Novelty Detection via Contrastive Learning with Negative Data Augmentation

Chengwei Chen[1*], Yuan Xie[1*], Shaohui Lin[1†], Ruizhi Qiao[2], Jian Zhou[2], Xin Tan[3], Yi Zhang[4] and Lizhuang Ma[1†]

[1]East China Normal University
[2]Tencent Youtu Lab
[3]Shanghai Jiao Tong University
[4]Zhejiang Lab

52184501028@stu.ecnu.edu.cn, {yxie,shlin}@cs.ecnu.edu.cn, {ruizhiqiao,darnellzhou}@tencent.com, tanxin2017@sjtu.edu.cn, zhangyi620@zhejianglab.com, lzma@cs.ecnu.edu.cn

Abstract

Novelty detection is the process of determining whether a query example differs from the learned training distribution. Previous generative adversarial networks based methods and self-supervised approaches suffer from instability training, mode dropping, and low discriminative ability. We overcome such problems by introducing a novel decoder-encoder framework. Firstly, a generative network (*a.k.a.* decoder) learns the representation by mapping the initialized latent vector to an image. In particular, this vector is initialized by considering the entire distribution of training data to avoid the problem of mode-dropping. Secondly, a contrastive network (*a.k.a.* encoder) aims to "learn to compare" through mutual information estimation, which directly helps the generative network to obtain a more discriminative representation by using a negative data augmentation strategy. Extensive experiments show that our model has significant superiority over cutting-edge novelty detectors and achieves new state-of-the-art results on various novelty detection benchmarks, *e.g.* CIFAR10 and DCASE. Moreover, our model is more stable for training in a non-adversarial manner, compared to other adversarial based novelty detection methods.

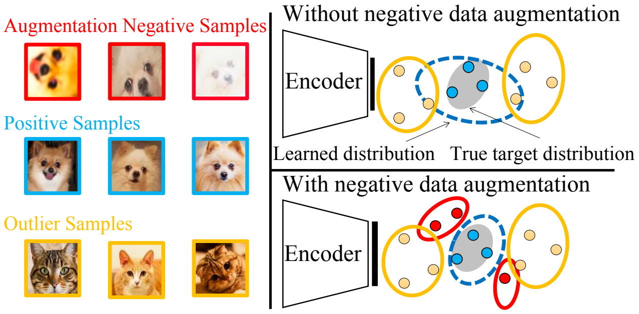

Augmentation Negative Samples
Positive Samples
Outlier Samples
Without negative data augmentation
Encoder
Learned distribution
True target distribution
With negative data augmentation
Encoder

Figure 1: Overview of negative data augmentation strategy. **Top**: without the negative data augmentation strategy, the learned distribution (blue oval) of target samples may include the outliers (yellow dots). **Bottom**: The novelty-like samples (red dots) are generated from target samples by negative data augmentation where we employ multiple transformations combination. Compared with outliers, the distribution of novelty-like samples (red oval) is more close to that of the target samples. Through mutual information estimation, the learned distribution of target class can be pulled to close to the true target distribution (gray oval), and pushed to far away from the distribution of hard novelty-like samples.

1 Introduction

Novelty detection can be described as a one-class classification, which aims to detect the samples whether drawing far away from the learned distribution of training samples from the target class. Generative adversarial networks (GANs) [Sabokrou *et al.*, 2018; Perera *et al.*, 2019; Chen *et al.*, 2020a] have been a common choice for novelty detection. Generator and discriminator compete mutually while collaborating to learn a representative latent space for the target class. In this latent space, the reconstructed features of novelty samples (outliers) have higher reconstruction errors than normal samples (inliers), which can be used for distinguishing between normal and novelty classes. Recently, self-supervised learning [Komodakis and Gidaris, 2018; Ji *et al.*, 2019] holds great for improving representations when labeled data are scarce. In the training process, the network learns useful feature representation by solving some specialized pretext tasks, such as geometric transformations prediction [Hendrycks *et al.*, 2019]. During inference, the model is transferred to the downstream task, like novelty detection.

However, some weaknesses are still existing in the previous work. For the GAN based methods, it suffers from three critical problems: mode-dropping, instable training, and low discriminative ability. First, various GAN based methods (*e.g.* ALOCC [Sabokrou *et al.*, 2018], OCGAN [Perera *et al.*, 2019] and DualGAN [Chen *et al.*, 2020a]), only learn the partial modes of target distributions, which causes the problem of mode dropping [Arora *et al.*, 2018]. Second, the imbalance capacity of generator and discriminator causes the training of model unstable [Zhao *et al.*, 2017], which affects the learning

[*]These authors contributed equally to this work.
[†]Corresponding authors.

of latent representation of normal samples. Third, the latent features with low discriminative ability [Liu *et al.*, 2020] are generated in self-representation, since the decoder of GANs tends to learn more structive representation than discriminative characteristics.

The existing self-supervised learning based novelty detection requires specialized implementation for the pretext tasks, such as design supervised labels, loss functions, and network architectures. Rotation prediction [Komodakis and Gidaris, 2018] and transformations prediction [Hendrycks *et al.*, 2019] could capture the semantic information of object shapes that are useful for target tasks. However, it lacks other semantic information such as object textures and colors, which leads to the low discriminative ability of features that fails to effectively detect novelty samples. Besides, some [Lim *et al.*, 2018; Sinha *et al.*, 2021] use data augmentation as an additional source of data in the GAN. Inspired by these works, contrasting shifted instance (CSI) [Tack *et al.*, 2020] contrasts distributionally-shifted augmentations with an auxiliary softmax classifier to learn the feature representation of encoder based on SimCLR [Chen *et al.*, 2020b]. However, the augmented images as negative samples only use one random augmentation, which cannot generate effective negative samples. Beside, the detection score function of CSI is complex with high computation and memory cost.

