One-Class Learned Encoder-Decoder Network with Adversarial Context Masking for Novelty Detection

John Taylor Jewell
Western University
London, ON, Canada
jjewell6@uwo.ca

Vahid Reza Khazaie
Western University
London, ON, Canada
vkhazaie@uwo.ca

Yalda Mohsenzadeh
Western University
London, ON, Canada
ymohsenz@uwo.ca

Abstract

Novelty detection is the task of recognizing samples that do not belong to the distribution of the target class. During training, the novelty class is absent, preventing the use of traditional classification approaches. Deep autoencoders have been widely used as a base of many novelty detection methods. In particular, context autoencoders have been successful in the novelty detection task because of the more effective representations they learn by reconstructing original images from randomly masked images. However, a significant drawback of context autoencoders is that random masking fails to consistently cover important structures of the input image, leading to suboptimal representations - especially for the novelty detection task. In this paper, to optimize input masking, we introduce a Mask Module that learns to generate optimal masks and a Reconstructor that aims to reconstruct masked images. The networks are trained in an adversarial setting in which the Mask Module seeks to maximize the reconstruction error that the Reconstructor is minimizing. When applied to novelty detection, the proposed approach learns semantically richer representations compared to context autoencoders and enhances novelty detection at test time through more optimal masking. Novelty detection experiments on the MNIST and CIFAR-10 image datasets demonstrate the proposed approach's superiority over cutting-edge methods. In a further experiment on the UCSD video dataset for novelty detection, the proposed approach achieves a frame-level Area Under the Curve (AUC) of 99.02% and an Equal Error Rate (EER) of 5.4%, exceeding recent state-of-the-art models. Code available at https://github.com/jewelltaylor/OLED.

1. Introduction

Novelty detection involves determining whether or not an unknown sample belongs to the distribution of the training data. In the case the sample is similar to the training data, it is referred to as an inlier or normal sample. Alternatively, if the sample does not follow the distribution defined in the training examples, it is referred to as an outlier or anomaly. Novelty detection differs from other machine learning tasks in that the outlier class is poorly sampled or nonexistent. Due to the unavailability of outlier samples, traditional classification approaches are not suitable.

Within computer vision, novelty detection is ubiquitous with subtasks that have widespread applications such as marker discovery in biomedical data [1] and video surveillance [2]. Anomaly detection in images is one such task that involves identifying whether an image is an inlier or an outlier based on training data that mostly consists of inlier images. To compensate for the unavailability of outlier samples, one-class classification methods aim to model the distribution of the inlier data [3]. New samples that do not conform to the target distribution are considered outliers. However, it is often hard to model the distribution of image data with conventional methods because of the high dimensionality in which the data points reside [3].

With the advent of deep learning, methods have been proposed that are able to effectively produce representations for high dimensional data [4]. Autoencoders (AE) are an unsupervised class of approaches that are well suited for modeling image data [5]. At a high level, an AE consists of two modules: an encoder and a decoder. The encoder learns a mapping from an image to a lower-dimensional latent space, and the decoder learns a mapping from the latent space back to the original image. In this way, AEs are trained in an unsupervised manner by minimizing the error between the original image and the reconstruction.

As a powerful unsupervised method for learning representations, AEs are the basis of many one-class classification approaches [6]. To detect anomalous images, the AE is first trained on a set of primarily normal images. At test time, the reconstruction error of a sample is used as an anomaly score. The underlying intuition is that the reconstruction error will be lower for inlier samples than outlier samples [7]. This follows from the fact that the AE is

