applied sciences

MDPI

*Article*

# Complex Background Reconstruction for Novelty Detection

**Kun Zhao [1], Man Su [2], Ran An [1], Hui He [1] and Zhi Wang [3,*]**

[1] The School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China; kunzhao@xjtu.edu.cn (K.Z.); ranan@stu.xjtu.edu.cn (R.A.); huihe@xjtu.edu.cn (H.H.)
[2] Beijing Institute of Tracking and Telecommunication Technology, Beijing 100094, China; suman_bcng@163.com
[3] The School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China
* Correspondence: zhiwang@xjtu.edu.cn

**Abstract:** Novelty detection aims to detect samples from classes different from the training samples (i.e., the normal class). Existing approaches predominantly make the target reconstruction better and choose the appropriate reconstruction error measurement method but ignore the influence of background information on this process. This paper proposes a novel reconstruction network and mutual information Siamese network. The reconstructed network aims to make the distribution of reconstructed samples consistent with that of original samples, intending to reduce background interference in the reconstruction process. After this, we measure the distance between the original and generated images based on a mutual information Siamese network, which extracts more discriminative features to calculate the similarity between the original images and their reconstructed ones. This part of the network uses global context information to improve the detection accuracy. We conduct extreme experiments to evaluate the proposed solution on two challenging public datasets. The experimental results show that the proposed method significantly outperforms the state-of-the-art methods.

**Keywords:** figure reconstruction; one-class novelty detection; mutual information fusion



**Citation:** Zhao, K.; Su, M.; An, R.; He, H.; Wang, Z. Complex Background Reconstruction for Novelty Detection. *Appl. Sci.* **2023**, *13*, 10702. https://doi.org/10.3390/app131910702

Academic Editor: Antonio Fernández-Caballero

Received: 21 July 2023
Revised: 16 September 2023
Accepted: 20 September 2023
Published: 26 September 2023



## 1. Introduction

Novelty detection tackles an important unsupervised learning problem where novelty samples are not known a priori and the majority of the training dataset consists of "normal" data [1]. This problem has been widely applied in many areas, including abnormality detection [2,3], intruder detection [4], biomedical data processing [5], imbalance learning [6], vehicle tracking [7], and specific sign detection [8]. Different from other machine learning tasks, methods for one-class novelty detection are trained on only one class (i.e., the normal class) and aim to determine whether the given sample is a novelty during the inference stage. The challenges associated with anomaly detection tasks encompass several key aspects. (1) Lack of manual supervision: Anomaly detection often operates in an unsupervised or semi-supervised manner, meaning that there is limited availability of labeled data specifically indicating anomalies. This scarcity of labeled data hampers the ability to train models effectively. (2) Limited novelty samples: Novelty or anomaly samples are typically rare and occur infrequently in real-world data. The small number of available anomaly samples makes it challenging for models to learn and generalize effectively. (3) Imbalanced training data: In most cases, the training dataset consists primarily of instances from the normal class, making it imbalanced. This imbalance can lead to model bias towards the majority class and the poor detection of anomalies. (4) Unknown anomaly patterns: Anomalies can take various forms, and their characteristics may not be well-defined or known in advance. This uncertainty about the nature and structure of anomalies adds complexity to the task.

In recent years, novelty detection tasks have received extensive attention. According to the recent literature, existing models for novelty detection can be generally divided into two

branches. The first branch is to extract latent features of an image as the input to traditional novelty detection algorithms such as one-class SVM [9]. Another one is to reconstruct the image using the deep generative networks [10], which are trained on samples from the normal class, and use the reconstruction error as a measure for novelty sample detection. The reconstruction approach using deep generative networks focuses on learning potential information for the retaining of normal samples, but, for the novelty samples, it will have large reconstruction errors. In this strategy, existing methods improve the performance by making image reconstruction better and choosing more appropriate distance metrics. Most of the existing methods ignore the effects of the image background. Specifically, the one-class novelty detection results for the CIFAR10 dataset are comparatively weaker for all methods. A complex background can affect the result of image reconstruction, which will affect the choice of distance metrics. However, background information can also help in object detection. In natural surroundings, the foreground and the background are strongly correlated. For instance, the sky background naturally correlates with birds and airplanes with a higher probability, whereas the ground background relates to terrestrial animals and vehicles. Hence, the background, especially a complex one, subtly indicates certain foreground elements, with distinctive backgrounds potentially indicating novel samples. Furthermore, owing to the expansive and non-discriminatory nature of neural networks, background information and foreground information exhibit a certain diffusion effect, resulting in the inclusion of some foreground information within the background. Hence, this paper proposes a method for complex background reconstruction, thereby facilitating the detection of novel targets.

Inspired by this, we offer improvements from the following aspects. First, we use the Maximum Mean Discrepancy (MMD) to make the distribution of reconstructed samples similar to that of the samples in the training set. Second, we propose a similarity metric network to measure the similarity between the sample and its reconstructed one. This network extracts more discriminant features as the basis for judging similarities and dealing with complex backgrounds and distributed discrete samples more effectively.

As highlighted in Figure 1, the uppermost group showcases the input images. The intermediate group depicts the reconstructed images exploiting L2 loss, whereas the lowermost group exhibits the reconstructions obtained using our solution. It is discernible that in the left segment, all the samples correspond to autonomous vehicles, which can be considered as standard instances. Conversely, the right segment contains a multitude of novel samples, including airplanes, trucks, and animals, etc. Notably, in both cases, it is evident that our solution improves the background reconstruction of the respective samples.

In this work, we propose a method for novelty detection. First, we propose a new reconstruction network (RNet) based on MMD-GAN [11] for the reconstruction of images. Second, we propose a similarity metric network (MNet) to learn the distance metric between latent representations of original images and reconstructed ones, which is used to determine whether the image is a novelty. In our method, RNet is trained on a dataset with only normal samples, and MNet takes samples and their reconstructed images generated by RNet as inputs. The goal of the whole model is to minimize the distance between normal samples and their reconstructed ones. At the testing stage, the detection result is based on the distance between the original sample and its reconstructed image.

Our contributions are three-fold.

- We propose a reconstruction network that not only can learn the features of the target, but also the distribution of the background. Reconstructing background information can reduce the interference of the background information on reconstructed objects
- We propose a Siamese neural network with mutual information to learn the similarity between original images and their reconstructions. This method not only can extract more discriminant features about objects than the previous method but can also obtain background context information. Therefore, the target classification accuracy is significantly improved.

- Experiments on challenging datasets with complex backgrounds demonstrate the superiority of the proposed model.

