluating their similarity or distance. Since a strong interconnection between clustering and

anomaly detection, it leads to significant research efforts focused on utilising clustering outcomes to identify anomalies. The underlying assumption of clustering-based anomaly detection methods is that normal instances adhere more strongly to clusters than anomalies, and the effectiveness of anomaly detection heavily relies on the clustering results [17]. Various approaches have been explored, including examining cluster size [30], measuring distances to cluster centres [31], and assessing cluster membership [32]. This category also includes Gaussian mixture model-based anomaly detection [33], which inherently relates to clustering methods. It is worth mentioning that the integration of clustering and deep learning for anomaly detection [34] has also appeared in recent years.

2.3 | End-to-end anomaly score learning-based method

This approach aims to develop an end-to-end one-class classifier that can effectively determine whether a given instance is normal or not. The key idea is to learn a classifier that can distinguish normal instances and given pseudo-exceptions, such as binary classification. ALOCC [35] aims to create and train two deep networks, where one network serves as the one-class model responsible for distinguishing normal samples from anomalies, while the other is trained to enhance normal and generate distorted outliers. The instantiation and optimisation of these two networks are achieved using the GANs. The generator of ALOCC can be regarded as a denoising autoencoder [36], which is the core part of the model. Fence GAN [37] is also used to learn an end-to-end classifier, and the key idea of the method is to generate data samples that closely align with the boundary of the training data distribution. To accomplish this, two loss functions are incorporated into the generator. These loss functions ensure that the generated instances are evenly distributed along a hypersphere boundary defined by the training data. SimpleNet [38]generates synthetic anomalies in a pre-trained feature space to train a discriminator network for detecting anomalous features.

2.4 | Reconstruction-based method

All of these approaches are based on one assumption: anomalies are unable to be effectively reconstructed from low-dimensional projections due to their incompressible nature. Early methods used autoencoders to reconstruct inputs [27, 39]. However, this approach often resulted in low-quality reconstructions, which led to a decline in model performance. With the rise of GANs, some methods have begun to apply adversarial learning within reconstruction networks to enhance the performance of the model [8, 9, 11]. The latter enhances the reconstruction quality and overall model performance by incorporating adversarial training based on the foundational approach. In addition, introducing regularisation terms into reconstruction-based anomaly detection models often results in larger reconstruction errors for anomalies. For example, works such as the Variational Autoencoder (VAE) [40] and the Denoising Autoencoder [36] have been utilised to enhance the performance of reconstruction-based anomaly detection. DRAEM [41] exemplifies reconstruction-based techniques. It synthesises abnormal images and reconstructs them as normal by utilising external datasets, which significantly enhances the generalisation capacity of the reconstruction network. Additionally, the model feeds both the original and reconstructed images into a segmentation network to identify abnormal regions, thereby greatly improving its segmentation capabilities. However, it is prone to failures when synthesising anomalies that are near the distribution. Inspired by saliency detection, Xing et al. [42] proposed the Saliency Augmentation Module to generate more realistic abnormal images than DRAEM, achieving better results. DSR [43] proposes an architecture based on quantised feature space representation and dual decoders to circumvent the requirement for image-level anomaly generation. By sampling the learnt quantised feature space at the feature level, the near-in-distribution anomalies are generated in a controlled way. NSA [44] does not use external data for augmentation but adopts more data augmentation methods, allowing it to outperform all previous methods learnt without utilising additional datasets. In contrast to other works that attempt to reconstruct abnormal images into normal images, Bauer [45] proposes reconstructing the abnormal areas of the image so that they deviate from the original image's appearance. This approach produces comparable results to others.

3 | PROPOSED METHOD: DualAD

The proposed DualAD consists of two modules: a reconstruction module and a FC module. The reconstruction module is implemented using the AAE, consisting of three components: an encoder G_{en} (encoding image as a latent vector z for training the FC module), a decoder G_{de} (reconstruction of image), and a discriminator D (adversarial training autoencoder); G_{en} and G_{de} are two parts of the generator in the AAE. As shown in Figure 1a, the image x is entered into the encoder to obtain the latent vector z, the decoder reconstructs z into the image $\hat{x}$, and adversarial learning is used to obtain high-quality reconstruction samples. We incorporate adversarial loss during the training of the reconstruction module to enhance the quality of reconstructed images and use feature matching loss to reduce the instability of adversarial training. Additionally, to suppress the generalisation of anomalies by the reconstruction model, we add a feature constraint module to impose constraints on the latent space after image encoding. The FC module (as shown in Figure 1b) is implemented using OCC (One-Class Classification), such as SVDD. The motivation for this module is to achieve a latent space where each instance from the latent space represents an image from the normal class. In the training phase, the reconstruction module and the FC module are trained iteratively through adversarial learning. In DualAD, dual adversarial refers to the intra-modular adversarial learning within the

reconstruction module and the inter-modular adversarial learning between the two modules.

