IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 25, 2019, accepted May 12, 2019, date of publication June 3, 2019, date of current version July 22, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2920251

A Novel and Efficient CVAE-GAN-Based Approach With Informative Manifold for Semi-Supervised Anomaly Detection

JIANG BIAN[1,2], XIAOLONG HUI[1], SHIYING SUN[1], XIAOGUANG ZHAO[1], AND MIN TAN[1]
[1]State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100190, China

Corresponding author: Xiaolong Hui (huixiaolong2015@ia.ac.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61421004 and Grant 61673378.

ABSTRACT Semi-supervised anomaly detection identifies abnormal (testing) observations which are different from normal (training) observations. In many practical situations, anomalies are poorly insufficient and not well defined, while the normal data are easily sampled, have a wide variety, and may not be classified. For this paradigm, we propose a novel end-to-end deep network as an anomaly detector only trained on normal samples. Our architecture consists of a conditional variational auto-encoder (CVAE), a feature discriminator (FD), and an adversarially trained WGAN-GP discriminator. The CVAE is designed as a generator to reconstruct images. It leverages underlying category information and multivariate Gaussian distributions to regularize the latent space, enabling a smooth and informative manifold. For anomalies which have a certain similarity to normal data, we perform active negative training by generating potential outliers from the latent space to limit network generative capability. In order to capture data characteristics, we maximize the mutual information between the inputs and the latent codes by the FD. It enhances the relationship between the high-dimensional image space and corresponding encoded vectors. To promote reconstruction, a structural similarity loss is applied to robustly recover local texture details and the WGAN-GP discriminator is employed to aid in generating photo-realistic images. We distinguish anomalies by computing a reconstruction-based anomaly score. Different from recent encoder–decoder or GAN-based architectures, our approach considers input categories, constructs, and exploits a useful manifold in an unsupervised manner and has a stronger reconstruction capability. The experimental results demonstrate that the proposed approach outperforms state-of-the-art methods over several benchmark datasets.

INDEX TERMS Semi-supervised anomaly detection, conditional variational auto-encoder, generative adversarial networks, informative manifold, structural similarity loss.

I. INTRODUCTION

In many practical applications, such as vision-based industrial fault monitoring [1], [2] and medical image based disease diagnosis [3]–[5], people would like to detect the anomalies that don't belong to any of the known classes so as to determine if the situation is within normal. In these cases, pictures of normal categories are available. However, the anomalies can't be well defined or sampled because they occur irregularly and also show great diversification. This is a typical semi-supervised anomaly detection task in the field of computer vision, which task only takes advantage of the normal observations for training. It attracts great interest of researchers and is closely similar to novelty detection [6], one-class classification [7], outlier detection [8] and irregularity detection [9] studies.

There has been a considerable volume of works proposing many different anomaly detection methods for videos and images, some of which are summarized in the overviews like [10]–[13]. Common popular approaches can be classified as self-representation learning [14]–[17] and statistical modeling [18]–[21]. Recently, the performance of anomaly detection is greatly improved by applying deep adversarial training process [3], [9], [22]–[25].

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

Self-representation learning is a powerful tool for anomaly detection, of which the feature representation and the data reconstruction are the two important components. The feature representation learns unique features of normal categories and rejects anomilies which don't conform to these features. Low level features such as gradient features [26], [27], mixtures of textures [28] and improved PCA [29] have been widely used in the last dozen years. High level features from deep networks, like auto-encoders [30], pre-trained networks [31] and PCAnet [32], [33], have achieved more excellent successes. Besides, due to lacking of sparse representations, anomalies can be distinguished by sparsity which is learned from the normal classes. For example, [34] and [17] utilize the sparse model for the detection of abnormal events and videos. And [15] detects anomalies in a union of subspaces. In addition, the data reconstruction method of self-representation learning can decide whether a sample belongs to normal classes or not by a reconstruction error of deep neural networks. Typically, the error is based on an encoder-decoder network and is minimized by training on normal samples [35], [36].

Statistical modeling learns the data distributions from normal samples, which distributions are usually expressed by parameters. Classical [37] and [38] are distance-based anomaly detection approach. Anomalies are identified by measuring their distances to the neighboring samples. LOF [39], a work also based on distance measurement, utilizes the k-nearest neighbors to estimate the local density. The CoP [40] identifies an anomaly which has a low mutual coherence with the rest of the data points. In [41], an anomaly measure of a sample is obtained by distances between its projection and the projected single point of training samples in each class.

In recent years, advances in Generative Adversarial Networks (GANs) [42] have opened new possibilities for semi-supervised anomaly detection. GANs can model complicated and high-dimensional distributions, especially the images [24], through a min-max game process. The learning models can successfully generate data with outstanding performance [42], [43]. Schlegl *et al.* [3] propose an AnoGAN which uses the similar convolutional structure of the DCGAN [24], to learn a generator only utilizing normal images. The posterior probability of a testing sample is optimized to reconstruct the sample by the generator. Finally, an anomaly score based on the reconstructed image and the feature map of the discriminator is calculated to discover abnormal markers in medical images. Later on, in order to reduce the complexity of mapping from image to latent space, Zenati *et al.* [25] jointly train them by making good use of the BiGAN [44] structure and distinguish anomalies with the same anomaly scores. In a follow-up study, Akcay *et al.* [45] propose GANomaly comprising an encoder-decoder-encoder network groups to explore the deep latent representation of the normal samples and adopt an anomaly score computed from the latent spaces. Their work announce achieving state-of-the-art performance statistically over benchmarks.

Different from the [3], [25], [45], Sabokrou *et al.* [22] present a new framework for anomaly detection, which leverages the reconstruction error to train a one-class classifier instead of computing the anomaly score. In [23], Pidhorskyi *et al.* propose a similar structure based on the GAN and the encoder-decoder network. But he computes an anomaly probability indicating the possibility that the sample belongs to normal data.

These most recent methods [3], [22], [23], [25], [45] are the successful GAN-based generative approaches for detecting unknown abnormal images. Nevertheless, they all suffer from mode collapsing problem [46] of GANs. Besides, the latent space of traditional encoders used in [3], [22], [25], [45], is not a disentangled representation nor a smooth manifold [47]. As a result, the latent manifold has less useful information to exploit and can't accurately describe the intrinsic characteristics of the normal samples [47], [48]. Furthermore, the reconstruction of [3], [22], [23], [25], [45] is performed by a pixel-wise L-norm loss which treats all pixels independent. It lacks consideration of the inter-pixel relationship, which prevents these methods from being applied to real-world scenes. In addition, the decoder of traditional auto-encoder is not robust to noises [47], [49] either.

In order to further improve the accuracy of GAN-based anomaly detection, negative training and manifold regulations can be used. The negative training utilizes the abnormal samples to increase the identification ability of anomalies. Munawar *et al.* [50] introduce a negative training stage to unlearn the anomaly reconstruction, which can also limit the strong generative capability of the GANs. Kimura and Yanagihara [51] exclude the distribution of abnormal images to reduce the influence of noisy normal data. However these methods assume the abnormal images are easily accessible. It's not an assumption for the semi-supervised problem. Therefore, we propose an active negative training approach. The active negative training can generate abnormal samples and utilizes the abnormal samples to conduct negative training.

With respect to the regulations of manifold, it can provide a desired representation of normal data. Gray *et al.* [52] present an IGMM-GAN model to cope with multiple types of data, in which the BiGAN [44] is combined with an infinite Gaussian mixture model [53]. Munawar *et al.* [50] use adversarial autoencoders (AAE) [54] to impose a supervised prior distribution on the latent space. Thus it can map the anomaly items away from the normal data. However, [52] requires the labels of normal samples and [50] needs to get some random inputs as anomalies. In this paper, the problem faced with is more general where normal samples are unlabeled and anomalies are not well defined.

With regard to the informative latent space, generative architectures are usually used to build good latent representations by virtue of reconstruction [44], [47], [55], [56]. A high-quality representation from the encoder is beneficial to the downstream tasks, such as classification task [57].