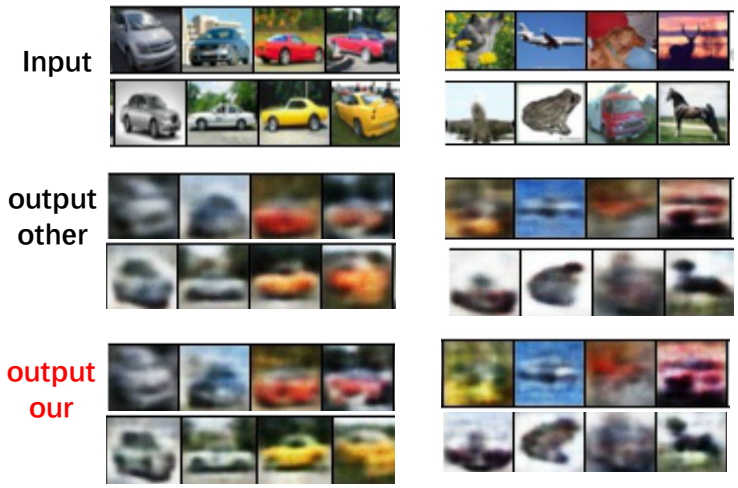



**Figure 1.** The result of our image reconstruction is compared with that of other methods. The left side shows normal samples (automobile) and the right side shows novel samples. For normal samples, our method has a better reconstruction effect than other methods (using $L2$ loss) on the whole image, because it not only learns the features of the target as in the previous method, but also the distribution of the background. Reconstructing background information can reduce the interference of the background information in reconstructed targets.

## 2. Related Work

In this section, we review the related work on novelty detection, MMD, and metric learning.

**Novelty Detection:** The one-class novelty detection task has attracted considerable attention in recent years. With the success of neural networks and deep learning, great progress has been made in novelty detection. According to the recent literature, approaches to novelty detection can be generally divided into two branches. One is to extract the latent features of the image as the input to a traditional anomaly detection algorithm such as one-class SVM (OC-SVM) [12]. Hoffmann [13] and Sakurada and Yairi [14] propose to leverage the mean squared error to conduct novelty detection. Gruhl et al. [15] leverage a self-adaptive and self-organizing paradigm for novelty detection. Deep One-Class Classification [16] jointly trains a deep neural network while optimizing a data-enclosing hypersphere in the output space, which firstly introduces the fully deep one-class classification objective for unsupervised anomaly detection.

Another approach is to reconstruct the image using a deep generative network. Since generative adversial networks (GANs) [17] have shown a strong ability to extract deep features, several works apply GANs in novelty detection. AnoGAN [2] hypothesizes that the latent vector mapped from the input of the GAN represents the high-dimensional distribution of the data. The detection process is based on the reconstruction error and the error between the intermediate discriminator feature of the test image and the reconstructed image. Meanwhile, in ADGAN [18], which shares a similar framework with AnoGAN, the network is optimized based on both the latent vector and the generator, and only the reconstruction error is considered as the anomaly score while inferencing. Sabokrou et al. [1] attempt to de-noise noisy samples of the given class, and the discriminator's prediction in the image space is used to quantify the reconstruction error. Li et al. [19] propose an augmented time-regularized generative adversarial network to generate effective artificial samples for novelty detection. Chen et al. [20] leverage an encoder–decoder reconstruction network and a CNN-based discrimination network to recognize noisy novel samples. Almohsen et al. [10] exploit an isometric adversarial auto-encoder to obtain stable detection results.

Nevertheless, the aforementioned approaches primarily concentrate on extracting and generating features from the foreground, neglecting the significant correlation that it shares with the background, especially for complex background cases. These methods depend solely on foreground information to detect novel targets. Highlighting the aforementioned concerns, this paper posits that complex backgrounds, with their implicit association with the foreground, can be leveraged to enable the detection of novel targets. By integrating both the foreground and background, the accuracy of novelty detection can be significantly enhanced.

**MMD**: MMD (Maximum Mean Discrepancy) is a statistical hypothesis aiming to measure the distance between two probability distributions. Schölkopf et al. [21] first proposed the criterion to replace the hard minimax objective function used in generative adversarial network training. The generative moment matching network (GMMN) [22,23] is a generative model that replaces the discriminator in the GAN with a two-sample test based on kernel MMD. The visualization result of the generated moment matching network is somewhat disappointing; Li et al. [22] improved it by combining the generated moment matching network with the auto-encoder. However, the empirical performance of GMMN and the computational efficiency of GMMN is not as competitive as that of GAN on challenging benchmark datasets. MMD-GAN [11] combines the key ideas of both GMMN and GAN; the authors proposed a method to improve both the expressive ability and computational efficiency of the model by replacing the fixed Gaussian kernel in the original GMMN with adversarial kernel learning techniques.

**Metric Learning**: Metric learning, also known as similarity learning, aims to learn by measuring the similarity between pairs of images, which is essential in many computer vision tasks, such as image retrieval, image matching, image verification, and multi-category tasks. Chopra et al. [24], Hadsell et al. [25] proposed the Siamese architecture for this purpose, and these are two important works often cited in this subject. Unlike the classification network, the goal of metric learning is to learn by measuring the similarity of the two instances in terms of the Euclidean distance, rather than the representation of the classification. Another popular architecture is the Triple Network [26]. For both of them, many authors have realized that it is important to mine a sample of the training set to find difficult or challenging pairs or triplets in order to converge faster or better reach the minimum [27].

## 3. Methodology

In this section, we introduce our model as a method for one-class novelty detection. The framework of the proposed method is illustrated in Figure 2. This model consists of two components: (1) the reconstruction network (RNet) for the reconstruction of the image; (2) the mutual information Simaese metwork (MNet) for the measurement of the similarity between the original image and the reconstructed image. These two components are trained in an adversarial and unsupervised manner.