To address these issues, we propose a novel decoder-encoder framework for one-class novelty detection to learn more discriminative latent representation by *contrastive learning*. Our framework consists of three parts: a generative network, a contrastive network, and a mutual information estimator. First, the generative network (decoder) aims to learn the representation of target class by mapping each initialized latent vector to each target image; The initialization of the latent vectors is obtained by encoding the entire distribution of training data to alleviate the problem of mode-dropping. Then, we employ the contrastive network (encoder) to extract both local feature maps from positive and negative samples. It also captures the global latent features from positive samples. In particular, we select the normal training data as positive samples and use negative data augmentation strategy to generate hard negative samples (*a.k.a.* novelty-like samples) from their corresponding positive samples by using multiple random transformations. Different from outliers, these novelty-like samples are more close to the normal ones. Therefore, it can better help to separate the distribution of normal and novelty samples by learning the disjointness between positive and novelty-like samples, as illustrated in Fig.1. Finally, mutual information estimator is adopted to generate discriminative latent features through contrastive learning on the features of input pairs by maximizing the local, global, and prior mutual information. Our decoder-encoder framework is presented in Fig. 2.

We summarize our main contributions of this paper:

- We propose a novel effective decoder-encoder framework for the novelty detection task. Contrastive learning is introduced to learn more discriminative latent features for distinguishing between positive and negative augmentation samples.
- The mutual information estimator is trained in a non-adversarial way by distinguishing between features of local parts and global context constituted by only positive and novelty-like samples, which helps the model to train more stably with faster convergence than GAN-based methods.
- Extensive experiments demonstrate the superior performance of our approach for novelty detection in several challenging datasets. For instance, our method achieves the highest mean AUC of 0.843 and 0.899 on CIFAR-10 and DCASE, compared to state-of-the-art methods.

2 Methodology

2.1 Our Decoder-encoder Framework

Our framework consists of three components: a generative network, a contrastive network and a mutual information estimator, as it shown in Fig. 2. The generative network learns a mapping from latent space to high-dimensional image space, while the contrastive network maps a positive/negative image to local feature maps and a global latent feature. The mutual information estimator is used for distinguishing between the features from the target samples and their corresponding hard negative samples to effectively learn more discriminative latent feature presentation.

Notations. Let $X = \{x_1, \cdots, x_N\}$ denote the original input images of target class with N samples, $Z = \{z_1, \cdots, z_N\}$ denote their corresponding global latent features where $z_i = \phi_{\theta_c}(x_i) \in \mathbb{R}^d$ is learned by the contrastive network $\mathbb{C}$ with parameters θ_c. d is the dimension of latent features. $A_i = \phi_{\theta_{fp}}(x_i) \in \mathbb{R}^{H \times W \times S}$ is the local feature maps extracted from contrastive network $\mathbb{C}$, where $\theta_{fp} \subset \theta_c$ and $H \times W \times S$ is the dimension of feature maps. $x'_i = \psi_{\theta_g}(z'_i)$ presents the reconstructed image by generative network $\mathbb{G}$ with parameters θ_g, where z'_i is the initialized latent vector. In mutual information estimator, we denote the global estimator, local estimator and prior estimator as $\mathbb{GE}, \mathbb{LE}$ and $\mathbb{PE}$ with parameters θ_{ge}, θ_{le} and θ_{pe}, respectively.

Generative network. The input latent vectors $z'_1, , \cdots, z'_N$ of generative network are first initialized by the PCA projection of all input images, which helps to alleviate the problem of mode dropping. Like a decoder, the generative network outputs $\psi_{\theta_g}(z'_i)$ should be regained close to the input image x_i. Instead of MSE loss, we employ the Laplacian pyramid loss [Ling and Okada, 2006] as the reconstruction loss; MSE is easy to yield the blurry image, while the Laplacian pyramid loss can generate better reconstructed images by Laplacian pyramid representation. Therefore, the reconstruction loss between the reconstructed output $\psi_{\theta_g}(z'_i)$ and the input image x_i is formulated as:

$$L_{lap} = \sum_j 2^{2j} \left| Lap^j(\psi_{\theta_g}(z'_i)) - Lap^j\,(x_i) \right|_1, \qquad (1)$$

where $Lap^j(\cdot)$ is the j-th level of Laplacian pyramid representation.