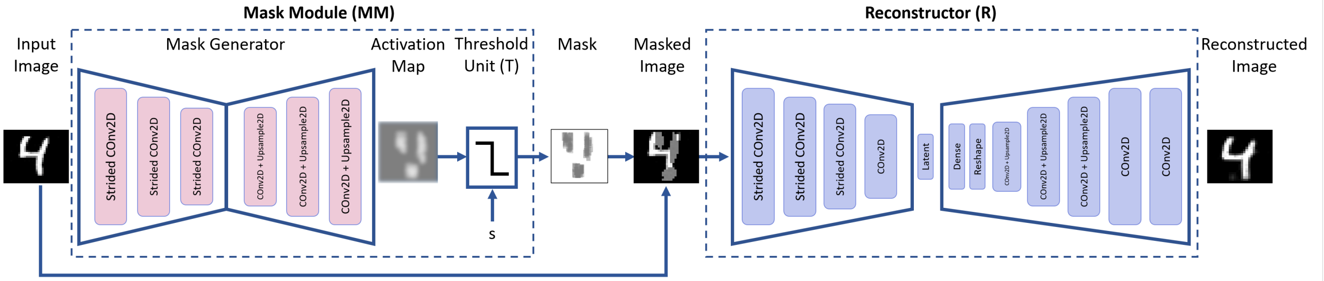

Figure 1. An overview of the architecture in OLED. The Mask Module adversarially learns to cover the important parts of the input image; it consists of an autoencoder that generates an activation map and a threshold unit to produce the binary mask. The Reconstructor aims to minimize the reconstruction error and the Mask Module aims to maximize the reconstruction error.

trained solely on inlier samples. However, this assumption is often violated, and the AE generalizes well to construct images outside of the distribution of the training data [8, 9]. This is especially evident in cases where anomalous images share similar compositional patterns as inlier images.

Recent methods introduce additional complexity into the autoencoders reconstruction task so that outliers are not reconstructed well [10, 11, 12]. To this end, denoising autoencoders (DAE) have been used. DAEs learn to reconstruct unperturbed images from images that have been perturbed by noise [13]. Beyond yielding more robust representations, the denoising task of the AE has been shown to induce a reconstruction error that approximates the local derivative of the log-density with respect to the input [14]. Thus, a sample's global reconstruction error reflects the norm of the derivative of the log-density with respect to the input. In this way, DAEs provide a more interpretable and theoretically grounded anomaly score.

Context autoencoders (CAE) [15], a specific type of DAE, have shown strong performance in the anomaly detection task [12, 16]. Instead of being perturbed by noise, input images are subjected to random masking. Consequently, the CAE learns to inpaint a randomly masked region of the input image in conjunction with the reconstruction task. This random masking is similar to using salt-and-pepper noise, which has been shown to yield better representations by implicitly enforcing the AE to learn semantic information about the distribution of the training data [15]. Despite these strengths, in some cases CAEs suffer from suboptimal representations leading to poor performance in the anomaly detection task.

Inspired by the drawbacks of CAEs [15], we proposed One-Class Learned Encoder-Decoder (OLED) Network with Adversarial Context Masking. OLED introduces a Mask Module MM that produces masks applied to images input into the Reconstructor R. The masks generated by MM are optimized to cover the most important parts of the input image, resulting in a comparable reconstruction score across samples. The underlying intuition is that the loss of the masked region will be low in the case of inlier images and high in the case of outlier images. This stems from the fact that the Reconstructor learns to inpaint masked regions using mostly inlier samples. Thus, the inpainted regions of outlier images will consist of patterns present in the inlier images, yielding a high reconstruction error.

At a high level, the Mask Module is a convolutional autoencoder, and the Reconstructor is a convolutional encoder-decoder. They are trained in an adversarial manner, where the Mask Module is trying to generate masks that yield higher reconstruction errors, and the Reconstructor is trying to minimize the reconstruction error of the masked image. The architecture of the proposed approach is shown in Figure 1. We applied OLED to several benchmark datasets for anomaly detection in addition to providing a formal analysis of the efficacy of the Mask Module. Experimental results demonstrate that OLED is able to outperform a variety of recent state-of-the-art methods and hints at the broader usefulness of the mask module in other core computer vision tasks. In this paper our contributions are the following:

- We proposed a novel approach for finding the most important parts of images for novelty detection.
- Our framework is optimized through adversarial setting which yields more efficient representations for novelty detection.
- Our method provides several anomaly scores which capture different aspects of normality
- Due to effectiveness of our method in masking important parts of the image, we can leverage it at the test time which yields better anomaly scores.