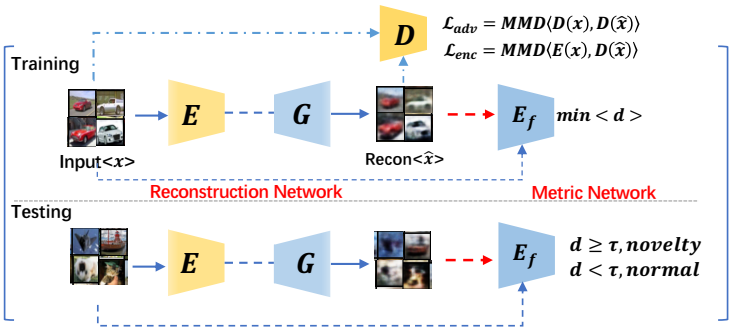



**Figure 2.** The framework of the our model. There are two cascaded components: the image reconstruction network (RNet) and metric network. Actually, the metric network is a mutual information

Siamese network (MNet). The model is trained on samples from the normal class. The RNet is optimized by the MMD loss in order to make the distribution of the reconstructed image $\hat{x}$ similar to the input image $x$'s. The inputs of MNet are $x$ and $\hat{x}$. We minimize the distance $d\langle f(x), f(\hat{x})\rangle$ between $f(x)$ and $f(\hat{x})$ while training MNet, where $f(\cdot)$ denotes a function of the feature extractor in $E_f$. During testing, we calculate the distance between $f(x)$ and $f(\hat{x})$; if $d\langle f(x), f(\hat{x})\rangle$ is not greater than the given threshold $\tau$, $x$ is considered as a novelty sample; otherwise, it is a normal sample.

### *3.1. Problem Settings*

In the one-class novelty detection task, our goal is to train an unsupervised novelty detection network that performs well on samples subjected to a dispersed data distribution. We notice that these samples tend to have large variance and the backgrounds show great variety.

The formal definition of the task is as follows. Consider a training dataset $\mathcal{D} = \{x_i^D\}_{i=1}^N$ with $N$ normal samples, and a testing dataset $\hat{\mathcal{D}} = \{(x_i^{\hat{D}}, y_i^{\hat{D}})\}_{i=1}^M$ with $M$ normal and novelty samples, where $y_i^{\hat{D}} \in \{0, 1\}$ denotes the sample label. Here, $y_i^{\hat{D}} = 0$ means that $x_i^{\hat{D}}$ is classified as a normal sample, whereas $y_i^{\hat{D}} = 1$ means that the corresponding $x_i^{\hat{D}}$ is classified as a novel sample. The goal of the task is to train a model on $\mathcal{D}$ to optimize a novelty score function $\mathcal{S}(\cdot)$. For a given testing sample $x^* \in \hat{\mathcal{D}}$ and a given threshold $\tau$, if $\mathcal{S}(x^*) > \tau$, $x^*$ is a detected novelty.

### *3.2. Reconstruction Network (RNet)*

In previous work [14], samples were typically reconstructed using auto-encoders and by generating normal samples against network training. The essential goal of reconstruction is to classify different object categories by calculating the reconstruction error between $x$ and $\hat{x}$. By reviewing the previous work [28], we find that most of the methods do not perform well on some datasets (e.g., CIFAR-10). There exist studies showing that making the distribution of real data similar to that of reconstructed data can benefit image generation [11]. Inspired by these studies, we still adopt the strategy of deep reconstruction networks, but we replace the discriminator in GAN with a two-sample test based on MMD.

Specifically, our RNet consists of three subnetworks: (1) the encoding subnetwork $E$ to map the image $x$ to the latent representation $z$, (2) the reconstruction subnetwork $G$ to reconstruct $\hat{x}$ by decoding $z$, and (3) the discrimination subnetwork $D$. In the reconstruction network, we employ a classical GAN architecture to extract background features. In the encoder part, we utilize the Maximum Mean Discrepancy (MMD) as a measure of dissimilarity between the original and generated images. This choice is motivated by the fact that, compared to other distance measures between distributions, such as the Kullback–Leibler (KL) divergence, which either require density estimation (either parametric or nonparametric) or space partitioning/bias correction strategies, MMD can be easily estimated as an empirical mean that converges to the true value of the MMD. Thus, MMD exhibits superior performance in the scenario of unknown image distributions. Moreover, MMD is equivalent to finding the Reproducing Kernel Hilbert Space (RKHS) function that maximizes the expectation difference between the two probability distributions, indicating its strong solution coherence. In the reconstruction part, we utilize the 1-norm to promote sparse features.

**Adversarial Loss:** Following the trend in current novelty detection methods [2], we use the feature distribution loss for adversarial learning. Studies have shown that feature matching loss can reduce the instability of GAN training [29].

MMD distance was first proposed for the two-sample test problem, which aims to determine whether two given distributions are the same. It uses kernel embedding $\phi(\boldsymbol{x}) = k(\cdot, \boldsymbol{x})$ associated with a characteristic kernel $k$, where $\phi$ is infinite-dimensional

and $\langle \phi(\boldsymbol{x}), \phi(\boldsymbol{y}) \rangle_{\mathcal{H}} = k(\boldsymbol{x}, \boldsymbol{y})$. Given two distributions $\mathbb{P}$ and $\mathbb{Q}$, and a kernel $k$, the squared MMD distance is defined as

$$M_k^2(\mathbb{P}, \mathbb{Q}) = \left\| \boldsymbol{\mu}_{\mathbb{P}} - \boldsymbol{\mu}_{\mathbb{Q}} \right\|_{\mathcal{H}}^2 = \mathbb{E}_{\boldsymbol{x}, \boldsymbol{x}' \sim \mathbb{P}} \left[ k(\boldsymbol{x}, \boldsymbol{x}') \right] + \mathbb{E}_{\boldsymbol{y}, \boldsymbol{y}' \sim \mathbb{Q}} \left[ k(\boldsymbol{y}, \boldsymbol{y}') \right] - 2\mathbb{E}_{\boldsymbol{x} \sim \mathbb{P}, \boldsymbol{y} \sim \mathbb{Q}} [k(\boldsymbol{x}, \boldsymbol{y})]. \quad (1)$$

where $\mu$ and $\mathbb{E}[\cdot]$ denote the mean value of the distribution and the expectation of the kernel function, respectively. The kernel $k(\boldsymbol{x}, \boldsymbol{y})$ measures the similarity between two samples $\boldsymbol{x}$ and $\boldsymbol{y}$. $\boldsymbol{x}$ and $\boldsymbol{x}'$ represent two random variables subjected to $\mathbb{P}$, and $\boldsymbol{y}$ and $\boldsymbol{y}'$ represent random variables subjected to $\mathbb{Q}$.

With respect to the square of the MMD distance, we have the theorem below.

**Theorem 1.** *Given a kernel $k$, if $k$ is a characteristic kernel, then $M_k^2(\mathbb{P}, \mathbb{Q}) = 0$ if $\mathbb{P} = \mathbb{Q}$.*

Theorem 1 [30] shows that the more similar the two distributions are, the smaller the MMD distance between them is. In this work, we input the original image and the reconstructed image into the discriminator $D$ to obtain the corresponding features $D(x)$ and $D(\hat{x})$. The feature distribution matching computes the similarity between the distributions of $D(x)$ and $D(\hat{x})$, which is measured by the MMD distance.