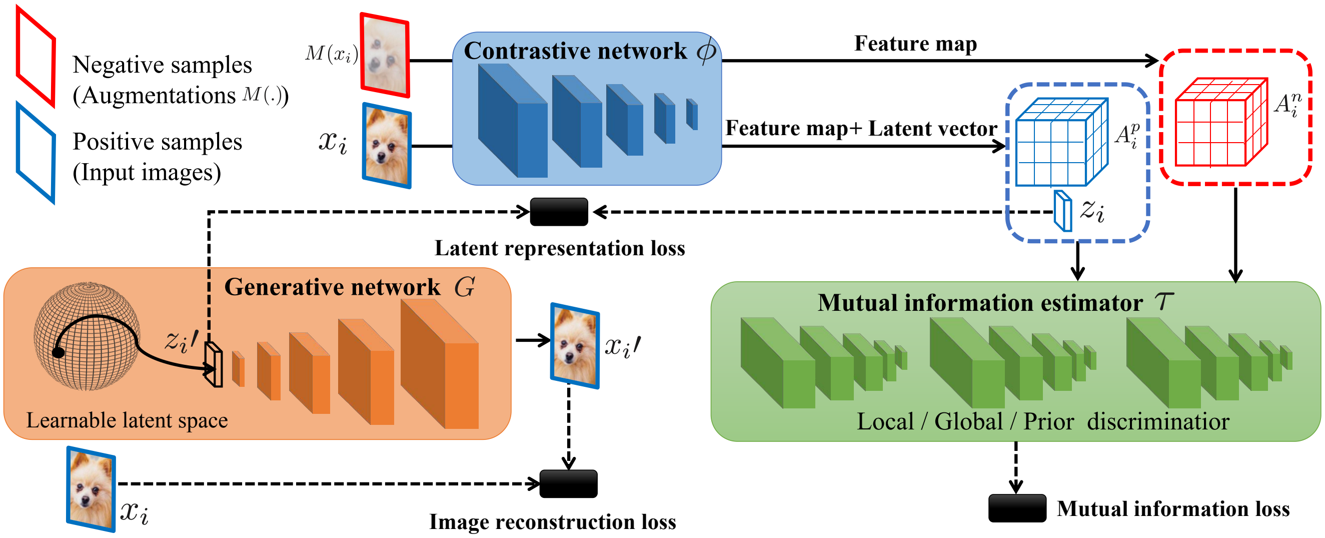


Figure 2: Illustration of our decoder-encoder framework for novelty detection. It consists of three components: a generative network, a contrastive network, and a mutual information estimator. The generative network (decoder) learns a reconstructed image representation by mapping each initialized latent vector to each target image. The contrastive network (encoder) encodes a pair of target and novelty-like samples from data augmentation to extract their latent global features and local feature maps. The mutual information estimator is adopted to generate discriminative latent features through contrastive learning on the features of input pairs by maximizing the local, global and prior mutual information.

Contrastive network. To improve the discriminative ability of the latent features of the target class, the contrastive network uses a pair of positive and negative samples as an input to extract both local feature maps (A_i^p, A_i^n) and a global latent vector z_i of the positive sample, where the samples from different classes compete each other through mutual information estimator. As shown in Fig. 2 (blue box), a positive samples x_i is an original image and its negative sample is generated by the negative data augmentation M as $M(x_i)$ (called novelty-like samples). Our transformations of augmentation include random resized crop, random color jitter, random grayscale, and random horizontal flip. We employ the combination of multiple random transformations to capture more rich semantic information, including object shape, textures, and colors. In addition, this augmentation can better help to separate the distribution of normal and novelty samples by learning the disjointness between positive and novelty-like samples, as the distribution of novelty-like samples is more close to positive ones than outliers. In our framework, we employ cooperative learning between the generative network and contrastive network to generate better reconstructed images and provide more discriminative global latent vectors. It also motivates us to use the decoder-encoder framework instead of the conventional encoder-decoder frameworks with unilateral learning [Vincent *et al.*, 2010; Marchi *et al.*, 2017]. Therefore, we need to minimize the distance between the initialized latent vector z_i' and the global latent vector $\phi_{\theta_c}(x_i)$ as:

$$L_{lat} = \|z_i' - \phi_{\theta_c}(x_i)\|_2^2. \tag{2}$$

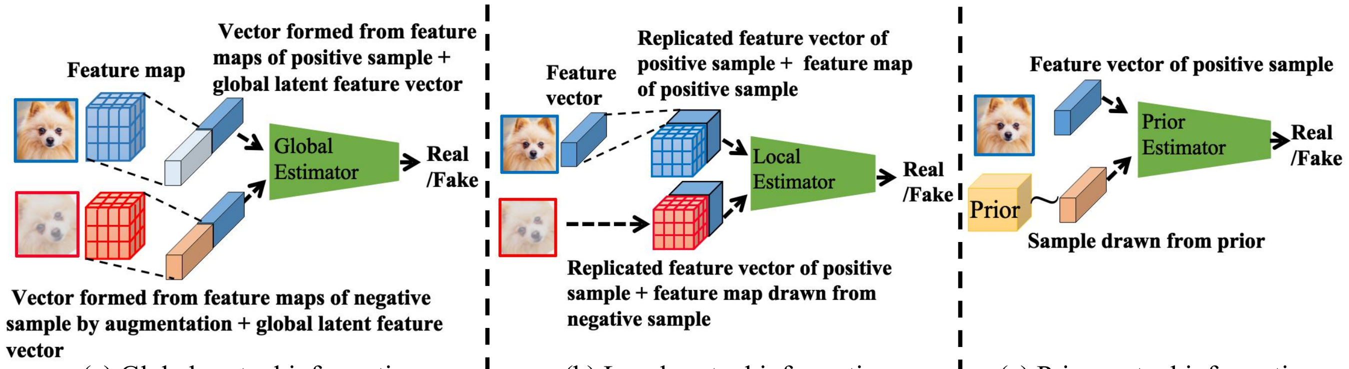
Inspired by [Sabokrou *et al.*, 2018], the reconstruction space is more effective for distinguishing the positive and novelty samples compared to the latent space. Therefore, a test example x first go through contrastive network and then generative network in a encoder-decoder manner during testing. The constraint by Eq. (2) makes the testing feasible. Without this constraint, the reconstructed images from positive samples will have significantly large reconstructed error even with the constraint of Eq. (1).

Mutual information estimator. Mutual information measures the essential relevance of two instances. The larger mutual information is, the more similar two variables have. Given two random variables x and y, mutual information [Belghazi *et al.*, 2018] can be estimated by the JS divergence between the joint $p(y \mid x)p(x)$ and the product of the margins $p(x)p(y)$. According to the definition of the variational estimation of JS divergence [Nowozin *et al.*, 2016], the maximization of the mutual information between variables X and Y can be formulated as:

$$\begin{aligned}\min_{\theta_e} -I(X,Y) = \min_{\theta_e} \{ &-(\mathbb{E}_{(x,y)\sim p(y|x)p(x)}[\log \sigma(T(x,y))] \\ &+\mathbb{E}_{(x,y)\sim p(y)p(x)}[\log(1-\sigma(T(x,y)))])\},\end{aligned} \tag{3}$$

where σ denotes the sigmoid function and $T(x) = \log \frac{2p(x)}{p(x)+q(x)}$. Here $p(z|x)p(x)$ and $p(z)p(x)$ are utilized to replace $p(x)$ and $q(x)$.