Compared to anomaly detection methods based on OCC-based methods, we have incorporated a reconstruction loss to prevent hypersphere collapse and employed a trainable hypersphere centre. Unlike purely reconstruction-based methods, we constrain the latent space of normal samples using the FC module and enhance the quality of reconstructed samples and the ability to model the normal distribution through adversarial training. In the inference phase, our method can detect anomalies in both feature and pixel spaces simultaneously, ensuring no anomalies are missed. Furthermore, the method presented in this paper is entirely unsupervised, not utilising any anomalous samples during the training phase.

3.1 | Adversarial training

Generative adversarial network (GAN) is a type of machine-learning model consisting of a generator and a discriminator. It aims to generate realistic samples by having the generator compete against the discriminator. This training process leads to the generation of increasingly realistic outputs. GANs are powerful tools for creating high-quality and diverse samples through this adversarial learning approach. The idea of GAN is the two-player minimax game. It's optimization goals are as follows:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log(D(x))] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \tag{1}$$

Where G represents the generator, D represents the discriminator, x represents the real sample, z represents the random noise, $p_{\text{data}}(x)$ represents the distribution of the real sample and $p_z(z)$ represents the distribution of noise.

GANs are very powerful in generating realistic images, so we introduce the generative adversarial idea of GANs into our model. The loss for adversarial learning is as follows:

$$\mathcal{L}_{adv} = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log(D(x))] + [\log(1-D(G_{de}(G_{en}(x))))] \tag{2}$$

By integrating adversarial training into the reconstruction process, we can significantly reduce fuzzy artefacts in the reconstructed images. This approach not only addresses the issue of large reconstruction errors for normal samples but also increases the distinguishability between normal and abnormal samples. However, adversarial training exhibits instability during the training process. Moreover, while adversarial loss $\mathcal{L}_{adv}$ ensures that the original and reconstructed images belong to the same distribution, it does not guarantee the high similarity between these images. This limitation stems from significant potential errors due to background variations within the same distribution. According to [46], feature matching loss has been shown to reduce the instability of GAN training. Therefore, feature matching loss is employed to not only align the images to the same distribution but also enhance their similarity, aligning closely with the goals of our reconstruction module. Unlike direct pixel-level comparisons, feature matching loss operates at a higher abstraction level, capturing the intrinsic structure and complexity of the samples, and reducing the impact of background variations or blurriness in the reconstructed images. This approach helps to prevent adverse effects on detection outcomes, thereby complementing the feature matching loss introduced in our model. f is a feature extraction function that can obtain a feature representation of the input sample. Specifically, $f(x)$ represents the feature of sample x extracted by the last fully connected layer ($f(.)$) of the discriminator. $f(.)$ is designed to enable feature matching between the original sample and the reconstructed sample. Our feature matching loss $\mathcal{L}_{fm}$ is defined as follows:

$$\mathcal{L}_{fm} = \mathbb{E}_{x\sim p_{\text{data}}(x)}\|f(x) - f(G_{de}(G_{en}(x)))\|_2 \tag{3}$$

In the reconstruction module, feature matching loss is $\mathcal{L}_{fm}$, and the reconstruction loss is as follows:

$$\mathcal{L}_{rec} = \mathbb{E}_{x\sim p_{\text{data}}(x)}\|x - G_{de}(G_{en}(x))\|_2 \tag{4}$$

The loss function of the generator is as follows:

$$\mathcal{L}_G = \lambda_1 * \mathcal{L}_{adv} + \lambda_2 * L_{fm} + \lambda_3 * \mathcal{L}_{rec} \tag{5}$$

where λ_1, λ_2, and λ_3 are the parameters of each loss item in the balance reconstruction module, and the larger the value is, the more the model pays attention to a certain item. The generator G composed of the encoder and decoder is trained against the discriminator D. The real sample is the input of the encoder G_{en}, and the generated sample is the output of the decoder G_{de}. The loss function of the discriminator is as follows:

$$\mathcal{L}_D = -\mathbb{E}_{x\sim p_{\text{data}}(x)}(\log(D(x)) + \log(1-D(G_{de}(G_{en}(x))))) \tag{6}$$

Through the adversarial game between the generator and discriminator, the quality of the reconstructed samples produced by the model continuously improves, and the reconstruction error of the normal samples becomes smaller.

3.2 | Feature constraint module

After adding adversarial learning, the reconstruction module has been able to reconstruct a more realistic normal sample. However, the reconstruction module has difficulty in finding a definitive generalisation boundary when there are no actual defective samples available for supervision. Therefore, as shown in Figure 1, we added the feature constraint module to the model. The underlying idea of the FC module stems from

SVDD's effectiveness in delineating a boundary around a dataset's normal class. It aims to leverage this approach by constructing a latent space where each data instance explicitly represents a typical image of the normal class. By constraining the latent space through the FC module, the reconstruction module can emphasise the representation of normal data characteristics and enhance the distinction between normal and anomalous instances. During inference, the distance of a sample representation from the centre of the hypersphere can serve as a criterion for identifying anomalies. Furthermore, the FC module can effectively prevent the degradation of model performance, which is often caused by mode collapse during the training process of *GAN*. By stabilising the training dynamics, the FC module ensures that the diversity of the generated samples is maintained, reducing the likelihood of the generator producing overly homogeneous outputs. This is crucial for enhancing the robustness and reliability of the GAN in various application scenarios. Define the encoding and decoding functions of the AAE as G_{en} and G_{de} respectively, and denote θ_e and θ_d as the parameters (weights and biases) associated with these functions.