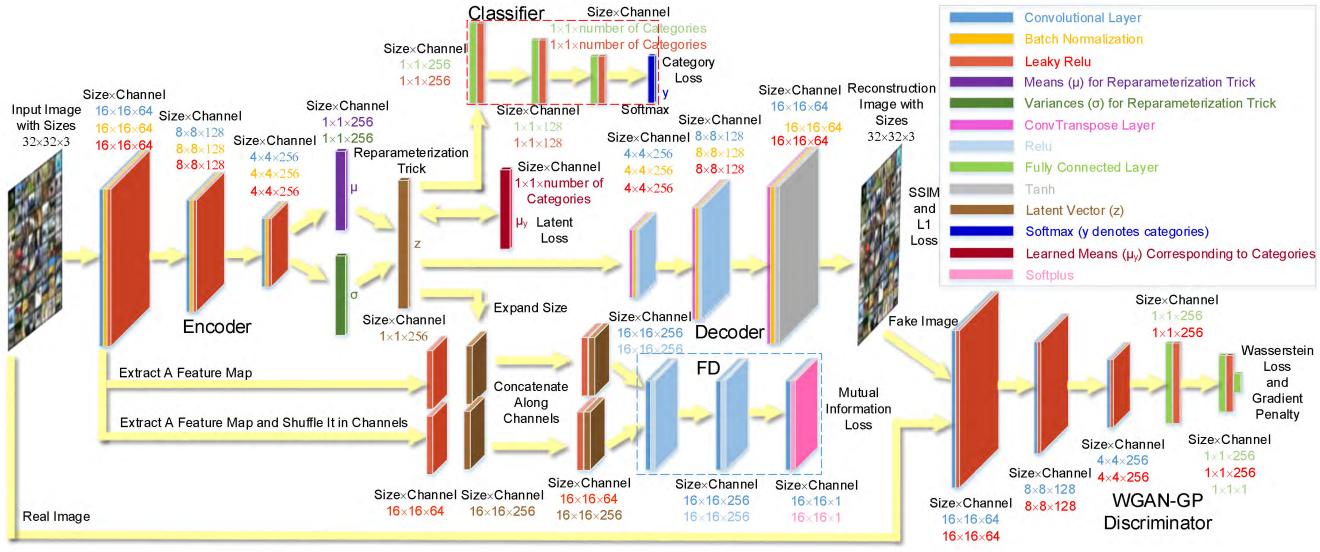

FIGURE 1. The proposed network architecture for anomaly detection. It is composed of a conditional variational auto-encoder(CVAE), a feature discriminator (FD) and a WGAN-GP discriminator. The CVAE comprises an encoder, a decoder, a classifier which is surrounded by the red dashed rectangle, and means corresponding to input categories. The FD is denoted by the blue dashed rectangle.

To obtain useful representations of normal samples without using labels, an unsupervised way is used to maximize the mutual information between the normal samples and the corresponding latent codes [57].

In this paper, we propose a novel end-to-end deep convolutional networks, which are only trained on unclassified normal images, to detect unknown anomalies. We don't use the auto-encoder structures but a designed conditional variational auto-encoder (CVAE). Different from variational auto-encoder (VAE) [47], it can take advantage of potential class information of training samples to generate a manifold which is regularized by multivariate Gaussian distributions with learnable mean values. The manifold of CVAE is smooth and disentangled. Besides the converged generative model is more robust to noises. We add the maximization of mutual information into the proposed approach. Therefore, the latent codes are encouraged to learn more useful representations associated with the inputs. In order to promote the convergence of network and the accuracy of reconstruction, we use the structural similarity (SSIM) metric [58] as a loss function and adopt the WGAN-GP [59] framework. SSIM is a measurement to capture salient differences between the input and the reconstruction. WGAN-GP can generate photo-realistic images. By virtue of Wasserstein loss [60] and gradient penalty [59], the WGAN-GP avoids the mode collapsing problem caused by vanilla GAN and feature matching based GAN [43]. In order to enhance the identification ability between normal samples and similar abnormal samples, potential outliers are generated from the regularized latent space and are involved in the active negative training. To describe the degree of abnormality, we present a reconstruction-based anomaly score, comprising the SSIM and L1-norm losses. Experimental results demonstrate that our approach achieves excellent results for anomaly detection, and outperforms several recent state-of-the-art works on various benchmarks.

The remainder of the paper is structured as follows. Section II describes the proposed approach including conditional variational auto-encoder, mutual information maximization, structural similarity, WGAN-GP, active negative training and anomaly score. Section III gives the experimental results and analyses in detail. The conclusions are summarized in Section IV.

II. THE PROPOSED APPROACH

Fig. 1 illustrates the overview of the proposed CVAE-GAN-based architecture. The CVAE learns the input data representation by the encoder and works as a generator to reconstruct the image via the decoder. The encoder adopts three identical pipelines which are composed of the convolutional layers followed by the batch-norm and the leaky ReLU activation. The decoder utilizes the similar structure of a DCGAN generator [24]. As shown in Fig. 1, the CVAE uses reparameterization trick [47], [61] to generate latent encoded vector z from u and σ, and the classifier denoted by red rectangle associates z with the input category y. The learnable means u_y are parameters of the multivariate Gaussian distributions which are imposed on z. The WGAN-GP discriminator is similar to the encoder but abandons the batch-norm layers and attaches fully connected layers to the convolutional layers as output. The feature discriminator (FD) uses 1×1 convolutional kernels to identify the combination between z and the true feature map. In this way, the mutual information between z and the inputs can be maximized.

A. CONDITIONAL VARIATIONAL AUTO-ENCODER

VAE [47] is a state-of-the-art image modeling technique and is known to generate a smooth, continuous and disentangled

latent manifold. (1) shows the reparameterization trick of VAE, where $\odot$ denotes an element-wise product, $N(0, 1)$ is a standard Normal distribution.

$$z = u + \sigma \odot \xi \quad \xi \in N(0, 1) \tag{1}$$

VAE utilizes this trick to realize random sampling from multivariate Gaussian distributions. As shown in (2), as shown at the bottom of this page, it learns to generate data which maximizes a variational lower bound of the model log-likelihood $E_{x\sim p_{data}(x)}\left[\log P_{model}(x)\right]$.

As illustrated in (2), the primary idea of the VAE is the stochastic variational inference which minimizes a reconstruction term $E_{x\sim p_{data}(x)}\left[E_{z\sim p(z|x)}\left[-\log q\left(x|z\right)\right]\right]$, a Kullback-Leibler (KL) divergence term $KL\left(p(z)||q(z)\right)$ and a mutual information term $I(z; x)$. In (2), x is a data point from the distribution $p_{data}(x)$, $p\left(z|x\right)$ is a variational approximate posterior [47] which is modeled by the encoder. The VAE can match an arbitrary prior $q(z)$, which is usually defined as standard Normal distribution, to the aggregated posterior distribution $p(z)$. Furthermore, $q\left(x|z\right)$ represents a data generator which is depicted by the decoder. Additionally, E denotes mathematical expectation and is calculated over the training batches.

In order to handle the inputs with multiple classes, we design a CVAE, a kind of recent advanced model, to replace VAE. It improves VAE and is able to generate data conditioned on certain attributes. To be more general, we assume the training data is not classified. Without labels, we implement our CVAE in an unsupervised way. We regard the latent code as a combination of z and y instead of only z, where y is a discrete latent variable that represents a category. As a result, we replace the z in (3), as shown at the bottom of this page, by (z, y). In this way, the CVAE not only keeps all the features of VAE but also learns a more informative manifold. (3) is another derivation of (2). Based on (3), the loss of CVAE is derived from (4), as shown at the bottom of this page. In (4), we model the $q(z|y)$ as multivariate Gaussian distributions with variances 1 and learnable means u_y for different categories y. Thus the $q(z|y)$ can be expressed by (5), as shown at the bottom of this page. In (5), the covariance matrix of z is a diagonal matrix. Because we hope that every component of the latent vector z is independent so that they are maximally informative. Besides, d represents the dimension of vector z. In (4), $p(y|z)$ is a classifier for hidden variables. Its architecture is denoted by the red dashed rectangle in Fig. 1. $q(y)$ is the prior distribution

$$\begin{aligned}
&E_{x\sim p_{data}(x)}\left[\log P_{model}(x)\right] \\
&\quad > -E_{x\sim p_{data}(x)}\left[E_{z\sim p(z|x)}\left[-\log q\left(x|z\right)\right]\right] - E_{x\sim p_{data}(x)}\left[KL\left(p\left(z|x\right)||q\left(z\right)\right)\right] \\
&\quad = -E_{x\sim p_{data}(x)}\left[E_{z\sim p(z|x)}\left[-\log q\left(x|z\right)\right]\right] - \int\int p(z|x)p_{data}(x)\log\frac{p(z|x)p(z)}{q(z)p(z|x)}dzdx - \int\int p(z|x)p_{data}(x)\log\frac{p(z|x)}{p(z)}dzdx \\
&\quad = -E_{x\sim p_{data}(x)}\left[E_{z\sim p(z|x)}\left[-\log q\left(x|z\right)\right]\right] - KL\left(p(z)||q(z)\right) - I(z; x) \\
&\quad = -E_{x\sim p_{data}(x)}\left[E_{z\sim p(z|x)}\left[-\log q\left(x|z\right)\right]\right] - \int p(z)\log\frac{p(z)}{q(z)}dz - I(z; x)
\end{aligned} \tag{2}$$

$$\begin{aligned}
&E_{x\sim p_{data}(x)}\left[\log P_{model}(x)\right] \\
&\quad > -E_{x\sim p_{data}(x)}\left[E_{z\sim p(z|x)}\left[-\log q\left(x|z\right)\right]\right] - E_{x\sim p_{data}(x)}\left[KL\left(p\left(z|x\right)||q\left(z\right)\right)\right] \\
&\quad = -E_{x\sim p_{data}(x)}\left[-\int p(z|x)\log q(x|z)dz + \int p(z|x)\log\frac{p(z|x)}{q(z)}dz\right] \\
&\quad = -E_{x\sim p_{data}(x)}\left[\int p(z|x)\log\frac{p(z|x)}{q(x, z)}dz\right]
\end{aligned} \tag{3}$$