Formally, let $f(\cdot)$ denote the function that outputs an intermediate layer of the encoder $D$ for a given input $x$ sampled from the input data distribution $\mathbb{P}$; the function of the discriminator $D$ can be considered as forming a new kernel with $k$: $k \circ D(\boldsymbol{a}, \boldsymbol{b}) = k(D(\boldsymbol{a}), D(\boldsymbol{b})) = k_D(\boldsymbol{a}, \boldsymbol{b})$. Thus, the adversarial loss can be defined as

$$\mathcal{L}_{adv} = M_{k \circ D}^2(f(x), f(\hat{x})). \quad (2)$$

**Reconstruction Loss:** In RNet, the reconstructed image is obtained by decoding the latent representation of the samples. In order to optimize the decoder, the reconstruction loss is defined as

$$\mathcal{L}_{con} = \mathbb{E}_{x \sim p_{\mathbf{X}}} \|x - \hat{x}\|_1. \quad (3)$$

**Encoder Loss:** Both losses above are designed to force the generator to reconstruct the input sample better. Moreover, we add a loss to minimize the MMD distance between the distributions of $E_1(x)$ and $E_2(\hat{x})$. The encoder loss is in the same form as the adversarial loss, which is defined as

$$\mathcal{L}_{enc} = M_k^2\Big(E(x), D(\hat{x})\Big). \quad (4)$$

To summarize, the loss function of RNet is a linear combination of the reconstruction loss, the adversarial loss, and the encoder loss:

$$\mathcal{L}_{RNet} = \mathcal{L}_{con} + \gamma \mathcal{L}_{adv} + \delta \mathcal{L}_{enc}. \quad (5)$$

where $\gamma$ and $\delta$ are weighting factors of the adversarial loss and the encoder loss, respectively.

### *3.3. Mutual Information Siamese Network (MNet)*

We proposed a network called the mutual information Siamese network (MNet) to measure the similarity between reconstructed images and original images. The similarity judges the quality of the reconstructed image and determines whether the images are novelties.

In the previous methods of novelty detection based on GAN/AE, they detect whether the sample is a novelty by calculating the reconstruction error of the input image $x$ and the reconstruction image $\hat{x}$. The general selection for the reconstruction error is the mean square error (MSE). However, RNet pays more attention to the object rather than the background for images with complex backgrounds. In this case, the quality of reconstruction is limited.

In this work, we employ ideas from metric learning based on deep features. Chopra et al. [24] introduced the Siamese neural network to solve signature verification as an image matching problem. The Siamese neural network is made up of twin networks, and the parameters between them are shared. These two networks' distinct inputs are joined by an energy function in the end. Note that the Siamese network effectively mitigates the issue of imbalanced samples, making it particularly suitable in the current one-shot environment, which is suitable for the present case. The energy function computes some metrics between the highest-level feature representation on each side. As Figure 3 shows, we feed $x$ and $\hat{x}$ into the encoder network $E$ to map them to the feature vectors in the feature space, and we determine whether the two samples are similar by calculating the distance between the two feature vectors.

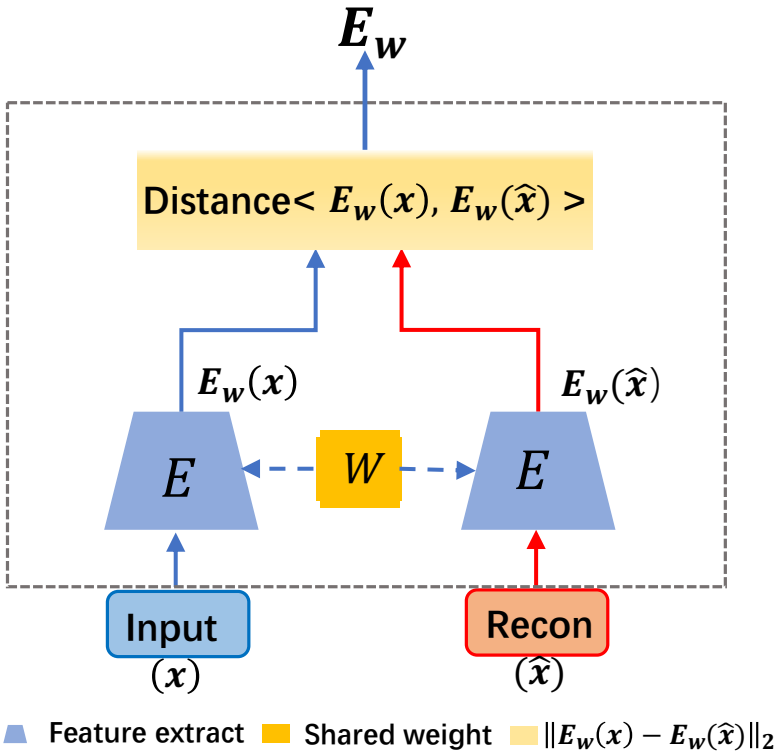



**Figure 3.** Mutual information Siamese network. Its purpose is to measure the similarity between the reconstructed image $\hat{x}$ and the original image $x$. $E$ is the mutual information feature extractor, which extracts the features of $x$ and $\hat{x}$. Then, we measure the distance of $E_w(x)$ and $E_w(\hat{x})$.

In one-class novelty detection, only samples from the normal class can be provided for training. Therefore, we lack the label when training with the Siamese neural network because we do not have an accurate measure of the reconstructed images of input samples. Thus, we train the network to extract the features of the training target in the input space without supervision.

**Mutual Information Maximization Loss:** Based on [31], we use a training encoder to maximize the mutual information between original samples and reconstructed samples, so that the features of reconstructed ones can better represent the features of input ones.

Let $X$ denote the set of original images, and $x \in X$ is an original sample. $Z$ is the combination of encoder vectors. $z \in Z$ is the encoding vector corresponding to $x$. $p(z|x)$ denotes the distribution of coding vectors produced by $x$, which estimates the mutual information by training the discriminator to distinguish between samples coming from the joint $\mathbb{J}$, and the product of marginals $\mathbb{M}$. $I(X, Z)$ denotes the mutual information between $x$ and $z$ and indicates the correlation between them. We hope that the encoder $E$ extracts $X$ more discriminative features $Z$ as much as possible, which lets $I(X, Z)$ be as large as possible:

$$p(z,x) = \arg\max_{p(z,x)} I(X,Z) \tag{6}$$

Following the formulation provided in Hjelm et al. [31], we take Jensen–Shannon divergence (JSD) as the mutual information estimation. Therefore, the loss is defined as

$$\begin{aligned}\mathcal{L}_{info} &= \max \widehat{\mathcal{I}}^{(\mathrm{JSD})}(X;Z) \\ &= \max \mathbb{E}_{\mathbb{J}}[\log \sigma(T(x,z))] - \mathbb{E}_{\mathbb{M}}\left[\log \sigma\left(T\left(x',z\right)\right)\right],\end{aligned} \tag{7}$$

where $x$ is an input sample, $z$ is the high-level representation related to $x$, and $x'$ is another input sample unrelated to $z$.