In our paper, we divide mutual information estimator into three parts: global estimator, local estimator and prior estimator. The goal of the mutual information estimator is to generate discriminative global latent features by contrastive learning between positive samples and negative samples. We first consider the global mutual information (see Fig. 3(a)). Based on Eq. (3), the maximization of global mutual information is also equivalent to minimize Eq. (4) by introducing


Feature map
Vector formed from feature maps of positive sample + global latent feature vector
Global Estimator
Real /Fake
Vector formed from feature maps of negative sample by augmentation + global latent feature vector
Feature vector
Replicated feature vector of positive sample + feature map of positive sample
Local Estimator
Real /Fake
Replicated feature vector of positive sample + feature map drawn from negative sample
Feature vector of positive sample
Prior Estimator
Real /Fake
Prior
Sample drawn from prior

(a) Global mutual information

(b) Local mutual information

(c) Prior mutual information

Figure 3: Local, global and prior mutual information estimation.

a global estimator $\tau_{\theta_{ge}}([A_z, z])$, which can be formulated as:

$$L_{global} = -\beta\Big(\mathbb{E}_{(A^p,z)\sim p(z|A^p)p(A^p)}[\log\sigma(\tau_{\theta_{ge}}([A^p_z,z]))] + \mathbb{E}_{(A^n,z)\sim p(z)p(A^n)}[\log(1-\sigma(\tau_{\theta_{ge}}([A^n_z,z])))]\Big), \quad (4)$$

where A^p_z and A^n_z are the vectors downscaled from the feature maps A^p and A^n from the positive sample and the negative sample, respectively. They all have the same dimension to z. $[A_z, z]$ is the concatenation between the downscaled feature A_z and z as an input pair. β is a hyperparameter. Similar to [Hjelm *et al.*, 2019], the optimization of Eq. (4) is to estimate the global latent feature distribution from positive samples by distinguishing the input from positive samples or negative samples.

Second, we consider local mutual information (see Fig. 3(b)), and also construct the relationship between the local feature map and the global latent feature. The process of estimation is the same as global mutual information. Thus, the objective function of local mutual information loss can be formulated as:

$$L_{local} = -\frac{\beta}{HW}\Sigma_{i,j}\Big(\mathbb{E}_{(A^p,z)\sim p(z|A^p)p(A^p)}\Big[\log\sigma\Big(\tau_{\theta_{le}}([A^p_{ij},z_A])\Big)\Big] + \mathbb{E}_{(A^n,z)\sim p(z)p(A^n)}\Big[\log\Big(1-\sigma\Big(\tau_{\theta_{le}}([A^n_{ij},z_A])\Big)\Big)\Big]\Big), \quad (5)$$

where $\tau_{\theta_{le}}([A_{ij}, z_A])$ is a local estimator to output the reverent probability between the local feature map A and latent representation z_A, which consists of a wide range of replicated feature vectors from z with the same dimension to A. An input pair is either from a positive sample $[A^p_{ij}, z_A]$ or a negative sample $[A^n_{ij}, z_A]$ at coordinates (i, j). H and W represent the height and width of the feature map.

Third, we employ the KL-divergence between the global latent feature and the prior distribution (*e.g.*, normal distribution) to encourage the latent feature to be more regular. Thus we can construct the following objective function:

$$L_{prior} = \gamma\mathbb{E}_{A\sim p(A)}[KL(p(z \mid A)\|q(z))], \quad (6)$$

where $q(z)$ is a prior distribution (*e.g.*, normal distribution) and $p(z|A)$ is the output of prior estimator $\tau_{\theta_{pe}}$. By combining the aforementioned three loss functions (*i.e.*, Eqs (3), (4) and (5)), we obtain the final mutual information estimator loss function:

$$L_{mie} = L_{global} + L_{local} + L_{prior}. \quad (7)$$

By minimizing the above function, we generate discriminative latent features z that helps to make the normal and novelty samples separable.

2.2 Overall Loss Function

According to the above discussion, we can construct the overall loss function for our decoder-encoder framework as:

$$L_{all} = \lambda_1 L_{lap} + \lambda_2 L_{lat} + \lambda_3 L_{mie} \quad (8)$$

where λ_1, λ_2 and λ_3 are the hyperparameters for balancing these three different terms. For solver, SGD optimizer can be directly used to minimize Eq. (8) in an end-to-end manner.

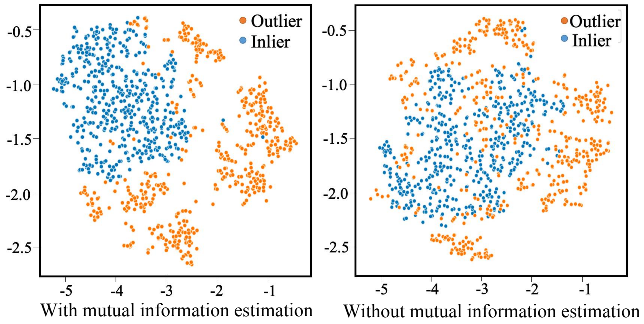
2.3 Implementation Details

We adopt the structure of DCGAN [Radford *et al.*, 2016] as the structure of the generative network and contrastive network. The global mutual information estimator is a fully-connected network with two 512-unit hidden layers. A 1 x 1 convnet with two 512-unit hidden layers is regarded as the local mutual information estimator. The prior mutual information estimator is a fully-connected network with two hidden layers of 1000 and 200 units.