$$\begin{aligned}
&\sum_y E_{x\sim p_{data}(x)}\left[\int p(z, y|x)\log\frac{p(z, y|x)}{q(x, z, y)}dz\right] \\
&\quad = E_{x\sim p_{data}(x)}\left[\sum_y\int p(y|z)p(z|x)\log\frac{p(y|z)p(z|x)}{q(x|z, y)q(z|y)q(y)}dz\right] \\
&\quad = E_{x\sim p_{data}(x)}\left[E_{z\sim p(z|x)}\left[-\sum_y p(y|z)\log q(x|z, y) + \sum_y p(y|z)\log\frac{p(z|x)}{q(z|y)} + KL\left(p(y|z)||q(y)\right)\right]\right] \\
&\quad = E_{x\sim p_{data}(x)}\left[E_{z\sim p(z|x)}\left[-E_{y\sim p(y|z)}\left[\log q(x|z, y)\right] + \sum_y p(y|z)\log\frac{p(z|x)}{q(z|y)} + KL\left(p(y|z)||q(y)\right)\right]\right]
\end{aligned} \tag{4}$$

$$q(z|y) = \frac{1}{(2\pi)^{d/2}}e^{-\frac{1}{2}||z-\mu_y||^2} \tag{5}$$

of categories. Without labels, we just estimate and set the number of the categories for y. This number is also the number of different multivariate Gaussian distributions. The $E_{x\sim p_{data}(x)}\left[E_{z\sim p(z|x)}\left[-E_{y\sim p(y|z)}\left[\log q(x|z,y)\right]\right]\right]$ is a reconstruction term. The latent loss of CVAE is $E_{x\sim p_{data}(x)}[E_{z\sim p(z|x)}[\sum_y p(y|z)\log\frac{p(z|x)}{q(z|y)}]]$. It is capable of forcing each z to be as close as possible to its corresponding multivariate Gaussian distribution which represents one of the classes. It is more natural than standard Normal distribution with 0 mean value in VAE and plays a role of unsupervised clustering in latent space. In VAE, the $p(z|x)$ is assumed to be multivariate Gaussian distributions with diagonal covariance matrixes [47]. So the $\frac{p(z|x)}{q(z|y)}$ in CVAE can be embodied by the ratio of the two Gaussian distributions. Moreover, the $E_{z\sim p(z|x)}\left[KL\left(p(y|z)||q(y)\right)\right]$ is a categorical loss term. The optimization of the categorical loss can reduce the KL divergence between $p(y|z)$ and $q(y)$. Without any prior distribution of categories y, $q(y)$ can be assumed to be uniformly distributed. So, the categorical loss is optional, it can be realized by the cross entropy and can force the normal classes to be evenly distributed.

Another method AAE [54] can also make the manifold continuous, smooth and disentangled. However, the AAE can't provide learnable means of multivariate Gaussian distributions as the CVAE does. The learnable means can cluster the same kind of objects in the latent space. They make the latent space disentangled and conditioned on the input categories. In this way, the CVAE enhances the relationship between the input normal data and the manifold in an unsupervised way and the network can learn more features of normal objects. So we designed the CVAE.

B. MUTUAL INFORMATION MAXIMIZATION

As shown in (2), the optimal solution of the VAE is to minimize the reconstruction term, simultaneously minimize the mutual information term. When a simple decoder is used, minimizing the reconstruction term will force the latent codes z relevant to input data, which leads to a maximization of $I(z;x)$ [62]. However, this good situation doesn't often occur. When the decoder is powerful, the mutual information $I(z;x)$ in (2) will be minimized without being affected by the minimization of the other two terms [63], [64]. Minimizing the mutual information makes the inputs x and the latent codes z independent, which means the latent codes don't learn useful representations. Therefore, the mutual information term needs to be maximized to learn an informative manifold and at the same time to aid in promoting the reconstruction accuracy of normal data. Finally, the trained network is able to extract more unique features from the training set.

$$\begin{aligned} I(x,z) &= \int\int p(x,z)\log\frac{p(x,z)}{p(x)p(z)}dxdz \\ &= \int\int p(z|x)p_{data}(x)\log\frac{p(z|x)}{p(z)}dxdz \\ &= KL(p(z|x)p_{data}(x)||p(z)p_{data}(x)) \end{aligned} \tag{6}$$

$$JS(P,Q) = \frac{1}{2}KL(P||\frac{P+Q}{2}) + \frac{1}{2}KL(Q||\frac{P+Q}{2}) \tag{7}$$

$$\begin{aligned} \min_G\max_D\ V(D,G) &= E_{x\sim p_r}\left[\log D(x)\right] \\ &\quad + E_{x\sim p_g}\left[\log(1-D(x))\right] \end{aligned} \tag{8}$$

$$\min_G V(D^*,G) = 2JS(p_r||p_g) - 2\log 2 \tag{9}$$

According to the mutual information definition, the mutual information can be derived as (6). Maximizing the mutual information between the inputs x and the latent codes z is equivalent to enlarging the KL divergence between $p(z|x)p_{data}(x)$ and $p(z)p_{data}(x)$. The larger KL divergence indicates a larger $\frac{p(z|x)}{p(z)}$. It means that for each data x, the encoder $p(z|x)$ can encode the unique z so that the probability of $p(z|x)$ is much larger than the probability of random prior distribution $p(z)$. That is to say, the encoder learns the information from the inputs x and has a high probability $p(z|x)$ to generate the unique latent codes z. When x is known, the uncertainty of z is greatly reduced. It is also the meaning of mutual information. Because the KL divergence doesn't have an upper bound. We replace the KL divergence by the Jensen-Shannon (JS) divergence which has a same effect to measure the distribution distance and has an upper bound of $\frac{1}{2}\log 2$. The upper bound enables a stable maximization process. The JS divergence is widely used in traditional GAN training and is shown in (7), where P and Q are two data distributions.

(8) is the classical GAN objective proposed in [42]. G and D are a generator and a discriminator respectively. Besides p_r and p_g are separately the real data distribution and the data distribution generated from G. The optimal (maximal) discriminator in (8) has a form of $D^*(x) = \frac{p_r(x)}{p_r(x)+p_g(x)}$. Then (9) can be derived by putting $D^*(x)$ into (8). So the maximization process in (8) is to force the GAN objective to approximate the JS divergence. From (9), it can be seen that minimizing $V(D^*,G)$ is to find an optimal G to reduce the JS divergence between the p_g and p_r [65]. On the contrary, the maximization of mutual information requires increasing the JS divergence. So by imitating (8), we replace the minimization process by maximization process to find a $p(z|x)$ which can maximize the JS divergence between $p(z|x)p_{data}(x)$ and $p(z)p_{data}(x)$. The JS-based mutual information loss can be denoted as (10), as shown at the top of the next page. The FD in (10) is a discriminator similar to D in (8). The architecture of the network FD is denoted by the blue dashed rectangle in Fig. 1.

C. STRUCTURAL SIMILARITY

SSIM is able to depict inter-dependencies between two $K\times K$ sized patches e and f of an image to make up for the widely used pixel-wise L1-norm loss. The SSIM considers image similarity in terms of luminance $l(e,f)$, contrast $c(e,f)$, and structure $s(e,f)$. It can be expressed by (11), where μ_e and μ_f are mean intensities of patches, σ_e^2 and σ_f^2 denote the patch variances, σ_{ef} represents the covariance of two patches, $c1$ and $c2$ are two constants to ensure numerical stability and are

$$\max_{p(z|x),FD} E_{(x,z)\sim p(z)p_{data}(x)}\left[\log FD(x,z)\right] + E_{(x,z)\sim p(z|x)p_{data}(x)}\left[\log\left(1 - FD(x,z)\right)\right] \tag{10}$$

typically 0.01 and 0.03. In our proposed approach, the entire reconstruction loss is constituted of the negative log of SSIM loss and the per-pixel L1-norm loss.

$$\begin{aligned} SSIM(e,f) &= l(e,f)c(e,f)s(e,f) \\ &= \frac{2\mu_e\mu_f + c_1}{\mu_e^2 + \mu_f^2 + c_1} \cdot \frac{2\sigma_e\sigma_f + c_2}{\sigma_e^2 + \sigma_f^2 + c_2} \cdot \frac{2\sigma_{ef} + \frac{c_2}{2}}{\sigma_e\sigma_f + \frac{c_2}{2}} \\ &= \frac{\left(2\mu_e\mu_f + c_1\right)\left(2\sigma_{ef} + c_2\right)}{\left(\mu_e^2 + \mu_f^2 + c_1\right)\left(\sigma_e^2 + \sigma_f^2 + c_2\right)} \end{aligned} \tag{11}$$

D. WASSERSTEIN GAN

$$W(p_r, p_g) = \frac{1}{K} \sup_{||f||_L \le K} E_{x\sim p_r}\left[f\left(x\right)\right] - E_{x\sim p_g}\left[f\left(x\right)\right] \tag{12}$$

Traditional GAN optimizes the Jensen-Shannon (JS) divergence which leads to a mode collapse problem and unstable generator gradients [65] for GANs. The Wasserstein loss, also known as the earth-mover distance, is better for GAN training [65]. It can be expressed as (12), where $W(p_r, p_g)$ represents the Wasserstein distance of two distributions p_r and p_g, sup indicates maximization, f is defined in the real number field and denotes a set of functions which satisfy the Lipschitz condition ($||f||_L \le K,\ K \ge 0$). By utilizing the Wasserstein loss, the Wasserstein GAN (WGAN) [60] has stable convergent performance. It can reduce blurriness and add more local details to the generated images.