2. Related Works

One-class classification is primarily associated with the domain of novelty, outlier, and anomaly detection. In these types of problems, a model attempts to capture the distribution of the inlier class to finally detect the unknown outliers or novel concepts. The conventional methods in the anomaly detection field utilized one-class SVM [17, 18] and Principal Component Analysis (PCA) and its variations [19, 20] to find a subspace that best represents the distribution of normal samples. Unsupervised clustering techniques like k-means [3] and Gaussian Mixture Models (GMM) [21] also have been used to formulate the distribution of normal data for identifying the anomalies, but they normally fail in dealing with high-dimensional data.

Several other proposed methods benefit from self-representation learning, such as reconstruction-based approaches. They usually rely on the hypothesis that the outlier samples cannot be reconstructed precisely by a model that only learned the distribution of inlier samples. For example, Cong et al. [22] suggested a model for video anomaly and outlier detection by learning sparse representations for distinguishing between inlier and outlier samples. In [23, 24], test samples are reconstructed using the representations learned from inlier samples, and the reconstruction error is employed as a metric for novelty detection. Most of the deep learning-based models with encoder-decoder architecture [25, 26, 27, 8, 28] also used this score to detect anomalies. Although effective, these methods are limited by the under-designed representation of their latent space. Gong et al. [9] proposed a deep autoencoder augmented with a memory module to encode the input to a latent space with the encoder. The resulting latent vector is used as a query to retrieve the most relevant memory item for reconstruction with the decoder.

In [1], a deep convolutional generative adversarial network (GAN) is leveraged to learn a manifold of normal images with a novel anomaly score based on the mapping from image space to a random distribution. Sabokrou et al. [29] took advantage of Generative Adversarial Networks (GAN) [30] along with denoising autoencoders to use the discriminator's score for the reconstructed images for the novelty detection task. Zaheer et al. [31] redefined the adversarial one-class classifier training setup by modifying the role of the discriminator to distinguish between good and bad quality reconstructions and improved the results even further. Perera et al. used denoising auto-encoder networks to enforce the normal samples to be distributed uniformly across the latent space [11]. Abati et al. suggested a deep autoencoder model with a parametric density estimator that learns the probability distribution underlying its latent representations through an autoregressive procedure [10].

Some recent works [32], [33] have tried to leverage pretrained deep neural networks by distilling the knowledge. In [32], they utilized a VGG16 [34] to compute a multilevel loss for training the student network to calculate the anomaly score and perform anomaly segmentation. Even though these methods could achieve high performance, they benefit from the knowledge attained by training on millions of labeled images and also may not work well on other modalities of data. As our proposed method does not leverage pretrained networks, we consider our work complimentary, and thus do not compare against this class of approaches.

3. Method

3.1. Motivation

Previous works have demonstrated that the reconstruction error of an Autoencoder (AE) acts as a good indicator of whether or not a sample conforms to the distribution defined in the training examples [7]. As such, AEs are commonly used for anomaly detection. To this end, Denoising Auotoencoders (DAE) have often been used because of the more robust representations they offer [14]. Context Autoencoders (CAEs), a subclass of DAEs, have been particularly successful in the anomaly detection task by offering representations that capture the semantics of the underlying training distribution [12].

However, CAEs have a number of disadvantages. The first drawback of CAEs is that they learn suboptimal representations by failing to consistently mask important parts of the image during training. Furthermore, they perform poorly at test time if they include random masking. This is because the mask placement is closely related to the reconstruction score. An outlier with a simple part of the image masked may have a lower reconstruction error than an inlier image with a difficult part of the image masked. Thus, random masking cannot be effectively used at test time for more robust anomaly detection. Conversely, our approach avoids these drawbacks by learning to mask intelligently. Experimental results from the ablation study in section 4.6 support this conclusion. In order to mitigate these shortcomings while leveraging the benefits offered by CAEs, we propose a One-Class Learned Encoder-Decoder Network with adversarial context masking, which we call OLED.

3.2. Overview

Our proposed framework, OLED, consists of two modules: the Reconstructor R and the Mask Module MM. An overview of the architecture is available in 1. R and MM are trained in an adversarial manner, where R seeks to reconstruct images that have been covered by masks generated by MM. Masks have the same spatial resolution as input images with a single channel of 0 or 1 activations. As such, a masked image is easily obtained by taking an element-wise product of an image and its corresponding mask.