Our image augmentation M contains cropping, horizontal flip, color jitter, rotation, and grayscale for random augmentations. During inference, the test sample x first goes through the contrastive network to be encoded into a latent vector. The generative network then upscales the latent vector to reconstruct the image from the learned discriminative latent feature space. Finally, the abnormal score is calculated by image reconstruction error between test image sample x and the corresponding generated image x'. The test sample x is regarded as a novelty instance if the image reconstruction error is larger than a predefined threshold T, and a normal instance otherwise. We use PyTorch [Paszke *et al.*, 2019] to implement our method. For training parameters, the learning rate and the number of total epochs are set to 0.002 and 100, respectively. SGD optimizer with momentum is adopted to

Local MI Global MI Prior MI		√	 √	 √	√ √	√ √	 √ √	√ √ √
CIFAR-10	0.750	0.832	0.838	0.823	0.841	0.842	0.842	0.843

Table 1: The effect of different mutual information estimation.



With mutual information estimation Without mutual information estimation

Figure 4: The visualization of latent space learned from target class (Frog class) by using proposed method with/without mutual information estimation.

optimize the parameters of our framework. Batch size, momentum and weight decay are set to 128, 0.9 and 0.005, respectively. For hyperparameters, β and γ are set to 0.5 and 0.1, respectively. λ_1, λ_2 and λ_3 are all set to 1.

3 Experiments

3.1 Experimental Setting

Public novelty detection dataset. We select CIFAR-10 [Krizhevsky and Hinton, 2009], COIL 100 [Nene *et al.*, 1996], MNIST [Lecun and Bottou, 1998], fMNIST [Xiao *et al.*, 2017] and DCASE [Mesaros *et al.*, 2017] as the standard evaluation dataset. COIL100 contains 100 objects with multiple different poses, where each class has less than 100 images. MNIST contains 60K/10K 28 × 28 training/test grayscale handwritten digit images from 0 to 9. fMNIST contains 60K/10K 28 × 28 training/test images of fashion appar-els/accessories. CIFAR-10 contains 50K/10K 32 × 32 training/test images with diverse content, background and complexity from 10 classes. In DCASE [Mesaros *et al.*, 2017] dataset, all abnormal event audios are artificially mixed with background audios (e.g. home, bus, and train) which contains 15 different acoustic background scenes. We use the original mixed audio files from the challenge, which contains 491, 496 and 500 audio files of roughly 30 seconds in the training, validation and test dataset, respectively. This task aims to distinguish abnormal acoustic signals from the normal ones.

Face anti-spoofing detection dataset. Replay-Attack [Chingovska *et al.*, 2012] dataset and CASIA-MFSD [Zhang *et al.*, 2012] dataset contain different attacks *e.g.* printed paper face, replaying a video and wearing a mask. Replay-Attack dataset contains 1.2K videos (200 real access videos and 1K attack videos). CASIA-MFSD dataset contains 50 subjects where each subject has 12 videos under 3 different image resolutions and varied lightings. The goal of this task is to detect whether a face is "alive" or just a fraudulent reproduction.

Dataset	Removal Feature			Reconstruction Loss		Different Prior			AUC
	z_A	z	None	MSE	Laplacian	N	C	P	
CIFAR-10	√				√			√	0.828
		√			√			√	0.821
			√		√			√	0.843
			√					√	0.530
			√	√				√	0.810
			√		√	√			0.745
			√		√		√		0.795

Table 2: Ablation study for proposed method. Part1:Removal of information inputs for the estimators. Part2:Comparsion of different reconstruction losses in generative network. Part3:The effect of different priors in generator inputs. N=Normal distribution, C=Contrastive prior, P=PCA prior.

Evaluation methodology. The protocol in the literature is proposed for one-class novelty detection [Perera *et al.*, 2019]. All of in-class training samples from only one class are used for training, and all samples in test set are used for testing. We use Area Under Curve (AUC) to evaluate the performance in novelty detection and acoustic anomaly detection. We also use Half Total Error Rate (HTER) [Bengio and Mariéthoz, 2004] for spoofing face detection.

3.2 Ablation Study

In this part, we evaluate the effect of each mutual information loss in Eq. (7), the effect of information fusion between feature maps and latent feature, the effect of Laplacian pyramid loss and the effect with/without PCA initialization. We conduct the experiments on CIFAR-10 for ablation study.

The effect of mutual information estimation. We first evaluate the effect of each component in mutual information estimation. The results are summarized in Tab. 1. Obviously, our model achieves a significantly higher AUC score by 0.843 with all mutual information estimation compared to that without any mutual information estimation (*i.e.* 0.75 AUC). This is due to the improvement of discriminative ability for the latent feature learned by the mutual information estimators. We also observe that global mutual information estimation (*i.e.* global loss Eq. (4)) achieves the best performance when using only one mutual information estimation loss. Note that training our model with two mutual information estimation losses achieves comparable results to the full three losses (see the 6th, 7th and 8th column in Tab. 1). Overall, three mutual information estimation losses indeed help to improve the performance of our model.

We further evaluate the mutual information estimation can help the model to obtain more discriminative feature representation by visualization. To this end, we randomly select 500 in-class samples (frog class) and 500 out-of-class samples from the dataset for testing [1]. As illustrated in Fig. 4, two global latent features are generated by our framework with/without mutual information estimation on CIFAR-10 by using t-SNE [van der Maaten and Hinton, 2008]. The normal samples (indicated by blue dots in Fig. 4 left) are significantly separated from novelty samples (indicated by yellow dots in Fig. 4 right) by using mutual information estimation, compared to that without this estimator.

[1] We run data selection 5 times and obtain a similar visualization result. For simplicity, we select one of them for visualization.