$$\begin{aligned} L_G &= -E_{z\sim p(z)}\left[D\left(G\left(z\right)\right)\right] \\ L_D &= -E_{x\sim p_{data}(x)}\left[D\left(x\right)\right] + E_{z\sim p(z)}\left[D\left(G\left(z\right)\right)\right] \end{aligned} \tag{13}$$

The WGAN is composed of a generator network G (the CVAE) and a discriminator network D. The discriminator is updated several times, and subsequently, the generator is updated once. In this way, the training loss can be reduced. As illustrated in (13), L_G and L_D are the generator loss and the discriminator loss respectively. Minimizing L_D in (13) is exactly the same as approximating the Wasserstein distance $W(p_r, p_g)$ in (12) [60]. In further, training D via the L_D can let the D distinguish the difference between distribution $p_{data}(x)$ and the generated distribution $p_{x\sim G(z)}\left(x\right)$. Because, to satisfy the L_D, the $E_{x\sim p_{data}(x)}\left[D\left(x\right)\right]$ is maximized and the $E_{z\sim p(z)}\left[D\left(G\left(z\right)\right)\right]$ is minimized. With respect to the L_G, small value of L_G means samples generated from the latent space $p(z)$ are almost the same as the samples from $p_{data}(x)$ from perspective of the discriminator.

$$\begin{aligned} GP|_{\hat{x}} &= E_{\hat{x}}\left[\left(\left\|\nabla_{\hat{x}} D\left(\hat{x}\right)\right\|_2 - 1\right)^2\right] \\ \hat{x} &= tx + (1-t)G(z) \\ x &\sim p_{data}(x) \\ z &\sim p(z) \end{aligned} \tag{14}$$

$$L_D = -E_{x\sim p_{data}(x)}\left[D\left(x\right)\right] + E_{z\sim p(z)}\left[D\left(G\left(z\right)\right)\right] + \lambda GP|_{\hat{x}} \tag{15}$$

In WGAN, the f in (12) uses weight clipping to satisfy the Lipschitz condition. However, it leads to optimization difficulties due to gradient exploding or vanishing problems [59]. WGAN-GP is an improved strategy to replace the weight clipping. It imposes a gradient penalty on the discriminator to enforce the Lipschitz constraint. The gradient penalty is shown in (14), where t is sampled uniformly from $[0, 1]$. With the gradient penalty, the discriminator loss is finally defined as (15), where λ is a proportionality coefficient. It further improves the training stability and can generate higher-quality images.

E. ACTIVE NEGATIVE TRAINING

The GAN-based architecture usually has a great generative ability. In particular, the anomalies similar to normal data may have a good reconstruction result. In order to prevent a high-quality generation of abnormal samples, abnormal samples can be used in the negative training to limit their reconstruction accuracy. In this paper, we assume the anomalies cannot be sampled or well defined. So, abnormal samples need to be generated actively. Due to the complexity and high dimensions of the image data, it is very difficult to generate abnormal samples in the image data space. Instead of searching in image space, the proposed active negative training method finds potential anomalies from the regularized low-dimensional manifold.

We consider an inherent property that the distributions of anomalies are very different from normal data and are also usually scattered. In the proposed approach, the CVAE imposes multivariate Gaussian distributions on the latent manifold for normal samples, forcing the latent codes of normal data to be clustered. As a consequence, the anomalies are not as concentrated as the normal data. It is also consistent with many other researches [66]. The latent vectors that are far away from the Gaussian means are considered as possible anomalies.

We adopt a random sampling strategy in the latent subspace to generate potential latent codes of anomalies. Their distances to the learnable means are required to be larger than 2 (twice the corresponding variance value 1). This distance above can be tuned based on the degree of similarity between abnormal and normal samples. For example, if abnormal and normal samples aren't similar to each other, the distance can be larger. In further, the latent loss can reflect the distribution distance between the latent vectors of normal samples and the multivariate Gaussian samples. So if the latent loss is larger than a threshold, the active negative training can be suspended until the latent loss is smaller.

F. ANOMALY SCORE

Our approach can reconstruct normal data with high accuracy and has poor reconstruction effect for unknown abnormal data. So based on the reconstruction, the proposed anomaly score is the sum of the SSIM and L1-norm losses. The SSIM loss compares the texture similarity between local regions of two images and the L1-norm loss examines the single pixel value of two images. The two losses are computed for each individual test sample. In order to balance the impacts of the two losses, each loss is normalized between 0 and 1 within the whole test set.

III. EXPERIMENTS AND ANALYSES

In this section, we evaluate and analyze the performance of the proposed anomaly detection architecture in detail. The results are compared with several state-of-the-art approaches qualitatively and quantitatively over four different benchmarks.

A. DATASETS DESCRIPTION

The anomaly detection tasks are constructed from the following public datasets.

1) MNIST

MNIST [67] consists of 70000 28×28 grayscale handwritten digits from 0 to 9. Each digit has 7000 images, among them 6000 for training and 1000 for testing. This dataset is the simplest in the four benchmarks and is not difficult to be trained by the neural networks. As a consequence, the anomalies can be reconstructed via learned information from normal data, leading to a decrease of detection accuracy.

2) CIFAR10

CIFAR10 [68] contains 60000 32×32 color images in 10 classes. Each category has 5000 training samples and 1000 testing samples for a total of 6000 images. The images in this benchmark have nature objects and backgrounds and are like real world photos. It is the most difficult dataset to be trained in the four benchmarks. As a result, the generalization ability of networks needs to be improved so that other normal objects which don't exist in the training set can be reconstructed well. Moreover, the anomaly reconstruction needs to be degraded, especially for anomalies which have similar structures to normal objects.

3) FASHION-MNIST

Fashion-MNIST [69] is a recent proposed dataset designed for machine learning algorithms. It is composed of 70000 28×28 grayscale fashion products, associated with 10 kinds of labels. It has the same structure of training and testing splits as MNIST. Fashion-MNIST is becoming popular, because it's more challenging than traditional MNIST and can better represent modern CV tasks.

4) COIL-100

The Coil-100 [70] is a dataset containing 7200 128×128 color images of 100 real-world objects. Each object is rotated by a turntable through 360 degrees and the image is taken at pose intervals of 5 degrees. So, this corresponds to only 72 images per object. The small data size gives great challenges to the proposed architecture.

B. QUANTITATIVE EVALUATION METHODOLOGY

For anomaly detection problem, methods generate a value (anomaly score) for each data point, indicating the degree of its abnormality. If the anomaly score is higher than a given threshold, it is identified as an abnormal sample. Otherwise, it is a normal sample. In combination with a threshold value, the testing samples are divided into four types: true positive (TP: abnormal samples are correctly detected), false negative (FN: abnormal samples are considered normal), true negative (TN: normal samples are correctly detected), and false positive (FP: normal samples are considered abnormal). In further, the true positive rate (TPR) and false positive rate (FPR) are defined as follows:

$$TPR = \frac{TP}{TP + FN} \tag{16}$$

$$FPR = \frac{FP}{FP + TN} \tag{17}$$

Given different threshold values, a receiver operating characteristic (ROC) curve plots all pairs of the true positive rate and the false positive rate. We utilize the area under the ROC curve (AUC) as performance measurement. The larger the values are, the better the model is. The two metrics are all between 0 and 1 and a perfect model has a value of 1.

$$F1 = \frac{2 \times PRECISION \times RECALL}{PRECISION + RECALL} \tag{18}$$

With respect to another metric, we adopt the F1-score. As shown in (18), it is the harmonic mean of precision and recall. The F1-score is related to the choice of thresholds, and we use the largest F1-score to evaluate the models. With a certain threshold, a perfect model provides 1 to both precision and recall, and thus gives F1-score equal to 1.

All values of the performance metrics take an average result of 20 trials. In each trial, training samples and testing samples are re-selected using different random seeds.

C. IMPLEMENTATION DETAILS

The input images are all resized to 32×32. The input images of other works mentioned in the experiments are all resized to 32×32 pixels. For anomaly score, the weights of SSIM loss and L1-norm loss are equal. The patch size K for SSIM is set as 11. The architecture of our approach is implemented based on the deep learning framework PyTorch [71]. Our network is optimized by Adam with a learning rate of 1×10^{-3}, momentums $\beta_1 = 0.50$, $\beta_2 = 0.999$. According to the numbers of normal samples in different datasets, the batch size is set as 64 for MNIST, CIFAR10, Fashion-MNIST while 1 for Coil-100. The size of latent code is set as 256. The weight values for SSIM loss and L1-norm loss are 50 and 10 respectively. The weight of mutual information loss is set as 10. The latent loss

and categorical loss separately have weight coefficients of 5 and 1 on CIFAR10, Fashion-MNIST and Coil-100 datasets. But for the simplest MNIST dataset, the weight of latent loss is 15 so as to increase the difficulty of reconstruction. For the training of WGAN-GP, the discriminator updates 5 times, and then the generator (CVAE) updates once. In each of the first four steps of the discriminator update, we successively use the L1-norm loss, SSIM loss, latent loss and mutual information loss to optimize the weights of CVAE. For a better and fast convergence of reconstruction, the L1-norm loss and the SSIM loss are used twice. The negative training process can be added when the latent loss is smaller than a threshold. Because when the latent loss is smaller, the latent codes of the normal samples are clustered well and the potential outliers can be easily and well generated. The categorical loss is optional, because the optimization of it means the normal classes are evenly distributed. The whole training process is shown in Algorithm 1.