We also perform an a priori constraint on the encoder to give the encoder an a priori specific expected statistical feature. Here, we use a variational auto-encoder (VAE) to constrain this a priori information. $q(z)$ is a standard normal distribution as

$$\mathcal{L}_{pri} = \min \mathbb{E}_{x\sim\tilde{p}(x)}[KL(p(z|x)\|q(z))]. \tag{8}$$

Thus, we arrive at our complete objective for mutual information feature extraction as

$$\mathcal{L}_E = \min_{p(z|x)} \left\{ -\beta \cdot \mathcal{L}_{info} + \gamma \cdot \mathcal{L}_{pri} \right\}. \tag{9}$$

**Similarity Metric Loss:** We directly take the reconstructed image from RNet as one of the inputs in MNet. However, since we did not discriminate between the reconstructed images during the training process, samples with large reconstruction errors can degrade the performance of MNet. To tackle this problem, we use Huberś loss [32] as the similarity metric loss to reduce the penalty for samples with large reconstruction errors. In our method, the loss is defined as

$$\begin{aligned}\mathcal{L}_{metric} &= \min L_\delta(E_w(x), E_w(\hat{x})) \\ &= \begin{cases} \frac{1}{2}d^2, & \text{for} |d| \le \delta \\ \delta \cdot (|d| - \delta/2), & \text{otherwise} \end{cases}\end{aligned} \tag{10}$$

where $d = (E_w(x) - E_w(\hat{x}))$.

As shown in Equation (10), when the predicted value is less than $\delta$, Huberś loss becomes a square error or it becomes a linear error, which reduces the penalties on samples with large reconstruction errors to reduce the effect caused by them. Therefore, Huber's loss enhances the robustness of our model.

To summarize, the loss function of MNet is a linear combination of the mutual information maximization loss and the similarity metrics loss.

$$\mathcal{L}_{MNet} = \mathcal{L}_E + \alpha \mathcal{L}_{metric} \tag{11}$$

where $\alpha$ is the weighting factor of the similarity metric loss.

## 4. Discussion

In this section, we deliberate on the restrictions and validity of our proposed solution. The Maximum Mean Discrepancy (MMD) technique requires the presence of comparable distributions that can be effectively mapped to Hilbert space and are amenable to kernel methods. The Siamese Network, while proficient in measuring similarity and distance, displays dependence on the assigned task. Consequently, the network's efficacy may fluctuate depending on the particular undertaking it is employed for. However, the prerequisites for the Siamese network to exhibit commendable performance are satisfactorily fulfilled within the context of this study.

It is important to note that the proposed methods, such as MMD and the L1-norm for sparsity promotion, demonstrate considerable robustness and can be applied to diverse scenarios. The use of empirical means in MMD avoids the need for a priori knowledge of the data distribution, making it suitable for scenarios with unknown image distributions. The L1-norm encourages parameter sparsity, aiding in the extraction of essential features. The choice of the Siamese network helps to mitigate the imbalance issue typically found in one-shot learning environments. This network architecture is less sensitive to sample imbalance, which is beneficial for the novelty detection problem discussed.

The primary focus of this work is image background detection, which involves decomposing motion images into frames for processing. This allows for minimal requirements on the dynamism of the environment, making the proposed approach adaptable to varying levels of motion or scene changes.

## 5. Experiment

In this section, we validate the effectiveness of our method on two publicly available multi-class object recognition datasets.

### *5.1. Dataset and Measurement Metric*

#### 5.1.1. Dataset Description

**CIFAR-10:** CIFAR-10 [33] contains 60,000 color images in 10 classes with the resolution of 32×32. CIFAR-10 contains objects in the real world. CIFAR-10 is not only highly noisy, but also varies in terms of the appearance and scale of target objects, which brings great difficulties to identification. Therefore, several methods of one-class novelty detection obtain relatively poor results on CIFAR-10. In our experiment, we choose one of the classes as the normal class, and samples from the remaining classes are taken as novelty classes; 90% of samples from the normal class are used for training. The remaining 10% of samples from the normal class and samples randomly selected from the novelty classes with a proportion ranging between 10% and 50% are used for testing.

**Caltech-256:** Caltech-256 [34] is a dataset for image object recognition containing 30,608 images in 256 classes. The number of images per class ranges between 80 and 827. In Caltech-256, there is also an additional class called "clutter" that contains 827 images that are considered as outliers. Similar to the previous work [1,35], we randomly select $n$ classes as normal classes, where $n \in \{1, 3, 5\}$. If there are enough images in the selected class, we use first 150 images, or we use the entirety of the images in it. At the testing stage, we randomly select a certain number of images from the "clutter" class as novelty samples with a proportion of 50%.

#### 5.1.2. Measurement Metric

ROC analysis provides tools to select possibly optimal models independently of (and prior to specifying) the cost context or the class distribution. The receiver operating characteristic (ROC) curve is the plot of the true positive rate against the false positive rate, at various threshold settings. The area under the curve (AUC) represents the overall performance of the model ("1" means perfect and "0.5" means uselessness) and is a measure of aggregated classification performance. Hence, we evaluate these methods by the area under the curve (AUC) of the ROC.

### *5.2. Training Strategy*

The detailed training strategy is as follows: (1) We train RNet in an alternative way first. Specifically, we first optimize $D$. After this, we fix the parameters of $D$ and switch to optimize $E$ and $G$. We also use the mutual information loss to update the feature extraction subnetwork of $E_f$. (2) We train MNet after RNet converges. We use the mutual information loss and Huber's loss to optimize $E_f$. We apply a novel self-looping training trick to optimize the training process. In detail, we take the reconstructed image as an input in the metric network. The trick enhances the robustness of the reconstruction network.

### 5.3. Novelty Detection Result

**Result on CIFAR-10:** We compare our method with several conventional deep-learning-based methods for general anomaly detection as baselines, including one-class SVM (OC-SVM) [12], kernel density estimation (KDE) [36], and deep variational auto-encoder (VAE) [37].

Table 1 shows the experimental results on CIFAR-10, where the samples are complicated. Among the baseline methods, the performance of OCGAN is [28] considerably greater than that of the others. The method that we propose has comparable performance to OCGAN, with an average AUC (see Section 5.1.2) of 73.48%, as shown in Table 2(d).