Methods	**PLANE**	**CAR**	**BIRD**	**CAT**	**DEER**	**DOG**	**FROG**	**HORSE**	**SHIP**	**TRUCK**	**Mean**
OCSVM ('01)	0.630	0.440	0.649	0.487	0.735	0.500	0.725	0.533	0.649	0.508	0.586
VAE ('13)	0.700	0.386	0.679	0.535	0.748	0.523	0.687	0.493	0.696	0.386	0.583
AnoGAN ('17)	0.671	0.547	0.529	0.545	0.651	0.603	0.585	0.625	0.758	0.665	0.618
DSVDD ('18)	0.617	0.659	0.508	0.591	0.609	0.657	0.677	0.673	0.759	0.731	0.648
ALOCC ('18)	0.620	0.717	0.537	0.560	0.587	0.563	0.612	0.605	0.744	0.671	0.622
RotNet ('18)	0.719	**0.945**	0.784	0.700	0.772	**0.866**	0.816	**0.937**	0.907	**0.888**	0.833
Neighbour ('19)	0.690	0.442	0.683	0.513	0.767	0.500	0.724	0.511	0.691	0.433	0.613
AND ('19)	0.717	0.494	0.662	0.527	0.736	0.504	0.726	0.560	0.680	0.566	0.617
IIC ('19)	0.684	0.894	0.498	0.653	0.605	0.591	0.493	0.748	0.818	0.767	0.674
Geometric ('19)	0.762	0.848	0.771	0.732	0.828	0.848	0.820	0.887	0.895	0.834	0.823
OCGAN ('19)	0.757	0.531	0.640	0.620	0.723	0.620	0.723	0.575	0.820	0.554	0.657
AE-EN ('20)	0.791	0.602	0.644	0.596	0.724	0.638	0.712	0.615	0.701	0.724	0.675
DROCC ('20)	0.817	0.767	0.667	0.671	0.736	0.744	0.744	0.714	0.800	0.762	0.742
DualGAN ('20)	0.875	0.548	0.719	0.639	0.833	0.643	0.810	0.581	0.872	0.503	0.703
Ours	**0.985**	0.765	**0.796**	**0.791**	**0.924**	0.717	**0.975**	0.691	**0.985**	0.752	**0.843**

Table 3: AUC of different novelty detection methods on CIFAR-10. Plane and car denote Airplane and Automobile in CIFAR-10, respectively.

	MNIST	**COIL**	**fMNIST**
ALOCC DR ('18)	0.88	0.809	0.753
ALOCC D ('18)	0.82	0.686	0.601
DCAE ('14)	0.899	0.949	0.908
GPND ('18)	0.932	0.968	0.901
Rot ('18)	0.933	0.970	0.935
OCGAN ('19)	0.977	0.995	0.924
DualGAN ('20)	0.985	**1.0**	0.995
Ours	**0.986**	**1.0**	**0.999**

Table 4: Mean AUC results of the different methods on MNIST, COIL and fMNIST.

Scenarios	**RotNet ('18)**	**Geometric ('19)**	**WaveNet ('19)**	**DualGAN ('20)**	**Ours**
Beach	0.508	0.522	0.72	0.82	**0.86**
Bus	0.562	0.585	0.83	**0.96**	**0.96**
restaurant	0.507	0.560	0.76	**0.80**	0.78
Car	0.606	0.664	0.82	**0.99**	**0.99**
City center	0.510	0.532	0.82	0.89	**0.90**
Forest path	0.515	0.530	0.72	0.78	**0.80**
Grocery store	0.511	0.530	0.77	0.90	**0.95**
Home	0.504	0.511	0.69	**0.90**	0.68
Library	0.514	0.531	0.67	0.89	**0.97**
Metro station	0.512	0.533	0.79	0.89	**0.93**
Office	0.508	0.523	0.78	0.87	**0.94**
Park	0.520	0.550	0.80	0.95	**0.99**
Residential area	0.512	0.527	0.78	0.78	**0.81**
Train	0.522	0.567	0.84	0.92	**0.95**
Tram	0.568	0.577	0.87	**0.97**	**0.97**

Table 5: AUC scores for all methods on DCASE dataset with 15 scenarios.

The effect to remove z_A or z. In local and global mutual information estimation, the concatenation of feature maps and latent vector are used as the inputs of their estimators during training. The discriminative feature representation is learned by distinguishing the concatenated features all from positive samples and part from negative samples. To evaluate the effectiveness of the concatenated features, we remove the information of latent matrix z_A in the local estimator or latent vector z in the global estimator. In the first part of Tab. 2, the (first 3 rows) combination without any removal in our estimators achieves the highest AUC, compared to the removals of z_A or z.

The effect of Laplacian pyramid loss. As shown in the second part of Tab. 2, the framework without any reconstruction loss results in the worst average AUC score of 0.53. We further compare the Laplacian pyramid loss with the MSE loss. As presented in the second part of Tab. 2, the performance of Laplacian pyramid loss significantly outperforms the MSE loss (*i.e.* 0.843 AUC *vs.* 0.810 AUC). To explain, Laplacian pyramid loss is a perception-level error, which is more effective for novelty detection than MSE loss.

The effect with/without PCA initialization. We compared our PCA initialization with normal distribution initialization and contrastive prior. Contrastive prior only uses the output z of the contrastive network at the 100th epoch training as initialization. As shown in the third part of Tab. 2, PCA initialization achieves the best performance, compared to other initialization methods. This is because PCA encodes the information of the entire training data, which alleviates the problem of mode dropping.

3.3 Comparison with State-of-the-art Methods

Novelty Detection

We first evaluate the effectiveness of our method on CIFAR-10. As shown in Tab. 3, our method achieves the highest mean AUC of 0.843, compared to other SOTA methods. We also found that the self-supervised learning methods based on the pretext tasks (e.g., RotNet [Komodakis and Gidaris, 2018], Geometric [Hendrycks *et al.*, 2019]) achieves higher performance, compared with GAN-based methods (e.g., OCGAN [Perera *et al.*, 2019] and DualGAN [Chen *et al.*, 2020a]). For the individual class, our method also shows the best results, except the class car, dog, horse and truck. The semantic information of these classes has high related to the object shape in CIFAR-10. Since the RotNet focuses on the semantic information of object shape, it achieves high performance in these classes. However, it lack capturing other semantic feature(*e.g.* object color and object texture).