For input x, the latent representation z of the encoder output is as follows:

$$z = G_{en}(x;\theta_e) \tag{7}$$

The hypersphere centre c can be initialised by averaging over the latent representations of the normal samples:

$$c = \frac{1}{|B|}\sum_{i=1}^{|B|} G_{en}(x;\theta_e) \tag{8}$$

Where $|B|$ is the number of samples in a batch. The optimisation goal of the FC module is as follows:

$$\min_{c}\mathcal{L}_{fc} = \frac{1}{m}\sum_{i=1}^{m}\|G_{en}(x_i;\theta_e) - c\|^2 \tag{9}$$

Where m is the number of data samples required to train the FC module in a batch, and x_i denotes one of the samples within these m samples. The optimisation goal of Equation (9) is to adjust the centre c of the hypersphere in such a way that it aggregates all feature representations of the normal samples within the hypersphere as closely as possible. During testing, it is commonly assumed that the proximity of a sample to the hypersphere's centre indicates a higher likelihood of it being normal. On the contrary, the greater the distance from the centre, the higher the degree of anomaly associated with the sample. Within the DualAD algorithm, to avoid the risk of hypersphere collapse, it is critical not to train the encoder in the reconstruction module and the hypersphere centre c in the FC module simultaneously. Therefore, we recommend separate training for these two modules. Section 3.3 provides more details about this training strategy.

3.3 | Optimisation objective

With the addition of the feature constraint module, we need to optimise the hypersphere centre c, the autoencoder network (G_{en} and G_{de}), and the discriminator network(D). In the training phase, we can train the adversarial-autoencoder(AAE) first and then the feature constraint module, or vice versa. Since the latent representation generated by an untrained autoencoder could be random and meaningless vectors that would not make sense for training the FC module, we chose to train the AAE in a batch by first selecting part of the sample and then training the FC module with the remaining sample. We set the sample scale for the first half to η, and the second half as $1-\eta(\eta > 0, \eta < 1)$, which is an adjustable hyperparameter. This training procedure is summarised in Algorithm 1.

The proposed model is trained using unlabelled samples, which is based on the assumption that there are many types of anomalies and it is difficult to collect abnormal samples, so the anomaly detection problem cannot be solved as a binary classification problem.

Assume that the total number of samples in a batch is $|B|$, and the total number of samples trained in the first part (reconstruction module) is $n = \eta\ |B|$. The overall optimisation goals for the generator G in the training phase are as follows:

$$\begin{aligned}\min_{\theta_e,\theta_d}\mathcal{L}_{G_{total}} = \lambda_1 * \mathcal{L}_{rec} + \lambda_2 * L_{adv} + \lambda_3 * \mathcal{L}_{fm} \\ + \alpha\cdot\|G_{en}(x;\theta_e) - c\|^2\end{aligned} \tag{10}$$

Where α is the parameter of loss items between the balance reconstruction module and the feature constraint module; the larger the value, the more attention the model pays to the feature constraint module. λ_1, λ_2 and λ_3 and α are adjustable hyperparameters and $\lambda_1+\lambda_2+\lambda_3 = 1$. The optimisation goals for the discriminator are defined as follows:

$$\min_{\theta_D}\mathcal{L}_D = -\mathbb{E}_{x\sim p_{\text{data}}(x)}(\log(D(x)) + \log(1 - D(G_{de}(G_{en}(x))))) \tag{11}$$

Where θ_D denotes the parameters of the discriminator network. In Equation (10) and Equation (11), x represents one of the n samples. The optimisation goal for the FC module is defined in Equation (9). In this paper, we adopt an adversarial training strategy to train the discriminator and generator. Additionally, the reconstruction module and the FC module are also optimised through adversarial training methods.

3.4 | Anomaly score

To prevent the omission of anomalies, the model enhances its detection capabilities and accuracy by simultaneously identifying anomalies in both pixel and feature spaces during the inference phase. This dual-detection mechanism ensures that

potential anomalies are captured from various dimensions, significantly reducing the risk of missed detections. Therefore, given the test sample x, the anomaly score A (x) obtained by the model is defined as follows:

$$A(x) = \|x - G_{de}(G_{en}(x;\theta_e),\theta_d)\|^2 + \varepsilon\|G_{en}(x;\theta_e) - c_f\|^2 \tag{12}$$

In Equation (12), θ_e, θ_d, and c_f are all the optimal solutions obtained after training. To assess the overall anomaly performance, the anomaly scores of the individual test sample x in the test set D_{test} are computed, resulting in a set of anomaly scores $A = \{A(x_i), x_i \in D_{test}\}$. Since the outlier scores can be too large or too small to compare, we scale them to the probabilistic range of [0,1]:

$$A_{standard}(x_i) = \frac{A(x_i) - \min(A)}{\max(A) - \min(A)} \tag{13}$$

Where $A_{standard}(x_i)$ is a standardised outlier score that can be used to evaluate the test dataset.