**Table 1.** Average AUCs in % one-class novelty detection for CIFAR-10 dataset. The top three results are marked in **bold**.

| Normal Class | OC-SVM | KDE | VAE | GAN | AnoGAN | DSVDD | OCGAN | Ours (MSE) | Ours (MNet) |
|---|---|---|---|---|---|---|---|---|---|
| AIRPLANE | 61.6 | 61.2 | 70.0 | 70.8 | 67.1 | 61.7 | **75.7** | **89.6** | **86.7** |
| AUTOMOBILE | 63.8 | **64.0** | 38.6 | 45.8 | 54.7 | **65.9** | 53.3 | 47.3 | **70.5** |
| BIRD | 50.0 | 50.1 | **67.9** | **66.4** | 52.9 | 50.8 | 64.0 | **71.5** | 63.4 |
| CAT | 55.9 | 56.4 | 53.5 | 51.0 | 54.5 | 59.1 | **62.0** | **62.1** | **60.1** |
| DEER | 66.0 | 66.2 | **74.8** | 72.2 | 65.1 | 60.9 | 72.3 | **83.5** | **78.4** |
| DOG | **62.4** | **62.4** | 52.3 | 50.5 | 60.3 | **65.7** | 62.0 | 61.2 | **64.7** |
| FROG | 74.7 | **74.9** | 68.7 | 70.7 | 58.5 | 67.7 | 72.3 | **88.8** | **86.2** |
| HORSE | **62.6** | **62.6** | 49.3 | 47.1 | 62.5 | **67.3** | 57.5 | 56.0 | **65.7** |
| SHIP | 74.9 | 75.1 | 69.6 | 71.3 | 75.8 | 75.9 | **82.0** | **89.5** | **82.4** |
| TRUCK | **75.9** | **76.0** | 38.6 | 45.8 | 66.5 | 73.1 | 55.4 | 50.8 | **76.8** |

**Table 2.** Ablation experiment for our method performed on CIFAR-10.

| Experiment | Average AUC % |
|---|---|
| (a) RNet + self-looping | 70.03 |
| (b) MNet + self-looping | 68.41 |
| (c) RNet + MNet | 71.73 |
| (d) RNet + MNet +self-looping | 73.48 |

**Result on Caltech-256:** Table 3 shows the results on Caltech-256. We compare our method with several previous methods designed specifically for detecting novelties, including Coherence Pursuit (CoP) [38], OutlierPursuit [39], REAPER [40], Dual Principal Component Pursuit (DPCP) [41], Low-Rank Representation (LRR) [42], OutRank [43], R-graph [35], and ALOCC [1]. The results show that our method has better performance on Caltech-256.

**Table 3.** Results on Caltech-256. Normal samples are images from $n$ randomly selected classes, and novelty samples are randomly selected from the "clutter" class. The top three results are marked in **bold**.

| | CoP | REAPER | OutlierPursuit | LRR | DPCP | R-Graph | ALOCC | Ours (MSE) | Ours (MNet) |
|---|---|---|---|---|---|---|---|---|---|
| $AUC_1$ | 90.5 | 81.6 | 83.7 | 90.7 | 78.3 | **94.8** | **94.2** | **94.2** | **94.1** |
| $AUC_3$ | 67.6 | 79.6 | 78.8 | 47.9 | 79.8 | 92.9 | **93.8** | **93.8** | **94.1** |
| $AUC_5$ | 48.7 | 65.7 | 62.9 | 33.7 | 67.6 | **91.3** | **92.3** | 90.8 | **92.7** |

### 5.4. Ablation Study

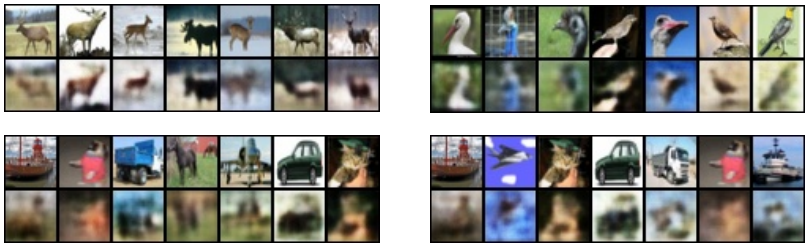
In order to investigate the effectiveness of each subnetwork of our method, we conduct several further ablation studies.

**How important is MNet?** We firstly use MSE instead of MNet to measure the reconstruction error. The penultimate columns of Tables 1 and 2(a) show the results of the method without MNet. The result shown in Table 2(a) is also slightly greater than the one of OCGAN but is inferior to ours. We conjecture that the reason is that the metric network extracts discriminative characteristics of the target, especially for objects with high complexity.

However, the comparison between the last two columns of Table 1 indicates that MSE is superior to MNet when some classes (e.g., deer, birds) are selected as the normal classes and MNet performs better when others (e.g., car, truck) are. We conduct an analysis on this result. First, we calculate the variance of each sample by class. The variance represents the dispersion level of the data. The larger the variance, the more dispersed the data distribution is. As can be seen from Tables 1 and 4, for the three classes with the most discrete image distribution, TRUCK, AUTOMOBILE, and HORSE, our method with the MNet function outperforms the state-of-the-art methods. However, the results obtained using the MSE function are reduced by 33.2%, 28.2%, and 16.8% compared to the state-of-the-art methods. At the same time, for the other seven classes with a relatively concentrated image distribution, our method can achieve very good results by applying both the MSE function and MNet function. Therefore, compared with the MSE function, our proposed MNet function can more stably achieve excellent results on different discrete-level datasets. Some different discrete-level samples are shown in Figure 4.

**Table 4.** Variance of each class on CIFAR-10.

| Normal Class | R-Channel | G-Channel | B-Channel |
|---|---|---|---|
| airplane | 0.500 | 0.481 | 0.531 |
| automobile | 0.536 | 0.531 | 0.548 |
| bird | 0.454 | 0.441 | 0.486 |
| cat | 0.513 | 0.504 | 0.515 |
| deer | 0.434 | 0.413 | 0.423 |
| dog | 0.500 | 0.487 | 0.497 |
| frog | 0.457 | 0.437 | 0.440 |
| horse | 0.486 | 0.487 | 0.503 |
| ship | 0.499 | 0.481 | 0.502 |
| truck | 0.536 | 0.538 | 0.562 |

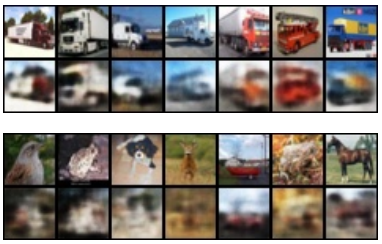


(a) The normal class is Deer



(b) The normal class is Bird

(c) The normal class is Car



(d) The normal class is Truck

**Figure 4.** Reconstructed images of classes. In each group of images, a, b, c, and d, the first row is the normal sample and its reconstruction sample, and the second row is the corresponding novelty sample

and its reconstruction sample. In (**a**,**b**), the variance of normal samples is small and the distributions' discrete levels are lowest. In (**c**,**d**), the variance of novelty samples is large and the distributions' discrete levels are highest. The reconstructed image of the normal class sample in the first row of each group also shows that our method can reconstruct the background.