Following [Perera *et al.*, 2019], we also evaluate the performance of novelty detection methods on COIL100, MNIST, fMNIST by using the following evaluation setting: The 80% of in-class samples are regarded as a normal class for training, while the rest of 20% of in-class samples is adopt for testing. Out-of-class test samples have the same number of in-class test samples, which are randomly selected from the test set. As shown in Tab. 4, our method achieves the best performance across different datasets, compared to state-of-the-art methods. For example, in MNIST dataset and fMNIST dataset, we achieve the improvement of average AUC score 0.1% and 0.4% respectively.

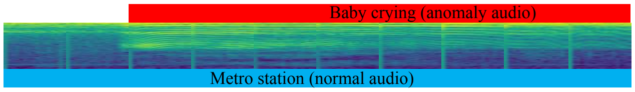


Figure 5: The visualization of test sample: Metro background audio mixed with baby crying.

Acoustic Anomaly Detection

The original WaveNet [Oord *et al.*, 2016] has been successfully applied into raw audio generation and music synthesis, which benefits from its powerful convolutional autoregressive architecture. Recently, [Rushe and Mac Namee, 2019] has extended its structure for anomaly detection in raw audio. Thus, we also make a comparison with it on the DCASE dataset, which is denoted by WaveNet for convenience. As shown in Tab. 5, we present the results of different methods across 15 classes/scenarios. Obviously, our method achieves the best performance in most of the scenarios, except the restaurant and home scenarios, compared to CAE, WaveNet, and DualGAN. For example, our method outperforms the best DualGAN by 8% AUC on library scenario. Note that DualGAN achieves amazing performance with 0.9 AUC. We conjecture this is due to the randomness of the DualGAN method in the home background. Interestingly, self-supervised learning based methods (*e.g.* RotNet [Komodakis and Gidaris, 2018] and Geometric [Hendrycks *et al.*, 2019]) shows worse performance on DCASE. This is due to the failure to presentation of object colors and object textures. Actually, as shown in Fig. 5, the main discriminative features between normal and novelty audio in spectrogram are from object colors and textures. Rich semantic information of object shapes obtained by these methods cannot separate the normal samples from novelty samples.

Spoofing Face Detection

Our method also works for spoofing face detection. Actually, we formulate face anti-spoofing detection as a novelty detection task by only using the normal (live faces) samples for training. We intuitively take the optical flow obtained from consecutive video frames as input, since the data source of spoof face detection is from the video. To this end, we first extract the optical flow using FlowNet2.0 [Ilg *et al.*, 2017] from each video with 30 fps.

Following [Tu *et al.*, 2020], the detection models select the training set from one of the training set in CASIA-MFSD and Replay-Attack dataset, and the test set from the other test dataset. The results are summarized in Tab. 6. The proposed method achieves the best performance (HTER = 0.175) on the Replay-Attack test set of which includes different types of spoofing attacks. On the other dataset setting, our method achieves a competitive performance (HTER = 0.308) on the testing set of the CASIA-MFSD dataset. Actually, the HTER achieved by our unsupervised method significantly lower than supervision methods except auxiliary method [Liu *et al.*, 2018]. This is probably due to the help of the additional depth information and rPPG signal. Nevertheless, our method achieves better performance by simultaneously evaluating these two tasks on the average HTER. We also found that the generalization of these models trained on the CASIA-MFSD dataset is better than the models trained on the Replay-Attack dataset. We speculate that the detection scenario on the CASIA-MFSD dataset is more complex than the Replay-Attack dataset, which leads to the easy learning of more knowledge during training.

Methods	Train on CASIA-MFSD &Test on Replay-Attack	Train on Replay-Attack &Test on CASIA-MFSD
LBP ('13)	0.470	0.396
LBP-TOP ('13)	0.497	0.606
Motion ('13)	0.502	0.479
CNN ('14)	0.485	0.455
Color LBP ('15)	0.379	0.354
Color Tex ('16)	0.303	0.377
Auxiliary ('18)	0.276	0.284
De-Spoof ('18)	0.285	0.411
DA ('18)	0.274	0.360
D-texture ('18)	0.222	0.350
OF Domain ('18)	0.301	0.368
ADA ('19)	**0.175**	0.416
GFA-CNN ('20)	0.214	0.343
DualGAN ('20)	0.223	**0.246**
Ours	**0.175**	0.308

Table 6: Classification performance of the proposed approach in terms of HTER.

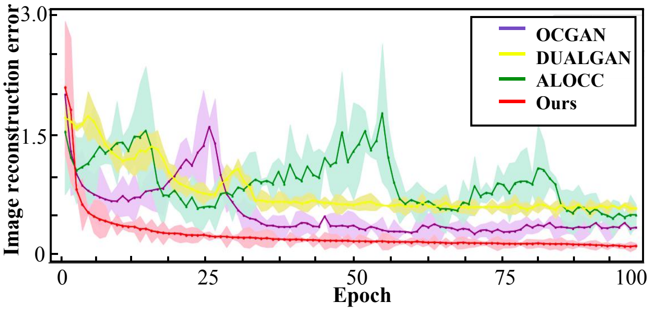


Figure 6: The visualization of training loss comparison between previous methods and proposed method on CIFAR-10.

3.4 Stability and Convergence

We further evaluate the effectiveness of our method during training. Our decoder-encoder framework achieves more stable training and fast convergence speed in CIFAR-10 dataset, compared to other encoder-decoder frameworks, *e.g.* OCGAN, ALOCC and DualGAN. As shown in Fig. 6, the proposed decoder-encoder framework tends to be convergent after the 25-th epoch, while the convergent value of other methods is significantly larger (*e.g.* about 40 epochs in OCGAN). In addition, the image reconstruction error is reduced steadily using the decoder-encoder framework, while the significant large fluctuations occur in other methods. This is due to the usage of the mutual information estimator that adversarial optimization is removed in our framework.