4 | EXPERIMENTS

4.1 | Implementation details

The DualNet is implemented in the Pytorch framework and Python ran on a 64-bit Ubuntu 20.04 LTS system with an Xeon(R) Platinum 8358P processor, 24 GB RAM and NVIDIA RTX 3090 GPU. We suggest to use stochastic gradient descent or its variants such as Adam [47] as the optimiser for the autoencoder's parameters. For training the hyper-sphere centre c, we recommend using an algorithm with an adaptive learning rate, such as AdaGrad [48]. This usually results in a faster convergence of the training procedure since it allows to assign higher weights to tuning c in the first few training epochs. The learning rate of G_{en}, G_{de} and D is the same and equal to 0.00001, and Beta1 and Beta2 use the default values. In the reconstruction module, we focus on the reconstruction quality of the image, and the model should pay more attention to the adversarial loss: $\lambda_1 = 0.05$, $\lambda_2 = 0.90$, $\lambda_3 = 0.05$ (confirmed through experimentation). We compare the performance of our method using Area Under the Curve (AUC) of the Receiver Operating Characteristics curve.

During the experiment, 10 primary baselines were used to compare directly with RACA. The traditional machine learning methods: OCSVM [13]. The classic anomaly detection: DAE [6] andVAE [40]. Some existing deep learning methods: AnoGAN [9], OCGAN [8], DSVDD [15], and DASVDD [23].

4.2 | Datasets

We evaluated the performance of our model on three datasets, which are common benchmark datasets for anomaly detection tasks: **MNIST** dataset [49], **FashionMNIST** dataset [50] and **CIFAR10** dataset [51]. Here, we briefly overview the three datasets in Figure 2.

Algorithm 1. The training process of DualAD.

Input: Set of training data x,iteration size N, Batch size $|B|$, parameters: λ_1,λ_2, λ_3,η, α.
Output: parameters: θ_e,θ_d,θ_D,c_f.

initialization:
Randomly initialize θ_e,θ_d,θ_D
$c = \frac{1}{(1-\eta)\cdot|B|}\sum_{i=1}^{(1-\eta)\cdot|B|} E_n(x;\theta_e)$

for *iteration 1* ***to*** $\rightarrow N$ **do**
  Take the number of samples $\eta \cdot |B|$ to train the adversarial autoencoder.
  Fixed hypersphere center c.
  Discriminator update:
  $z = G_{en}(x)$
  $\hat{x} = G_{de}(z)$
  update θ_D by Eq. (11)
  Generator update:
  update θ_e, θ_d by Eq. (10)
  Pick the number of samples (1-η) $\cdot$ $|B|$ (the remaining samples) to train the feature constraint module.
  Fixed the parameters of generator G_{en}, G_{de} and discriminator D.
  update hypersphere center c by Eq. (9)
end
$c_f = c$
c_f is the optimal value of c.

4.3 | Results and analysis

The results of the DualNet model compared with other benchmark models are shown in Table 1 and Figure 3a. Figure 3b shows t-SNE visualisation of the features extracted from G_{en}, which indicates that our method can make the abnormal and normal samples distinguishable in the latent space. In Table 1, each case highlights the algorithm with the best performance, indicated in bold. The average performance of the DualNet model is higher on all datasets than on other benchmark datasets.

On average, the DualNet and other benchmark models perform significantly worse on the CIFAR10 dataset than on the MNIST and FashionMNIST datasets. This underperformance

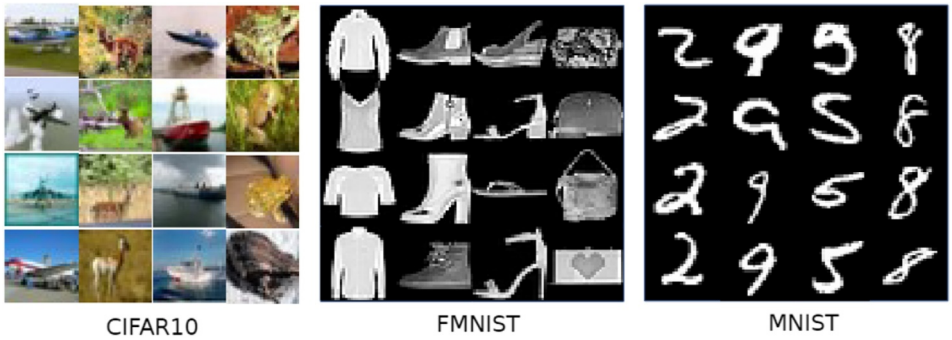


FIGURE 2 Representative images from the datasets used for evaluation. Images in each column belong to the same class.