Algorithm 1 Training Process of the Proposed Network

Input: A threshold T for active negative training

```
Initialize the network with random weights. count = 0
for each epoch do
    for each data batch do
        if count <= 3 then
            Optimize WGAN-GP discriminator by L_D.
            Optimize CVAE by SSIM and L1-norm loss.
            Optimize CVAE by latent loss.
            if latent loss < T then
                Conduct negative training.
            Optimize CVAE by mutual information loss.
            Optimize CVAE by SSIM and L1-norm loss.
        else if count <= 4 then
            Optimize WGAN-GP discriminator by L_D.
        else
            Optimize WGAN-GP generator by L_G.
        if count == 5 then
            count = 0
        else
            count = count+1
```

D. EXPERIMENTS IN MNIST AND CIFAR10 DATASETS

We compare our proposed CVAE-GAN-based anomaly detection (CVGAD) approach with the three recent state-of-the-art models [3], [25], [45] over two reference benchmarks MNIST [72] and CIFAR10 [68]. We follow the protocol as described in [25], [45]. Each of the ten categories is treated as an anomaly, while the rest of the categories are regarded as normal classes. For each dataset, 80% of normal samples are randomly sampled to constitute the training set. The testing set is composed of all abnormal samples and the rest normal samples.

Fig. 2 and Fig. 3 show the AUC results obtained on MNIST and CIFAR10 respectively. It can be seen that the proposed

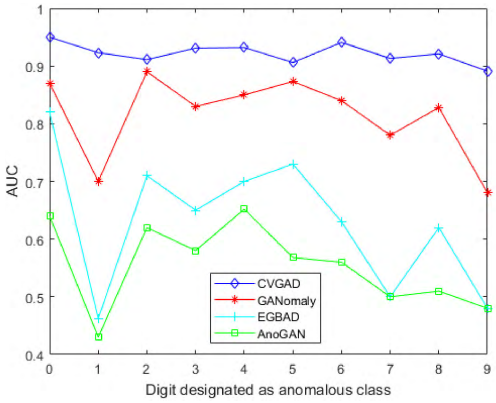

FIGURE 2. Experimental results of AUC performance in MNIST dataset.

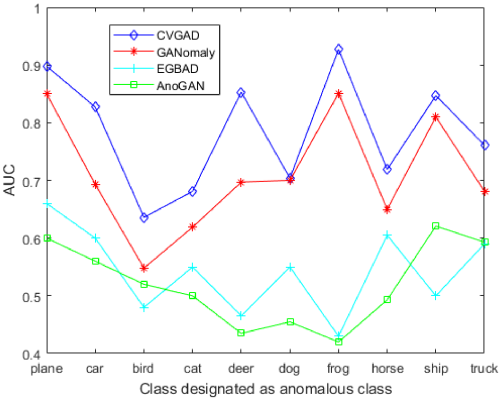

FIGURE 3. Experimental results of AUC performance in CIFAR10 dataset.

CVGAD is superior to other approaches and achieves the best AUC. In Fig. 2, compared with other digits, the digit 1 has a relatively lower AUC result for the models of [3], [25], [45]. The reason is that 1 has the simplest structure and the deep convolutional networks have a strong ability to learn extra information from other digits to reconstruct 1. In our framework, the reparameterization trick of CVAE in the training process provides variances in the latent space. The variances give some uncertainty to increase the reconstruction difficulty and finally make the trained model more robust to noises. In this way, CVAE prevents the model from easily reconstructing other objects not in the training set. Fig. 4 shows the original input and its reconstruction results. The designed CVAE module makes the abnormal digit 1 distinguishable. Therefore, our approach achieves better performance when treating 1 as an anomaly.

A challenge reflected in Fig. 3 is that there are relatively large differences between the AUC results in CIFAR10 dataset when designating different abnormal classes. One reason is that we adopt only one set of parameters to verify the model generality. The other one is the similarity between normal and abnormal objects. For example, the cat is similar to the dog. To distinguish the anomalies that are similar to normal samples, the architecture is required to learn more local details and reconstruct the fine-grained images. In [22], [73], they show that an added

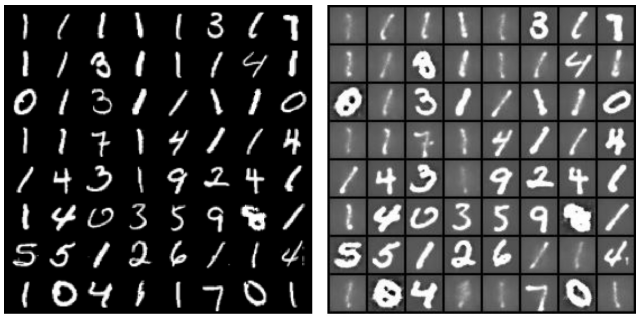

FIGURE 4. Experimental results in MNIST dataset. (a) Input samples in testing set. (b) Reconstruction results of (a) when designating digit 1 as an anomalous class.

adversarial module is capable of improving generative ability of decoded images. In our framework, the WGAN-GP is used for the adversarial training. It can further relieve the GAN problems of training instability and mode collapse. Besides, it can accelerate the convergence and reduce hyper-parameter tuning [74]. Traditional image reconstruction task by auto-encoder tends to use the L2-norm loss. However they bring blurriness to the decoded images. This may be because the pixel-wise L2-norm loss is too rigorous. The adopted SSIM value reflects image similarity from multiple perspectives and can measure the similarity degree smoothly and intuitively. Fig. 5 illustrates the reconstruction results of the proposed approach in CIFAR10 dataset. Fig. 5 (a) and Fig. 5 (b) show the normal testing samples and their corresponding decoded images. The blurriness is successfully reduced and more local details of images are recovered. Conversely, compared with abnormal testing samples in Fig. 5 (c), the reconstructed images in Fig. 5 (d) cannot accurately recover the texture details. Therefore, the accuracy improvement of the proposed detection approach attributes much to the fine-grained reconstruction.

E. ANALYSES OF LOSS TERMS

The latent loss takes advantage of hidden category information to regularize the latent space by multivariate Gaussian distributions with learnable means. It can generate a much better data manifold which is smooth, continuous and disentangled [47], [75]. That means the coding space is filled, exhibits no holes and can also reflect class information. The learned means enhance the relationship between the input and the manifold to potentially cluster the same kind of objects. We use the latent loss to regularize the latent space and use the SSIM, L1-norm and Wasserstein loss for reconstruction. Fig. 6 shows the clustering result. The three pictures represent three different categories clustered in latent space. The digits in each picture are sampled near the three clustered centers. It can be seen that the digits align their own corresponding means of different multivariate Gaussian distributions. In addition, based on the clustered means in Fig. 6, the potential abnormal samples can be generated reasonably in the

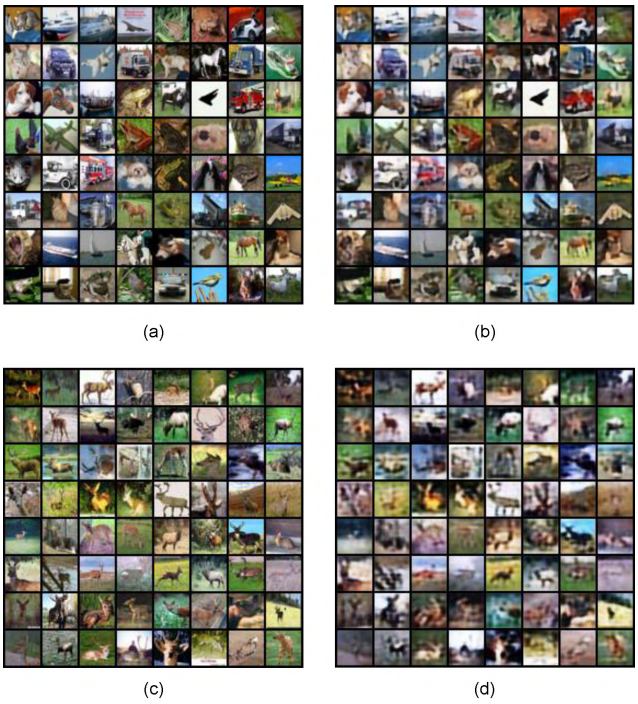

FIGURE 5. Experimental results in CIFAR10 dataset. (a) Input samples of normal classes in testing set. Deers are designated as anomalous class. (b) Reconstruction results of (a). (c) Input samples of abnormal class deer in testing set. (d) Reconstruction results of (c).

FIGURE 6. In the latent space, we randomly select three different categories. The digits in the three pictures are sampled near the three clustered centers respectively.

latent space for the negative training process. It provides a good assistant for the semi-supervised learning task.