4 Conclusion

In this paper, we propose a novel decoder-encoder framework for novelty detection. To alleviate mode dropping of GANs, each latent initialized vector is mapped to an image by the PCA initialization in the generative network. To learn a more discriminative latent feature representation, we introduce a contrastive network to learn to compare the different input pairs through mutual information estimation. Specifically, our framework is trained without adversarial optimization, which benefits fast convergence and stable training of our model. We have comprehensively evaluated the performance of our method on a variety of novelty detection tasks over different datasets, which demonstrates the superior performance gains over the state-of-the-art methods.

Acknowledgments

We thank for the support from National Natural Science Foundation of China (61772524, 61902129, 61972157, 61876161, 61701235, 61373077), Shanghai Pujiang Talent Program (19PJ1403100), the Science and Technology Commission of Pudong (NO. PKJ2018-Y46) and Shanghai Jiaotong University Translational Medicine Cross Foundation (ZH2018ZDA25), Shanghai Sailing Program (21YF1411200), Natural Science Foundation of Shanghai (20ZR1417700), the Fundamental Research Funds for the Central Universities, CAAI-Huawei MindSpore Open Fund.

References

[Arora *et al.*, 2018] Sanjeev Arora, Andrej Risteski, and Yi Zhang. Do gans learn the distribution? some theory and empirics. In *ICLR*, 2018.

[Belghazi *et al.*, 2018] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *ICML*, 2018.

[Bengio and Mariéthoz, 2004] Samy Bengio and Johnny Mariéthoz. A statistical significance test for person authentication. In *Odyssey*, 2004.

[Chen *et al.*, 2020a] Chengwei Chen, Jing Liu, Yuan Xie, Yin Xiao Ban, Chunyun Wu, Yiqing Tao, and Haichuan Song. Latent regularized generative dual adversarial network for abnormal detection. In *IJCAI*, 2020.

[Chen *et al.*, 2020b] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[Chingovska *et al.*, 2012] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *BIOSIG*, 2012.

[Hendrycks *et al.*, 2019] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. In *NeurIPS*, 2019.

[Hjelm *et al.*, 2019] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019.

[Ilg *et al.*, 2017] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, 2017.

[Ji *et al.*, 2019] Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *ICCV*, 2019.

[Komodakis and Gidaris, 2018] Nikos Komodakis and Spyros Gidaris. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018.

[Krizhevsky and Hinton, 2009] Alex. Krizhevsky and Geoffrey. Hinton. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

[Lecun and Bottou, 1998] Yann Lecun and Leon Bottou. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[Lim *et al.*, 2018] Swee Kiat Lim, Yi Loo, Ngoc-Trung Tran, Ngai-Man Cheung, Gemma Roig, and Yuval Elovici. Doping: Generative data augmentation for unsupervised anomaly detection with gan. In *ICDM*, 2018.

[Ling and Okada, 2006] Haibin Ling and Kazunori Okada. Diffusion distance for histogram comparison. In *CVPR*, 2006.

[Liu *et al.*, 2018] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *CVPR*, 2018.

[Liu *et al.*, 2020] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Zhaoyu Wang, Li Mian, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *arXiv preprint arXiv:2006.08218*, 1(2), 2020.

[Marchi *et al.*, 2017] Erik Marchi, Fabio Vesperini, Stefano Squartini, and Björn Schuller. Deep recurrent neural network-based autoencoders for acoustic novelty detection. *Computational intelligence and neuroscience*, 2017, 2017.

[Mesaros *et al.*, 2017] Annamaria Mesaros, Toni Heittola, Aleksandr Diment, Benjamin Elizalde, Ankit Shah, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen. Dcase 2017 challenge setup: Tasks, datasets and baseline system. In *DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events*, 2017.

[Nene *et al.*, 1996] Samer A. Nene, Shree K. Nayar, and Hiroshi Murase. Columbia object image library (coil-20). Technical Report CUCS-005-96, Department of Computer Science, Columbia University, February 1996.

[Nowozin *et al.*, 2016] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *NeurIPS*, 2016.

[Oord *et al.*, 2016] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[Paszke *et al.*, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

[Perera *et al.*, 2019] Pramuditha Perera, Ramesh Nallapati, and Bing Xiang. Ocgan: One-class novelty detection using gans with constrained latent representations. In *CVPR*, 2019.

[Radford *et al.*, 2016] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.

[Rushe and Mac Namee, 2019] Ellen Rushe and Brian Mac Namee. Anomaly detection in raw audio using deep autoregressive networks. In *ICASSP*, 2019.

[Sabokrou *et al.*, 2018] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. Adversarially learned one-class classifier for novelty detection. In *CVPR*, 2018.

[Sinha *et al.*, 2021] Abhishek Sinha, Kumar Ayush, Jiaming Song, Burak Uzkent, Hongxia Jin, and Stefano Ermon. Negative data augmentation. *arXiv preprint arXiv:2102.05113*, 2021.

[Tack *et al.*, 2020] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. In *NeurIPS*, 2020.

[Tu *et al.*, 2020] Xiaoguang Tu, Zheng Ma, Jian Zhao, Guodong Du, Mei Xie, and Jiashi Feng. Learning generalizable and identity-discriminative representations for face anti-spoofing. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–19, 2020.

[van der Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[Vincent *et al.*, 2010] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(Dec):3371–3408, 2010.

[Xiao *et al.*, 2017] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[Zhang *et al.*, 2012] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z Li. A face antispoofing database with diverse attacks. In *ICB*, 2012.

[Zhao *et al.*, 2017] Junbo Jake Zhao, Michaël Mathieu, and Yann LeCun. Energy-based generative adversarial networks. In *ICLR*, 2017.