Meanwhile, we also conduct experiments aimed at adjusting the number of selected normal classes on Caltech-256. We randomly select $n \in \{1, 3, 5, 7, 9\}$ classes as normal classes. During testing, we randomly select a certain number of images from the "clutter" category as novelty samples, with a proportion of 50%. As shown in the Figure 5, as the amount of classes increases, since the data distribution becomes more dispersed, MNet performs better than MSE.

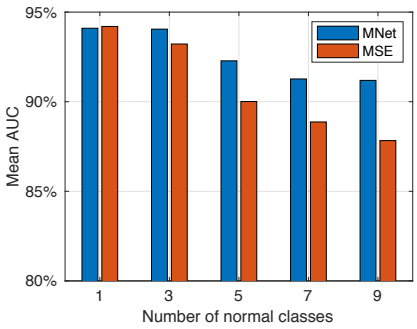



**Figure 5.** Mean AUC in % for one-class novelty detection. $n \in \{1, 3, 5, 7, 9\}$ classes are randomly selected as normal classes from Caltech-256.

**How important is RNet?** We replace the image reconstruction network with the auto-encoder and GAN from the previous work and use our proposed work metric network part to detect the novelty samples. Table 2(b) shows the results of our method without the image reconstruction network. The last two columns in Table 1 also show the results of using our reconstruction network. The experimental results show that the method of using MMD loss can improve the detection results. MMD loss minimizes the distance between the reconstructed sample and the original sample distribution. Specifically, the two samples are closer together in terms of statistics. The background can also be reconstructed to some extent for images with background. Then, in the process of novel class detection, we can not only use the characteristics of the target itself, but also use global context information to assist detection. Figure 4 shows the results of using RNet to reconstruct the sample. It can be seen that the background is reconstructed to some extent in the image. We believe that our approach focuses on background information, so our results show good performance.

**How important is self-looping training?** We do not use the above training trick. The results of our training without the self-looping trick are shown in Table 2(c). Compared with our overall experiment, using this method can improve the accuracy of novelty detection. We feed reconstructed images into the network for training, which takes advantage of the ability to generate network for data augmentation. Thus, we find that these data are enhanced by the advantages of the network itself.

## 6. Conclusions

In this paper, we present a novel and sophisticated network architecture for complex background reconstruction, specifically tailored to the task of one-class novelty detection. We highlight the inherent correlation between complex backgrounds and the foreground object, as well as the rich information encapsulated within the foreground. As a result, we assert that harnessing the power of background reconstruction significantly enhances the detection of novel samples. To achieve this, we employ the MMD loss function to effectively mitigate background interference and we introduce MNet, a metric to measure the background similarity to minimize their divergence. The experiments show that our

method exhibits an enhancement of approximately 0.3% in both the CIFAR-10 and Caltech-256 datasets.

**Author Contributions:** Conceptualization and investigation, K.Z.; methodology and writing—original draft preparation, M.S.; software and data curation, R.A.; Visualization and validation, H.H.; supervision and writing—review and editing, Z.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by NSFC, grant number 62072367.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** We leveraged two open datasets for evaluation, named CIFAR-10 and Caltech-256. They can be downloaded from http://www.cs.toronto.edu/~kriz/cifar-10-python.tar.gz (accessed on 25 December 2022) and https://www.kaggle.com/datasets/jessicali9530/caltech256 (accessed on 25 December 2022), respectively.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| OC-SVM | one-class support vector machine |
| MMD | Maximum Mean Discrepancy |
| RNet | reconstruction network |
| MNet | metric network |
| GMMN | generative moment matching network |
| VAE | variational auto-encoder |
| AUC | area under the curve |
| ROC | receiver operating characteristic |
| DPCP | Dual Principal Component Pursuit |

## References

1. Sabokrou, M.; Khalooei, M.; Fathy, M.; Adeli, E. Adversarially Learned One-Class Classifier for Novelty Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3379–3388.
2. Schlegl, T.; Seeböck, P.; Waldstein, S.M.; Schmidt-Erfurth, U.; Langs, G. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In Proceedings of the International Conference on Information Processing in Medical Imaging, Boone, NC, USA, 25–30 June 2017; Springer: Berlin/Heidelberg, Germany, 2017; pp. 146–157.
3. Xia, X.; Pan, X.; Li, N.; He, X.; Ma, L.; Zhang, X.; Ding, N. GAN-based anomaly detection: A review. *Neurocomputing* **2022**, *493*, 497–535. [CrossRef]
4. Javaid, A.; Niyaz, Q.; Sun, W.; Alam, M. A deep learning approach for network intrusion detection system. In Proceedings of the 9th EAI International Conference on Bio-Inspired Information and Communications Technologies (formerly BIONETICS), New York, NY, USA, 3–5 December 2016; ICST (Institute for Computer Sciences, Social-Informatics and Technology: Budapest, Hungary, 2016; pp. 21–26.
5. Min, S.; Lee, B.; Yoon, S. Deep learning in bioinformatics. *Briefings Bioinform.* **2017**, *18*, 851–869. [CrossRef] [PubMed]
6. Bandaragoda, T.R.; Ting, K.M.; Albrecht, D.; Liu, F.T.; Zhu, Y.; Wells, J.R. Isolation-based anomaly detection using nearest-neighbor ensembles. *Comput. Intell.* **2018**, *34*, 968–998. [CrossRef]
7. Ellouze, A.; Ksantini, M.; Delmotte, F.; Karray, M. Single Object Tracking Applied to an Aircraft. In Proceedings of the 2018 15th International Multi-Conference on Systems, Signals & Devices (SSD), Yasmine Hammamet, Tunisia, 19–22 March 2018; pp. 1441–1446.
8. Triki, N.; Ksantini, M.; Karray, M. Traffic Sign Recognition System based on Belief Functions Theory. In Proceedings of the 13th International Conference on Agents and Artificial Intelligence (ICAART 2021), Online, 4–6 February 2021; Volume 2, pp. 775–780.
9. Yerima, S.Y.; Bashar, A. Semi-supervised novelty detection with one class SVM for SMS spam detection. In Proceedings of the 2022 29th International Conference on Systems, Signals and Image Processing (IWSSIP), IEEE, Sofia, Bulgaria, 1–3 June 2022; pp. 1–4.
10. Almohsen, R.; Keaton, M.R.; Adjeroh, D.A.; Doretto, G. Generative probabilistic novelty detection with isometric adversarial autoencoders. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–23 June 2022; pp. 2003–2013.