Through the adversarial training process, R learns representations that encode semantic information of the training distribution through the inpainting task. Alternatively, MM learns to mask the most important parts of the input image by maximizing the reconstruction error of R. At test time, new samples are subjected to masks generated by MM and fed to R where the reconstruction error is used as an anomaly score. Accordingly, the reconstruction error will be low for the inlier class because R is trained to reconstruct and inpaint inlier samples. However, in the case of anomalies, the reconstruction error will be higher primarily. This stems from the fact that R learns to reconstruct and inpaint masked regions using mostly inlier samples.

3.3. Reconstructor

R is a convolutional encoder-decoder network that is trained to reconstruct masked images. Following some of the previous works [16], a dense bottleneck is used. The full connectivity of the dense layer is helpful for the inpainting task, especially for shallow networks with low receptive fields. Moreover, R does not include max-pooling layers for greater stability in training. To further promote stability, Leaky ReLU and batch normalization are used in each convolutional block. The values after the last convolution layer are clipped to in between -1 and 1.

3.4. Mask Module

MM consists of a mask generator MG followed by a threshold unit T that generates masks of the same resolution as the input image. These masks are applied to the corresponding input image prior to being fed into R. MM seeks to produce a mask that maximizes the reconstruction error of the input. In this way, it learns to mask the most optimal parts of the image. Thus, masks generated by MM yield more comparable anomaly scores across samples in contrast to random masking.

3.4.1 Mask Generator

MG is a convolutional autoencoder that takes an input image and generates a corresponding activation map. This activation map is input into the threshold unit to produce a binary mask. Similar to R, MG avoids the use of max pooling. Additionally, batch normalization and Leaky ReLU are used in each convolutional block, with the exception of the last convolution block that uses ReLU. In contrast to R, MG has a spatial bottleneck and contains much fewer parameters. This reflects the fact that R has a substantially more complex task than MG.

3.4.2 Threshold Unit

Activation maps generated by MG are input into T to generate a mask. T requires a threshold hyperparameter that determines what percentage of the pixels in the image will not be masked. In this way, the same amount of pixels are masked in each image, ensuring that the reconstruction errors are comparable between samples.

For each activation map, pixels with activations in the top 1 - t percent are set to 0. The final mask is obtained by setting the remaining activations to one. More formally, given an activation map A and a scalar s that represents the numeric value of the pixel with the t highest activations:

$$A_{ij} = \begin{cases} 0, & \text{if } A_{ij} \geq s \\ 1, & \text{otherwise} \end{cases} \tag{1}$$

As it stands, this is a discontinuous function, which is known to have less stable optimization. In order to eliminate this problem, the threshold operation is reformulated in terms of continuous ReLU activation function:

$$A_{ij} = \frac{max(A_{ij} * -1 + s, 0)}{max(A_{ij} * -1 + s, 0) + \epsilon} \tag{2}$$

where max(A_{ij}, 0) represents the ReLU activation, and ϵ is an infinitesimal positive scalar. The above formulation ensures continuity over the entire domain of the function enabling backpropagation through T into MG.

3.4.3 Masking Procedure

MG and T sequentially process an input image to create a mask. Masks generated by MM are single-channel binary images with the same spatial resolution as input images. The masked image is obtained by applying the mask to its corresponding image for each channel. More precisely, given an image x, the corresponding masked image x_m is defined as:

$$x_m = x \odot MM(x) \tag{3}$$

where $\odot$ denotes element-wise multiplication. In this way, activations in the mask that are 0 set the corresponding pixel in the input image to 0, otherwise the pixel remains unchanged. It is important to note that input images, and thus the reconstructions generated by R, are scaled between -1 and 1. Because of this, masked pixels are set to the midpoint of the color range.