In order to distinguish anomalies, the encoded manifold needs to well represent the characteristics of normal input data. Maximizing the mutual information can progressively increase the relationship between the latent codes and the inputs. We use the mutual information loss to make the manifold informative and use the SSIM, L1-norm and Wasserstein loss for reconstruction. Fig. 7 shows some relationship established by the mutual information loss. Fig. 7 (b) exhibits images that are close to the images of Fig. 7 (a) in latent space. The images of Fig. 7 (a) and (b) are similar in colors and structures. In contrast, the images in Fig. 7 (c) look different from Fig.7 (a) and they are further apart from each other in latent space.

In order to show the performance enhancement caused by each of these multiple loss terms, we reduce one of the losses and repeat the experiments in CIFAR10 dataset. We designate the deer as an anomaly and the rest categories

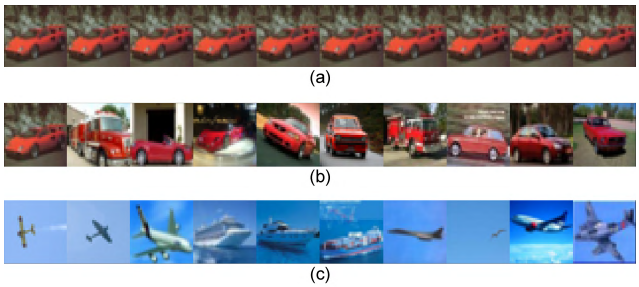

FIGURE 7. (a) shows a randomly selected testing sample. (b) shows 10 testing samples closest to (a) in latent space. (c) shows 10 testing samples furthest from (a) in latent space. The Euclidean distance is used as the distance metric.

TABLE 1. The impact of each loss on performance.

	performance index	treat deer as an anomaly
without mutual information loss	F1	0.781
	AUC	0.794
without latent loss	F1	0.785
	AUC	0.790
without SSIM loss	F1	0.722
	AUC	0.736
without L1-norm loss	F1	0.751
	AUC	0.767
without Wasserstein loss	F1	0.705
	AUC	0.718
without negative training process	F1	0.821
	AUC	0.838
use all losses	F1	0.845
	AUC	0.853

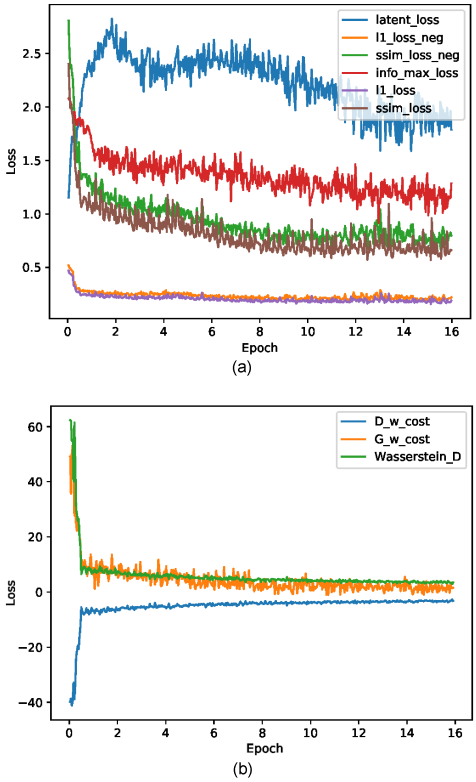

FIGURE 8. (a) Training losses of CVAE. (b) Training losses of WGAN-GP. D_w_cost represents the loss cost of discriminator weights. G_w_cost denotes the loss cost of generator weights. Wasserstein_D is the Wasserstein distance.

as normal classes. The results are shown in Table 1. The Wasserstein loss has the greatest impact on performance. It may attribute to the adversarial training style. Without the latent loss, the performance is degraded. That may be because the latent loss can improve the generalization ability of the model. Besides, the negative training process can't work well without the clustering effect of the latent loss. Compared with the L1-norm loss, the SSIM loss has a greater impact on performance. It may be because the network can learn more features of normal images by virtue of the SSIM loss. In addition, the mutual information loss can improve the accuracy of anomaly detection and the negative training process has a minimal effect on performance enhancement.

F. TRAINING CONVERGENCE

Fig. 8 shows the curves of training losses. The SSIM and L1-norm losses are decreased, making the manifold learn more reconstruction-based data characteristics. Minimizing the latent loss reduces the distance between the data representations and the multivariate Gaussian distributions, forcing the latent space to be smooth and disentangled. The reparameterization trick in the latent loss, provides noises to increase the generalization ability of the model. The latent loss and the reconstruction loss restrict each other. They form a hidden adversarial training style which is reflected by the curve jitter in Fig. 8 (a). The hidden adversarial training can endow the model with better performance. Fig. 8 (a) also illustrates the negative training process which increases the reconstruction (SSIM and L1-norm) losses for potential outliers. In addition, Fig. 8 (b) shows the training curves of WGAN-GP. The Wasserstein loss gradually decreases to reduce the distance between the distribution of generated samples and normal samples. All losses decrease and the converged model achieves an overall high likelihood.

G. EXPERIMENTS IN FASHION-MNIST DATASET

In this section, we evaluate the proposed architecture in FASHION-MNIST dataset. Different from the last experiment, the normal inputs in this experiment are images of one class but not nine classes. A similar setup in [23] is applied. The anomalies are randomly selected from the other classes. Besides, the 5-fold cross-validation is adopted with each fold taking 20% of each class. The partition ratio of the training set, validation set and testing set is 6:2:2.

Fig. 9 shows the normal and abnormal testing samples and their corresponding decoded images. In Fig. 9(a) and Fig. 9(b), the vertical linear structures of normal class trousers are captured by networks. Therefore, the anomalies circled in yellow are reconstructed like trousers, leading to a large reconstruction error. Because the shoes circled in red have a lateral structure, the network responds very little to them. In Fig. 9(c) and Fig. 9(d), the normal class sneaker circled in green is well reconstructed. The sandal circled in red and

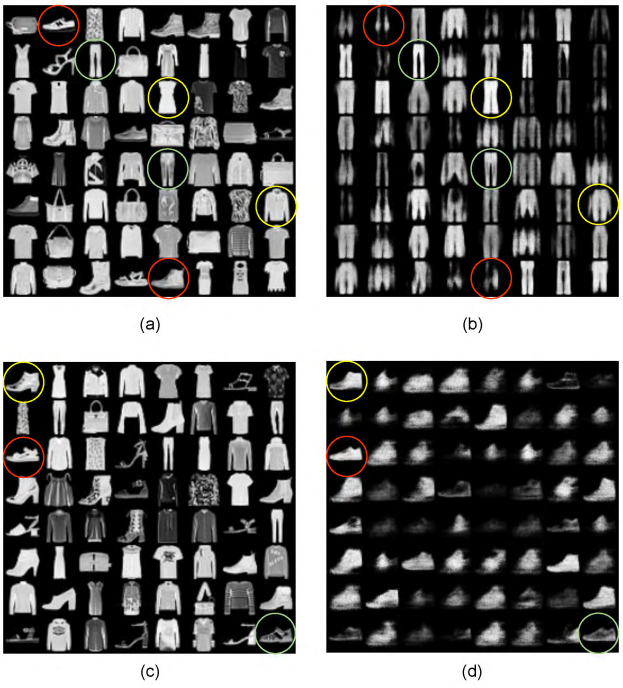

FIGURE 9. Experimental results in Fashion-MNIST dataset. (a) Input samples of testing set. Trousers are designated as normal class. (b) Reconstruction results of (a). (c) Input samples of testing set. Sneakers are designated as normal class. (d) Reconstruction results of (c). In (a), (b), (c), (d), the green circles denote normal classes while the red and yellow circles represent anomalous classes.

TABLE 2. Comparison results in fashion-MNIST dataset.

	% of outliers	10	20	30	40	50
GPND	F1	0.968	0.945	0.917	0.891	0.864
	AUC	0.928	0.932	0.933	0.933	0.933
CVGAD	F1	0.946	0.931	0.923	0.895	0.890
	AUC	0.921	0.917	0.940	0.946	0.937

the ankle boot circled in yellow are anomalies which are similar to the sneaker. However, they are reconstructed like the sneaker without the particular hollow structure and shoe heel, thus resulting in a high anomaly score.

We compare our methods with the state-of-the-art GPND [23]. Table 2 shows the performance indexes of the GPND and the proposed CVGAD. Very different from the CVGAD, GPND imposes a standard Normal distribution on latent space, uses auto-encoder loss, traditional GAN training and detects anomalies by calculating a data probability density function. The CVGAD achieves a better result when the percentage of anomalies is high. It performs better in abnormal samples. That may be because it takes into account the negative training and can effectively use SSIM loss to distinguish abnormal local structural details as illustrated in Fig. 9 (c) and Fig. 9 (d).