11. Li, C.L.; Chang, W.C.; Cheng, Y.; Yang, Y.; Póczos, B. Mmd gan: Towards deeper understanding of moment matching network. In *Proceedings of the Thirty-First Conference on Neural Information Processing Systems NIPS 2017, Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017*; pp. 2203–2213.
12. Schölkopf, B.; Platt, J.C.; Shawe-Taylor, J.; Smola, A.J.; Williamson, R.C. Estimating the support of a high-dimensional distribution. *Neural Comput.* **2001**, *13*, 1443–1471. [CrossRef] [PubMed]
13. Hoffmann, H. Kernel PCA for novelty detection. *Pattern Recognit.* **2007**, *40*, 863–874. [CrossRef]
14. Sakurada, M.; Yairi, T. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis, Gold Coast, QLD, Australia, 2 December 2014; ACM: New York, NY, USA, 2014; p. 4.
15. Gruhl, C.; Sick, B.; Tomforde, S. Novelty detection in continuously changing environments. *Future Gener. Comput. Syst.* **2021**, *114*, 138–154. [CrossRef]
16. Ruff, L.; Vandermeulen, R.; Goernitz, N.; Deecke, L.; Siddiqui, S.A.; Binder, A.; Müller, E.; Kloft, M. Deep one-class classification. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 4393–4402.
17. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the 28th Annual Conference on Neural Information Processing Systems (2014NIPS), Advances in Neural Information Processing Systems, Red Hook, NY, USA, 8–13 December 2014; pp. 2672–2680.
18. Deecke, L.; Vandermeulen, R.; Ruff, L.; Mandt, S.; Kloft, M. Anomaly detection with generative adversarial networks. *arXiv* **2018**, arXiv:1809.04758.
19. Li, Y.; Shi, Z.; Liu, C.; Tian, W.; Kong, Z.; Williams, C.B. Augmented time regularized generative adversarial network (atr-gan) for data augmentation in online process anomaly detection. *IEEE Trans. Autom. Sci. Eng.* **2021**, *19*, 3338–3355. [CrossRef]
20. Chen, D.; Yue, L.; Chang, X.; Xu, M.; Jia, T. NM-GAN: Noise-modulated generative adversarial network for video anomaly detection. *Pattern Recognit.* **2021**, *116*, 107969. [CrossRef]
21. Schölkopf, B.; Smola, A.J.; Bach, F. *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*; MIT Press: Cambridge, MA, USA, 2002.
22. Li, Y.; Swersky, K.; Zemel, R. Generative moment matching networks. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 1718–1727.
23. Dziugaite, G.K.; Roy, D.M.; Ghahramani, Z. Training generative neural networks via maximum mean discrepancy optimization. *arXiv* **2015**, arXiv:1505.03906.
24. Chopra, S.; Hadsell, R.; LeCun, Y. Learning a similarity metric discriminatively, with application to face verification. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; pp. 539–546.
25. Hadsell, R.; Chopra, S.; LeCun, Y. Dimensionality reduction by learning an invariant mapping. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), IEEE, New York, NY, USA, 17–22 June 2006; Volume 2, pp. 1735–1742.
26. Hoffer, E.; Ailon, N. Deep metric learning using triplet network. In Proceedings of the International Workshop on Similarity-Based Pattern Recognition, Copenhagen, Denmark, 12–14 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 84–92.
27. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 815–823.
28. Perera, P.; Nallapati, R.; Xiang, B. Ocgan: One-class novelty detection using gans with constrained latent representations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2898–2906.
29. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved techniques for training gans. In Proceedings of the 30th Annual Conference on Neural Information Processing Systems, Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 2234–2242.
30. Gneiting, T.; Raftery, A.E. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **2007**, *102*, 359–378. [CrossRef]
31. Hjelm, R.D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Trischler, A.; Bengio, Y. Learning deep representations by mutual information estimation and maximization. *arXiv* **2018**, arXiv:1808.06670.
32. Huber, P.J. *Robust Statistics*; Springer: Berlin/Heidelberg, Germany, 2011.
33. Krizhevsky, A.; Hinton, G. *Learning Multiple Layers of Features From Tiny Images; Technical Report; Citeseer: Toronto, ON, Canada, 2009*.
34. Griffin, G.; Holub, A.; Perona, P. *Caltech-256 Object Category Dataset*; California Institute of Technology: Pasadena, CA, USA, 2007.
35. You, C.; Robinson, D.P.; Vidal, R. Provable self-representation based outlier detection in a union of subspaces. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3395–3404.
36. Parzen, E. On estimation of a probability density function and mode. *Ann. Math. Stat.* **1962**, *33*, 1065–1076. [CrossRef]
37. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
38. Rahmani, M.; Atia, G.K. Coherence pursuit: Fast, simple, and robust principal component analysis. *IEEE Trans. Signal Process.* **2017**, *65*, 6260–6275. [CrossRef]

39. Xu, H.; Caramanis, C.; Sanghavi, S. Robust PCA via outlier pursuit. In Proceedings of the 24th Annual Conference on Neural Information Processing Systems, Advances in Neural Information Processing Systems 23 (NIPS 2010), Vancouver, BC, Canada, 6–9 December 2010; pp. 2496–2504.
40. Lerman, G.; McCoy, M.B.; Tropp, J.A.; Zhang, T. Robust computation of linear models by convex relaxation. *Found. Comput. Math.* **2015**, *15*, 363–410. [CrossRef]
41. Tsakiris, M.C.; Vidal, R. Dual principal component pursuit. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 7–13 December 2015; pp. 10–18.
42. Liu, G.; Lin, Z.; Yu, Y. Robust subspace segmentation by low-rank representation. In Proceedings of the roceedings of the 27th International Conference on Machine Learning, Haifa, Israel, 21–24 June 2010; Volume 1; p. 8.
43. Moonesignhe, H.; Tan, P.N. Outlier detection using random walks. In Proceedings of the 2006 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06), Arlington, VA, USA, 13–15 November 2006; IEEE: Piscataway, NJ, USA, 2006; pp. 532–539.