3.5. Adversarial Training

Adversarial training is a learning mechanism in which two networks compete in a minmax game that iteratively enhances the ability to model the underlying distribution of the data. Following this intuition, Generative Adversarial Networks (GANs) [30] have been proposed and shown immense success in generating samples with similar distribution of the training data. In order to do so, a generator network G and discriminator network D are trained in this manner. G takes as input a noise vector and seeks to produce samples that follow the distribution of the training data. Alternatively, D takes as input real samples from the training set along with fake samples generated by G and seeks to discriminate between the two. More formally, given an image x sampled from p_{data} and a random latent vector z sampled from p_z the objective of a GAN is:

$$\min_G \max_D \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z))] \tag{4}$$

$G(z)$ is a sample generated by G with input z. $D(x)$ and $D(G(z))$ are the discriminator's classification scores for a real and generated sample, respectively.

Similarly, we train MM and R adversarially. MM seeks to generate masks that yield the highest reconstruction error from R. The total reconstruction error L_{tot} consists of an L2 loss of the masked image L_{mask}, contextual loss of the masked region L_{cont} and an L2 loss of an unperturbed image L_{recon}. Given an inlier image x and the corresponding masked image x_m, L_{mask}, L_{cont} and L_{recon} are defined as:

$$L_{mask} = \|x - R(x_m)\|^2 \tag{5}$$

$$L_{cont} = \|x_c - R(x_c)\| \tag{6}$$

$$L_{rec} = \|x - R(x)\|^2 \tag{7}$$

where x_c is the masked region of the input image and $R(x_c)$ is the reconstruction of the masked region. $R(x_m)$ denotes the reconstruction of the masked image x_m. $R(x)$ is the reconstruction of the intact image x with the Reconstructor. The following are the components of the objective:

- L_{mask}: Forces the network to form a semantic understanding of characteristic elements of inlier samples.
- L_{cont}: Emphasizes that the masked region of the image is reconstructed properly to avoid blurry reconstructions of the masked region.
- L_{rec}: Helps the network learn the distribution of unmasked inliers.

As such, the objective function of OLED is given by:

$$\min_R \max_{MM} L_{mask} + \gamma L_{cont} + \lambda L_{rec} \tag{8}$$

where γ and λ are hyperparameters that weigh L_{cont} and L_{rec}, respectively. Since MM has no bearing on L_{rec}, it is not included in the error of MM.

3.6. Anomaly Scoring

The three distinct loss terms in the OLED objective present the opportunity for three anomaly scores to be defined: s_{mask}, s_{cont} and s_{rec}. s_{mask}, s_{cont} and s_{rec} are obtained through scaling L_{mask}, L_{cont} and L_{rec} between 0 and 1. By virtue of being derived from the respective losses, each anomaly score captures a different element of normality. s_{cont} and s_{mask} capture normality local to the masked region which tends to cover the most characteristic parts of the image. s_{rec} captures the global normality of the image, taking into account how good the entire reconstruction of the image is. s_{avg} is obtained by taking the average of s_{mask}, s_{cont} and s_{rec}.

4. Experiments

This section contains a detailed analysis of the proposed method, OLED. In particular, we evaluated OLED on three datasets that are benchmarks in the novelty/anomaly detection literature, and the results are compared to recent state-of-the-art methods. Additionally, we presented a formal analysis exploring the effectiveness of masks generated by MM.

4.1. Implementation Details

OLED is implemented in Python using the TensorFlow [35]. A detailed overview of the architecture of R and MM is available in Section 3.3 and Section 3.4, respectively. t, λ and γ are set to 87.5, 1 and 50 respectively. These hyperparameters were set based on experimentation and an ablation study showing the stability of the performance across different settings. The threshold parameter can be adjusted based on the difficulty of the dataset; where larger values of the threshold are more suitable for more complex datasets. The weights of the the loss function listed as the defaults are to balance out the effect of reconstruction losses since they are on different scales. R and MM use an Adam optimizer with a learning rate of $5e^{-4}$, b_1 of .5 and b_2 of .9. The networks are trained for 300 epochs. Following [29], a small validation set containing 150 samples from inliers and 150 samples from outliers from the training set is used to determine the best epoch to select models R and MM.

4.2. Datasets

The three datasets chosen for the experiments are MNIST [36], CIFAR-10 [37] and UCSD [38]. These particular datasets were chosen based on their popularity as benchmarks in the anomaly detection literature. The setups were chosen in a way that enables OLED to be compared to a variety of recent state-of-the-art methods.