H. EXPERIMENTS IN COIL-100 DATASET

In this experiment, we evaluate the performance of the proposed approach in Coil-100 dataset. The Coil-100 dataset has

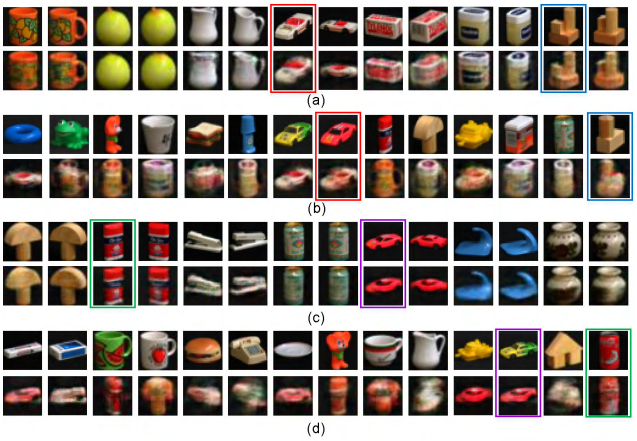

FIGURE 10. Experimental results in Coil-100 dataset. (a) and (b) are testing samples of an experiment which takes randomly 7 categories as normal classes. (c) and (d) belong to another experiment with the same setup. (a) and (c) are normal inputs and corresponding decoded images. (b) and (d) are anomalies and corresponding decoded images. The rectangles in the same colors denote similar normal and abnormal samples.

only 72 images per object. We follow the similar protocol in [14]. One, four and seven categories are randomly chosen as normal objects respectively. Due to the few samples in each kind of object, validation process is not necessary and 70 pictures are used for training while 2 for testing. Besides, a maximum of one sample is selected from each of the remaining categories as an anomaly. Fig. 10 (a) and Fig. 10 (c) show the testing samples and their reconstructions for seven different kinds of normal objects. In contrast, Fig. 10 (b) and Fig. 10 (d) are for the corresponding abnormal objects. The pictures boxed up in red and purple show that the reconstructed abnormal objects are similar to the normal objects. Moreover, the pictures boxed up in blue and green show that the abnormal objects, which are very similar to the normal objects, lose their clear textures and are blurred in their reconstructions.

Table 3 shows the quantitative results of different methods in Coil-100 dataset. Most of the numbers are borrowed from [14]. Compared with the l_1-thresholding, although the AUC of the proposed CVGAD is lower than the l_1-thresholding, the CVGAD has higher F1 scores. In particular, when the number of categories increases, the F1 score of the l_1-thresholding decreases rapidly while the F1 score of the CVGAD decreases a little. It may attribute to that the CVGAD enhances the relationship between the inputs and the latent codes and considers the number of categories. This can be observed from Fig. 10 (a) and Fig. 10 (c). Although there are seven different objects in Fig. 10(a) and Fig. 10(c), the CVGAD can accurately recover each of them without being influenced by other objects. In addition, the CVGAD is only trained from random weights with the limited 70 samples per class. The small data size can influence our training results.

We don't compare the proposed CVGAD with the R-graph [14]. Because, a VGG [81] network pretrained on

TABLE 3. Comparison results in coil-100 dataset.

	OutRank [76], [77]	CoP [40]	REAPER [29]	Outlier-Pursuit [78]	LRR [79]	DPCP [80]	l_1 thres-holding [15]	R-graph [14]	GPND [23]	CVGAD
Inliers: one category of images. Outliers: 50%										
AUC	0.836	0.843	0.900	0.908	0.847	0.900	0.991	0.997	0.968	0.973
F1	0.862	0.866	0.892	0.902	0.872	0.882	0.978	0.990	0.979	0.983
Inliers: four categories of images. Outliers: 25%										
AUC	0.613	0.628	0.877	0.837	0.687	0.859	0.992	0.996	0.945	0.949
F1	0.491	0.500	0.703	0.686	0.541	0.684	0.941	0.970	0.960	0.952
Inliers: seven categories of images. Outliers: 15%										
AUC	0.570	0.580	0.824	0.822	0.628	0.804	0.991	0.996	0.919	0.938
F1	0.342	0.346	0.541	0.528	0.366	0.511	0.897	0.995	0.941	0.945

ImageNet [82] is used in the R-graph. Besides, the R-graph has more complex network structures than the CVGAD.

The results of the CVGAD have similar performance with the GPND and the l_1-thresholding [15]. The three approaches are all based on the data self-expression. It seems that the self-expression is more suitable and powerful for the Coil-100 dataset.

IV. CONCLUSION

In conclusion, we propose a novel CVAE-GAN-BASED end-to-end framework to solve the semi-supervised anomaly detection problem. Our deep convolutional networks are only trained by normal images which are not classified and may have different categories. Our proposed framework takes advantage of the adversarial training ideas both in CVAE and GAN to control the latent encoded manifold and to provide the high-quality reconstructions. Based on the manifold, the active negative training can add extra anomaly information to the network training. The maximization of mutual information further improves the learning ability for normal samples. The SSIM and Wasserstein loss have a strong ability to ensure the convergence of the CVAE and GAN. In experiments, the normal samples are designed with large data size, small data size, several categories and one category respectively. All the experiments achieve satisfactory results. Compared with the state-of-the-art works on public dataset, our approach improves the detection accuracy without very deep and complex architectures. So, our work shows a great perspective to learn the intrinsic nature of normal data to distinguish anomalies.

In the future, the performance can be further improved by designing more anomaly metrics, such as manifold-based anomaly scores.

REFERENCES

[1] S. Liu, H. Wu, Y. Huang, Y. Yang, and J. Jia, "Accelerated structure-aware sparse Bayesian learning for 3D electrical impedance tomography," *IEEE Trans. Ind. Informat.*, to be published.

[2] H. Yetis and M. Karakose, "Adaptive vision based condition monitoring and fault detection method for multi robots at production lines in industrial systems," *Int. J. Appl. Math., Electron. Comput.*, vol. 4, no. 1, p. 271, 2016.

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

[4] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab, "Deep autoencoding models for unsupervised anomaly segmentation in brain MR images," in *Proc. Int. MICCAI Brainlesion Workshop*. Cham, Switzerland: Springer, 2018, pp. 161–169.

[5] S. Liu, J. Jia, Y. D. Zhang, and Y. Yang, "Image reconstruction in electrical impedance tomography based on structure-aware sparse Bayesian learning," *IEEE Trans. Med. Imag.*, vol. 37, no. 9, pp. 2090–2102, Sep. 2018.

[6] M. A. F. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal Process.*, vol. 99, pp. 215–249, Jun. 2014.

[7] S. S. Khan and M. G. Madden, "A survey of recent trends in one class classification," in *Proc. Irish Conf. Artif. Intell. Cognit. Sci.* Berlin, Germany: Springer, 2009, pp. 188–197.

[8] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, vol. 589. Hoboken, NJ, USA: Wiley, 2005.

[9] M. Sabokrou, M. Pourreza, M. Fayyaz, R. Entezari, M. Fathy, J. Gall, and E. Adeli, "AVID: Adversarial visual irregularity detection," 2018, *arXiv:1805.09521*. [Online]. Available: https://arxiv.org/abs/1805.09521

[10] M. Ahmed, A. N. Mahmood, and J. Hu, "A survey of network anomaly detection techniques," *J. Netw. Comput. Appl.*, vol. 60, pp. 19–31, Jan. 2016.

[11] B. R. Kiran, D. M. Thomas, and R. Parakkal, "An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos," *J. Imag.*, vol. 4, no. 2, p. 36, 2018.

[12] L. Akoglu, H. Tong, and D. Koutra, "Graph based anomaly detection and description: A survey," *Data Mining Knowl. Discovery*, vol. 29, no. 3, pp. 626–688, 2015.

[13] S. Ranshous, S. Shen, D. Koutra, S. Harenberg, C. Faloutsos, and N. F. Samatova, "Anomaly detection in dynamic networks: A survey," *Wiley Interdiscipl. Rev., Comput. Stat.*, vol. 7, no. 3, pp. 223–247, 2015.

[14] C. You, D. P. Robinson, and R. Vidal, "Provable self-representation based outlier detection in a union of subspaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3395–3404.

[15] M. Soltanolkotabi and E. J. Candés, "A geometric analysis of subspace clustering with outliers," *Ann. Statist.*, vol. 40, no. 4, pp. 2195–2238, 2012.

[16] Y. Xia, X. Cao, F. Wen, G. Hua, and J. Sun, "Learning discriminative reconstructions for unsupervised outlier removal," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1511–1519.

[17] M. Sabokrou, M. Fathy, and M. Hoseini, "Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder," *Electron. Lett.*, vol. 52, no. 13, pp. 1122–1124, 2016.

[18] M. Markou and S. Singh, "Novelty detection: A review—part 1: Statistical approaches," *Signal Process.*, vol. 83, no. 12, pp. 2481–2497, 2003.

[19] J. Kim and C. D. Scott, "Robust kernel density estimation," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 2529–2565, Jan. 2012.

[20] K. Yamanishi, J.-I. Takeuchi, G. Williams, and P. Milne, "On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms," *Data Mining Knowl. Discovery*, vol. 8, no. 3, pp. 275–300, 2004.

[21] V. Barnett and T. Lewis, *Outliers in Statistical Data*. Hoboken, NJ, USA: Wiley, 1974.

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

[23] S. Pidhorskyi, R. Almohsen, and G. Doretto, "Generative probabilistic novelty detection with adversarial autoencoders," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 6823–6834.

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

[25] H. Zenati, C. S. Foo, B. Lecouat, G. Manek, and V. R. Chandrasekhar, "Efficient GAN-based anomaly detection," 2018, *arXiv:1802.06222*. [Online]. Available: https://arxiv.org/abs/1802.06222

[26] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1446–1453.