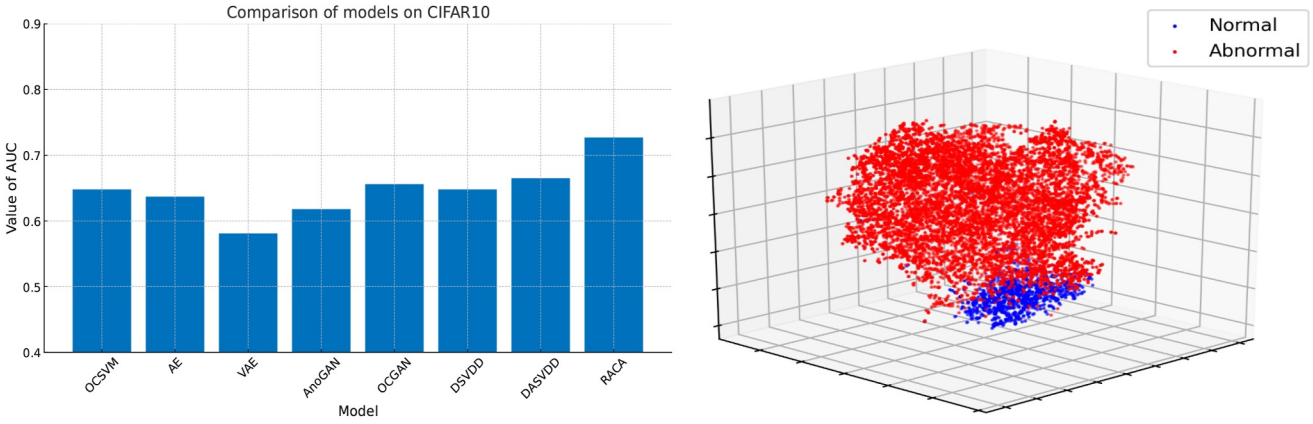
TABLE 1 Comparison of DualNet with previous works on MNIST and FashionMNIST(%).

Dataset	OCSVM	AE	VAE	AnoGAN	OCGAN	DSVDD	DASVDD	Ours
0	98.6	98.8	99.7	96.6	**99.9**	98.0	99.6	98.9
1	99.5	99.3	99.9	99.2	99.9	99.7	99.9	**99.9**
2	82.5	91.7	93.6	85.0	94.2	91.7	95.4	**97.7**
3	88.1	88.5	95.9	88.7	96.3	91.9	96.2	**96.9**
4	94.9	86.2	97.3	89.4	97.5	94.9	98.1	**98.3**
5	77.1	85.8	96.4	88.3	**98.0**	88.5	97.2	97.1
6	96.5	95.4	99.3	94.7	99.1	98.3	**99.6**	99.5
7	93.7	94.0	97.6	93.5	98.1	94.6	98.1	**98.1**
8	88.9	82.3	92.3	84.9	93.9	93.9	94.2	**96.0**
9	93.1	96.5	97.6	92.4	98.1	96.5	98.3	**98.9**
Mean	91.3	91.9	96.9	91.3	97.5	94.8	97.7	**98.1**
TShirt	90.6	71.6	87.4	89.0	91.1	79.1	91.2	**91.6**
Trouser	97.5	96.9	97.7	97.1	92.8	94.0	99.0	**99.1**
Pullover	88.1	72.9	81.6	86.5	88.0	83.0	89.3	**89.6**
Dress	91.3	78.5	91.2	91.2	87.6	82.9	**93.7**	93.4
Coat	88.5	82.9	87.2	87.6	82.3	87.0	90.7	**91.4**
Sandal	87.6	93.1	91.6	89.6	91.2	80.3	**93.8**	90.7
Shirt	81.4	66.7	73.8	74.3	**89.4**	74.9	82.8	83.5
Sneaker	98.4	95.4	97.6	97.2	**98.7**	94.2	98.6	98.5
Bag	86.0	70.0	79.5	81.9	73.7	79.1	89.4	**91.0**
Ankle-Boot	97.7	80.7	96.5	89.9	86.6	93.2	97.9	**98.1**
Mean	90.7	80.9	88.4	88.4	88.1	84.8	92.6	**92.7**

can be attributed to the characteristics of the CIFAR10 dataset, where images are typically blurrier, have lower resolution, and contain substantial background information, contributing to its complexity and large intra-class variance. These factors make the CIFAR10 dataset particularly challenging. The performance of the proposed DualNet method is better than previous models in the CIFAR10 dataset, which indicates that the DualNet model can well reduce the impact of sample background noise on anomaly detection tasks. Since the samples used for anomaly detection in the real world often have a lot of noise, our proposed method can reduce the impact of noise on the model performance. Figure 4a shows that our method performs better compared to other one-class classification methods. Figure 4b shows the histogram of anomaly score distribution of the test set in the inference stage, proving that the model's latent vector is separable in the feature space, proving the effectiveness of our method.

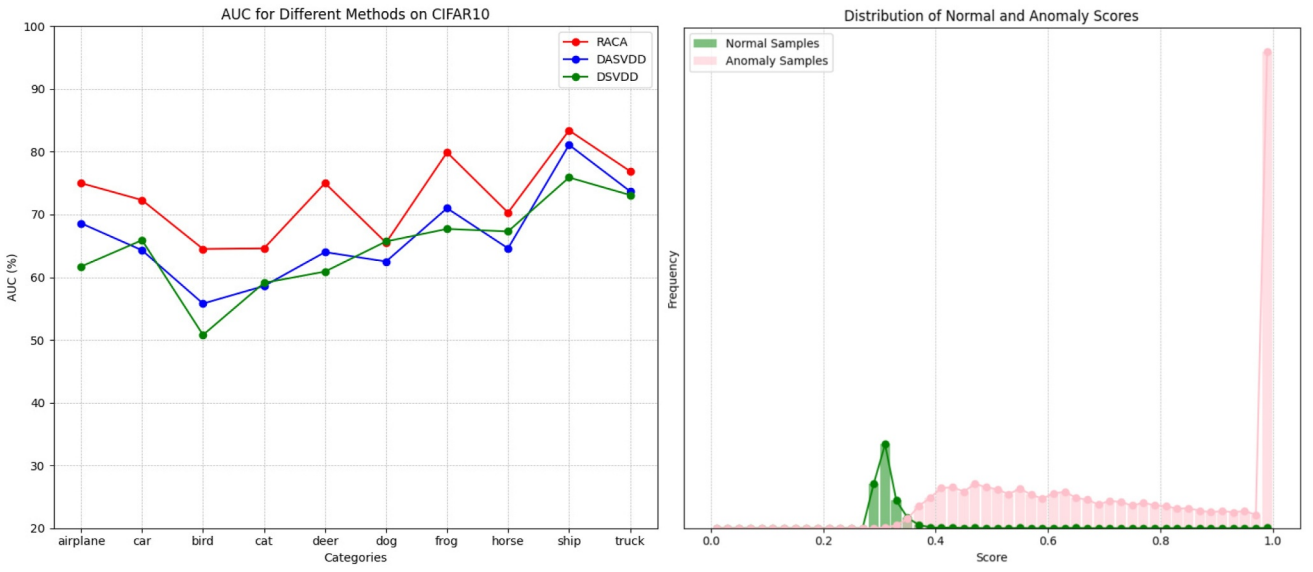
4.4 | Ablation study

To investigate the effectiveness of each additional component of the proposed work, we conducted an ablation study using the CIFAR10 dataset. In this ablation experiment, $AE + FC$ (AF) is the basic model. Specifically, we conducted ablation experiments in four different scenarios: AF, $AF + L_{adv}$,



(a) The value of AUC on CIFAR10. (b) Visualizing sample features.

FIGURE 3 (a) DualNet performance compared with other models on the CIFAR10 dataset. (b) t-SNE visualisation of the features extracted from G_{en}.



(a) AUC for Different Methods (b) The score distribution of the samples

FIGURE 4 (a) Performance and comparison between the models most similar to DualAD and the DualAD model on the CIFAR10 dataset. In addition, (b) the histogram of the anomaly score distribution of the test set in the inference stage.

TABLE 2 Ablation study for RACA performed on CIFAR10.

Target	AUC(%)
AF	66.5
$AF + L_{adv}$	71.7
$AF + L_{adv} + L_{fm}$	72.7

$AF + L_{adv} + L_{fm}$. The mean AUC for each class of the CIFAR10 dataset is tabulated in Table 2.

5 | CONCLUSION

In this paper, we introduce a novel unsupervised anomaly detection model that utilises a dual adversarial network (DualAD) for image anomaly detection. DualAD consists of a reconstruction module and an FC module and employs dual adversarial learning during the training process. The intra-module adversarial learning within the reconstruction module helps the model to obtain higher-quality reconstructed samples. The inter-module adversarial learning between the reconstruction module and the FC module enables the model to achieve a latent space more conducive to anomaly detection tasks. Additionally, the interaction between the reconstruction module and the FC module can alleviate issues related to the model's generalisation to anomalies and hypersphere collapse. In the inference phase, our method can detect anomalies in feature and pixel spaces simultaneously, ensuring no anomalies are missed. Experiments on multiple data sets demonstrate the effectiveness of the DualAD.

However, DualAD also faces several challenges: (i) In real industrial scenarios, training data may be contaminated with anomalous samples, leading the model to incorporate these anomalies as part of the normal pattern mistakenly. This can result in the model learning these anomalous features, causing a shift in the distribution of the latent space for normal data and, consequently, degrading the performance of the model. (ii) The data used for validation in this study consists of images that are relatively small in size and low in resolution and exhibit a uniform pattern. In real-world scenarios, more and larger images may encompass a wide variety of normal patterns and substantial variations, making it more challenging to learn accurate reconstruction for all normal images. (iii) Over-reliance on adversarial training and feature constraints may introduce increased complexity and computational demands, potentially affecting the scalability and efficiency of deployment on a larger scale. These challenges need to be adequately addressed when implementing the proposed algorithm in practical applications.