[27] M. Bertini, A. D. Bimbo, and L. Seidenari, "Multi-scale and real-time non-parametric approach for anomaly detection and localization," *Comput. Vis. Image Understand.*, vol. 116, no. 3, pp. 320–329, 2012.

[28] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1975–1981.

[29] G. Lerman, M. B. McCoy, J. A. Tropp, and T. Zhang, "Robust computation of linear models by convex relaxation," *Found. Comput. Math.*, vol. 15, no. 2, pp. 363–410, Apr. 2015.

[30] M. Sabokrou, M. Fathy, M. Hoseini, and R. Klette, "Real-time anomaly detection and localization in crowded scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2015, pp. 56–62.

[31] M. Sabokrou, M. Fayyaz, M. Fathy, and R. Klette, "Deep-cascade: Cascading 3D deep neural networks for fast anomaly detection and localization in crowded scenes," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1992–2004, Apr. 2017.

[32] Y. Feng, Y. Yuan, and X. Lu, "Learning deep event models for crowd anomaly detection," *Neurocomputing*, vol. 219, pp. 548–556, Jan. 2017.

[33] Z. Fang, F. Fei, Y. Fang, C. Lee, N. Xiong, L. Shu, and S. Chen, "Abnormal event detection in crowded scenes based on deep learning," *Multimedia Tools Appl.*, vol. 75, no. 22, pp. 14617–14639, 2016.

[34] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *Proc. CVPR*, Jun. 2011, pp. 3449–3456.

[35] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe, "Learning deep representations of appearance and motion for anomalous event detection," 2015, *arXiv:1510.01553*. [Online]. Available: https://arxiv.org/abs/1510.01553

[36] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 733–742.

[37] E. M. Knorr, R. T. Ng, and V. Tucakov, "Distance-based outliers: Algorithms and applications," *Very Large Data Bases J.*, vol. 8, nos. 3–4, pp. 237–253, 2000.

[38] V. Hautamaki, I. Karkkainen, and P. Franti, "Outlier detection using k-nearest neighbour graph," in *Proc. 17th Int. Conf. Pattern Recognit. (ICPR)*, vol. 3, Aug. 2004, pp. 430–433.

[39] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," *ACM SIGMOD Rec.*, vol. 29, no. 2, pp. 93–104, 2000.

[40] M. Rahmani and G. K. Atia, "Coherence pursuit: Fast, simple, and robust principal component analysis," *IEEE Trans. Signal Process.*, vol. 65, no. 23, pp. 6260–6275, Dec. 2017.

[41] P. Bodesheim, A. Freytag, E. Rodner, M. Kemmler, and J. Denzler, "Kernel null space methods for novelty detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3374–3381.

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

[43] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2234–2242.

[44] J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial feature learning," 2016, *arXiv:1605.09782*. [Online]. Available: https://arxiv.org/abs/1605.09782

[45] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "GANomaly: Semi-supervised anomaly detection via adversarial training," 2018, *arXiv:1805.06725*. [Online]. Available: https://arxiv.org/abs/1805.06725

[46] I. Goodfellow, "NIPS 2016 tutorial: Generative adversarial networks," 2016, *arXiv:1701.00160*. [Online]. Available: https://arxiv.org/abs/1701.00160

[47] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*. [Online]. Available: https://arxiv.org/abs/1312.6114

[48] D. Bouchacourt, R. Tomioka, and S. Nowozin, "Multi-level variational autoencoder: Learning disentangled representations from grouped observations," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 2095–2102.

[49] C. Doersch, "Tutorial on variational autoencoders," 2016, *arXiv:1606.05908*. [Online]. Available: https://arxiv.org/abs/1606.05908

[50] A. Munawar, P. Vinayavekhin, and G. De Magistris, "Limiting the reconstruction capability of generative neural network using negative learning," in *Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2017, pp. 1–6.

[51] M. Kimura and T. Yanagihara, "Anomaly detection using GANs for visual inspection in noisy training data," 2018, *arXiv:1807.01136*. [Online]. Available: https://arxiv.org/abs/1807.01136

[52] K. Gray, D. Smolyak, S. Badirli, and G. Mohler, "Coupled IGMM-GANs for deep multimodal anomaly detection in human mobility data," 2018, *arXiv:1809.02728*. [Online]. Available: https://arxiv.org/abs/1809.02728

[53] C. E. Rasmussen, "The infinite Gaussian mixture model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2000, pp. 554–560.

[54] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," 2015, *arXiv:1511.05644*. [Online]. Available: https://arxiv.org/abs/1511.05644

[55] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 1096–1103.

[56] D. J. Rezende, S. Mohamed, I. Danihelka, K. Gregor, and D. Wierstra, "One-shot generalization in deep generative models," 2016, *arXiv:1603.05106*. [Online]. Available: https://arxiv.org/abs/1603.05106

[57] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," 2018, *arXiv:1808.06670*. [Online]. Available: https://arxiv.org/abs/1808.06670

[58] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

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

[60] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," 2017, *arXiv:1701.07875*. [Online]. Available: https://arxiv.org/abs/1701.07875

[61] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," 2014, *arXiv:1401.4082*. [Online]. Available: https://arxiv.org/abs/1401.4082

[62] A. Makhzani and B. J. Frey, "Pixelgan autoencoders," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1975–1985.

[63] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," 2015, *arXiv:1511.06349*. [Online]. Available: https://arxiv.org/abs/1511.06349

[64] X. Chen, D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel, "Variational lossy autoencoder," 2016, *arXiv:1611.02731*. [Online]. Available: https://arxiv.org/abs/1611.02731

[65] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," 2017, *arXiv:1701.04862*. [Online]. Available: https://arxiv.org/abs/1701.04862

[66] I. Steinwart, D. Hush, and C. Scovel, "A classification framework for anomaly detection," *J. Mach. Learn. Res.*, vol. 6, pp. 211–232, Jun. 2005.

[67] Y. LeCun and C. Cortes. (2010). *MNIST Handwritten Digit Database*. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[68] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Tech. Rep., 2009.

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

[70] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia object image library (coil-20)," Dept. Comput. Sci., Columbia Univ., New York, NY, USA, Tech. Rep. CUCS-006-96, 1996.

[71] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *Proc. 31st Conf. Neural Inf. Process. Syst.*, 2017, pp. 1–4.

[72] 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.

[73] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," 2015, *arXiv:1512.09300*. [Online]. Available: https://arxiv.org/abs/1512.09300

[74] L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for GANs do actually converge?" 2018, *arXiv:1801.04406*. [Online]. Available: https://arxiv.org/abs/1801.04406

[75] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in β-VAE," 2018, *arXiv:1804.03599*. [Online]. Available: https://arxiv.org/abs/1804.03599
[76] H. D. K. Moonesignhe and P.-N. Tan, "Outlier detection using random walks," in *Proc. 18th IEEE Int. Conf. Tools Artif. Intell. (ICTAI)*, Nov. 2006, pp. 532–539.
[77] H. D. K. Moonesinghe and P.-N. Tan, "Outrank: A graph-based outlier detection framework using random walk," *Int. J. Artif. Intell. Tools*, vol. 17, no. 1, pp. 19–36, 2008,
[78] H. Xu, C. Caramanis, and S. Sanghavi, "Robust PCA via outlier pursuit," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 2496–2504.
[79] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 663–670.
[80] M. C. Tsakiris and R. Vidal, "Dual principal component pursuit," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2015, pp. 10–18.
[81] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556
[82] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

JIANG BIAN received the B.Sc. degree in information science and engineering from Central South University, Changsha, China, in 2015. He is currently pursuing the Ph.D. degree with the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

His current research interests include computer vision, pattern recognition, SLAM, and UAV applications.

XIAOLONG HUI received the B.Sc. and M.Sc. degrees in mechanical and electrical control engineering from Beijing Jiaotong University, Beijing, China, in 2012 and 2015, respectively. He is currently pursuing the Ph.D. degree with the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing.

His current research interests include mechanical design, electronics, robotic navigation, and robotic control.

SHIYING SUN received the B.Sc. degree in control science and engineering from Central South University, Hunan, China, in 2013, and the Ph.D. degree in control science and engineering from the Institute of Automation, Chinese Academy of Sciences (IACAS), Beijing, China, in 2019, where he holds a postdoctoral position with the State Key Laboratory of Management and Control for Complex Systems.

His current research interests include advanced robot control, navigation, and computer vision.

XIAOGUANG ZHAO received the B.Sc. degree in control engineering from the Shenyang University of Technology, Shenyang, China, in 1992, and the M.Sc. and Ph.D. degrees in control theory and control engineering from the Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, in 1998 and 2001, respectively. She is currently a Professor with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

Her current research interests include advanced robot control, wireless sensor networks, and robot vision.

MIN TAN received the B.Sc. degree in control engineering from Tsinghua University, Beijing, China, in 1986, and the Ph.D. degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 1990, where he is currently a Professor with the State Key Laboratory of Management and Control for Complex Systems.

His research interests include advanced robot control, multirobot systems, biomimetic robots, and manufacturing systems.