Future research, in addition to addressing the challenges mentioned above, should pay more attention to how to reduce the reconstruction error of normal samples and increase the reconstruction error of abnormal samples. The transference of the DualAD algorithm in other fields can also be explored.

AUTHOR CONTRIBUTIONS

Yonghao Wan: Formal analysis; Investigation; Methodology; Software; Supervision; Visualisation; Writing – original draft. **Aimin Feng**: Conceptualisation; Funding acquisition; Investigation; Methodology; Validation; Writing – review & editing.

ACKNOWLEDGEMENT

There are no funders to report for this submission.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in https://github.com/WanYonghaoSZ2216019/DualAD. These data were derived from the following resources available in the public domain: MNIST dataset: http://yann.lecun.com/exdb/mnist/FashionMNIST FashionMNIST dataset : https://github.com/zalandoresearch/fashion-mnist/tree/master/data/fashionCIFAR10 CIFAR10 dataset : http://www.cs.toronto.edu/~kriz/cifar.html.

ORCID

Yonghao Wan https://orcid.org/0009-0009-3389-5190
Aimin Feng https://orcid.org/0009-0009-3949-1883

REFERENCES

1. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. ACM Comput. Surv., 1–58 (2009). https://doi.org/10.1145/1541880.1541882
2. Liu, F., et al.: Log2vec: a heterogeneous graph embedding based approach for detecting cyber threats within enterprise. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, pp. 1777–1794 (2019)
3. Zhang, C., Costa-Perez, X., Patras, P.: Adversarial attacks against deep learning-based network intrusion detection systems and defense mechanisms. IEEE/ACM Trans. Netw. 30(3), 1294–1311 (2022). https://doi.org/10.1109/tnet.2021.3137084
4. Porwal, U., Mukund, S.: Credit card fraud detection in e-commerce: an outlier detection approach. arXiv: Learning,arXiv: Learning (2018)
5. Zong, B., et al.: Deep autoencoding Gaussian mixture model for unsupervised anomaly detection. In: International Conference on Learning Representations,International Conference on Learning Representations (2018)
6. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06) (2006). https://doi.org/10.1109/cvpr.2006.100
7. Goodfellow, I., et al.: Generative adversarial nets. Adv. Neural Inf. Process. Syst. 27 (2014)
8. Perera, P., Nallapati, R., Xiang, B.: Ocgan: one-class novelty detection using gans with constrained latent representations. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019). https://doi.org/10.1109/cvpr.2019.00301
9. Schlegl, T., et al.: Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery, pp. 146–157 (2017). https://doi.org/10.1007/978-3-319-59050-9_12
10. Zenati, H., et al.: Adversarially learned anomaly detection. In: 2018 IEEE International Conference on Data Mining (ICDM) (2018). https://doi.org/10.1109/icdm.2018.00088

11. Akcay, S., Atapour-Abarghouei, A., Breckon, T.P.: GANomaly: Semi-supervised Anomaly Detection via Adversarial Training, pp. 622–637 (2019). https://doi.org/10.1007/978-3-030-20893-6_39
12. Zaigham Zaheer, M., et al.: Old is gold: redefining the adversarially learned one-class classifier training paradigm. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020). https://doi.org/10.1109/cvpr42600.2020.01419
13. Schölkopf, B., et al.: Estimating the support of a high-dimensional distribution. Neural Comput., 1443–1471 (2001). https://doi.org/10.1162/089976601750264965
14. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. ACM Comput. Surv. 41(3), 1–58 (2009). https://doi.org/10.1145/1541880.1541882
15. Ruff, L., et al.: Deep one-class classification. In: International Conference on Machine Learning, pp. 4393–4402. PMLR (2018)
16. Makhzani, A., et al.: Adversarial autoencoders. arXiv preprint arXiv:1511.05644 (2015)
17. Pang, G., et al.: Deep learning for anomaly detection: a review. ACM Comput. Surv., 1–38 (2022). https://doi.org/10.1145/3439950
18. Yi, J., Yoon, S.: Patch SVDD: Patch-Level SVDD for Anomaly Detection and Segmentation, pp. 375–390 (2021). https://doi.org/10.1007/978-3-030-69544-6_23
19. Zhang, Z., Deng, X.: Anomaly detection using improved deep svdd model with data structure preservation. Pattern Recogn. Lett., 1–6 (2021). https://doi.org/10.1016/j.patrec.2021.04.020
20. Hu, C., Chen, K., Shao, H.: A semantic-enhanced method based on deep svdd for pixel-wise anomaly detection. In: 2021 IEEE International Conference on Multimedia and Expo (ICME) (2021). https://doi.org/10.1109/icme51207.2021.9428370
21. Massoli, F.V., et al.: Mocca: multi-layer one-class classification for anomaly detection. IEEE Transact. Neural Networks Learn. Syst., 2313–2323 (2022). https://doi.org/10.1109/tnnls.2021.3130074
22. Sauter, D., et al.: Defect detection of metal nuts applying convolutional neural networks. In: 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC) (2021). https://doi.org/10.1109/compsac51774.2021.00043
23. Hojjati, H., Armanfard, N.: Dasvdd: Deep Autoencoding Support Vector Data Descriptor for Anomaly Detection. Cornell University - arXiv, Cornell University - arXiv (2021)
24. Liznerski, P., et al.: Explainable Deep One-Class Classification. Cornell University - arXiv (2020). Learning
25. Zhou, D., et al.: Batch quadratic programming network with maximum entropy constraint for anomaly detection. IET Comput. Vis., 230–240 (2022). https://doi.org/10.1049/cvi2.12082
26. Sohn, K., et al.: Learning and Evaluating Representations for Deep One-Class Classification. Cornell University - arXiv,Learning (2020)
27. Zhou, Y., et al.: Vae-based deep svdd for anomaly detection. Neurocomputing, 131–140 (2021). https://doi.org/10.1016/j.neucom.2021.04.089
28. Chen, Y., et al.: Deep one-class classification via interpolated Gaussian descriptor. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 383–392 (2022). https://doi.org/10.1609/aaai.v36i1.19915
29. Gia, T.L., Yi, D., Ahn, J.: Robust deep support vector data description for unreliable data. In: 2022 IEEE/ACIS 7th International Conference on Big Data, Cloud Computing, and Data Science (BCD) (2022). https://doi.org/10.1109/bcd54882.2022.9900580
30. Jiang, M., Tseng, S., Su, C.: Two-phase clustering process for outliers detection. Pattern Recogn. Lett. 22(6–7), 691–700 (2001). https://doi.org/10.1016/s0167-8655(00)00131-8
31. He, Z., Xu, X., Deng, S.: Discovering cluster-based local outliers. Pattern Recogn. Lett. 24(9–10), 1641–1650 (2003). https://doi.org/10.1016/s0167-8655(03)00003-5
32. Schubert, E., et al.: Dbscan revisited, revisited: why and how you should (still) use dbscan. ACM Trans. Database Syst., 1–21 (2017). https://doi.org/10.1145/3068335
33. Artola, A., et al.: Glad: a global-to-local anomaly detector. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2023). https://doi.org/10.1109/wacv56688.2023.00546
34. Yang, X., et al.: Deep spectral clustering using dual autoencoder network. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019). https://doi.org/10.1109/cvpr.2019.00419
35. Sabokrou, M., et al.: Adversarially learned one-class classifier for novelty detection. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2001). https://doi.org/10.1109/cvpr.2018.00356
36. Vincent, P., et al.: Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research,Journal of Machine Learning Research (2010)
37. Ngo, C., et al.: Fence gan: towards better anomaly detection. arXiv: Learning,arXiv: Learning (2019)
38. Liu, Z., et al.: Simplenet: a simple network for image anomaly detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20402–20411 (2023)
39. Gong, D., et al.: Memorizing normality to detect anomaly: memory-augmented deep autoencoder for unsupervised anomaly detection. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019). https://doi.org/10.1109/iccv.2019.00179
40. Kingma, D., Welling, M.: Auto-encoding variational bayes. arXiv: Mach. Learn. (2013)
41. Zavrtanik, V., Kristan, M., Skočaj, D.: Draem – a discriminatively trained reconstruction embedding for surface anomaly detection. In: International Conference on Computer Vision,International Conference on Computer Vision (2021)
42. Xing, P., Sun, Y., Li, Z.: Self-supervised Guided Segmentation Framework for Unsupervised Anomaly Detection (2022)
43. Zavrtanik, V., Kristan, M., Skočaj, D.: Dsr -a dual subspace re-projection network for surface anomaly detection
44. Schlüter, H., et al.: Natural Synthetic Anomalies for Self-Supervised Anomaly Detection and Localization (2021)
45. Bauer, A.: Self-supervised Training with Autoencoders for Visual Anomaly Detection (2022)
46. Salimans, T., et al.: Improved Techniques for Training Gans. Cornell University - arXiv,Cornell University - arXiv (2016)
47. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. arXiv: Learning,arXiv: Learning (2014)
48. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. In: Conference on Learning Theory,Conference on Learning Theory (2010)
49. LeCun, Y.: The mnist database of handwritten digits. http://yann.lecun.com/exdb/mnist/ (1998)
50. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv: Learning,arXiv: Learning (2017)
51. Krizhevsky, A.: Learning Multiple Layers of Features from Tiny Images (2009)

How to cite this article: Wan, Y., Feng, A.: DualAD: dual adversarial network for image anomaly detection★. IET Comput. Vis. 18(8), 1138–1148 (2024). https://doi.org/10.1049/cvi2.12297