MNIST: MNIST is a dataset that contains 60,000 images of handwritten digits from 0 to 9. Images in MNIST are grayscale with a resolution of 28 x 28.

Method	AUCROC
OCSVM [17]	0.9499
AE [5]	0.9619
VAE [40]	0.9643
PixCNN [41]	0.6141
DSEBM [26]	0.9554
MemAE [9]	0.9751
OLED (Ours) s_{rec}	**0.9772**
OLED (Ours) s_{mask}	**0.9851**
OLED (Ours) s_{cont}	0.9650
OLED (Ours) s_{avg}	**0.9845**

Table 1. Average AUCROC values on all 10 classes sampled from MNIST image dataset.

CIFAR-10: CIFAR-10 is a dataset that contains 60,000 natural images of objects from across ten classes. Images in CIFAR-10 are RGB with a resolution of 32 x 32. Similar to MNIST, CIFAR-10 is also used widely as a benchmark in the anomaly detection literature. However, CIFAR-10 presents more of a challenge because images differ substantially across classes, and the background of images are not aligned.

UCSD: This dataset [39] consists of two subsets (Ped1 and Ped2) with different outdoor scenes. Available objects in the frames are pedestrians, cars, skateboarders, wheelchairs, and bicycles. Pedestrians are dominant in nearly all frames and considered as the normal class, while other objects are anomalies. We assessed our method on Ped2, which includes 2,550 frames in 16 training and 2,010 frames in 12 test videos, all with a resolution of 240×360 pixels. Following [31], we calculated frame-level area under the receiver operating characteristic (AUCROC) and Equal Error Rate (EER) to evaluate performance and compare against both patch-based and full-frame setups.

4.3. Novelty Detection in Image Datasets

MNIST: OLED is evaluated on MNIST using the protocol defined in [9]. This protocol involves dividing the dataset into ten different anomaly detection datasets corresponding to the ten predefined classes in MNIST. In each anomaly detection dataset, the inliers are sampled from 1 class, and the outliers are sampled from the remaining 9 classes. The normal data is split into train and test sets with a ratio of 2:1, and the anomaly proposition is set to be 30%. Following [9], AUCROC is the evaluation metric for this experiment.

Given the protocol in [9], OLED is compared against MemAE [9] and other methods [17, 40, 41, 26]. The results are reported in Table 1. OLED yields excellent results, surpassing MemAE and other approaches. In particular, s_{rec}, s_{mask} and s_{avg} exceed all other identified approaches, recording an AUCROC of 0.977, 0.985 and 0.984, respectively. A visualization of OLED applied to both inlier and outlier samples for MNIST is available in Figure 2. Additionally, in Figure 3, the reconstructions of OLED are compared to that of a normal AE, further demonstrating the superiority of the representations offered by OLED for the anomaly detection task.

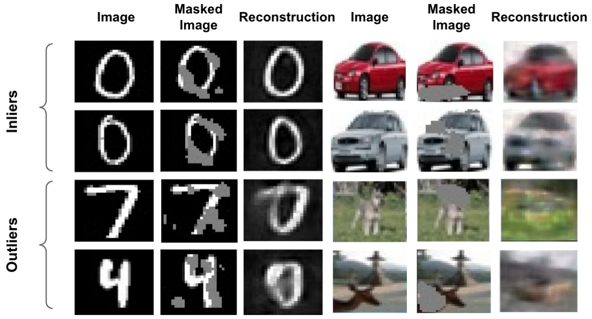

Figure 2. OLED Reconstructions. For both MNIST and CIFAR-10, the original image, perturbed image after applying the mask generated by MM (masks are illustrated in gray) and the final reconstruction are shown. Inlier samples are in the top two rows and outlier samples are in the bottom two rows.

Method	AUCROC
OCSVM [17]	0.5856
DAE [13]	0.5358
VAE [40]	0.5833
PixCNN [41]	0.5506
GAN [1]	0.5916
AND [10]	0.6172
AnoGAN [1]	0.6179
DSVDD [42]	0.6481
OCGAN [11]	0.6566
OLED (Ours) s_{rec}	**0.6622**
OLED (Ours) s_{mask}	**0.6711**
OLED (Ours) s_{avg}	**0.6683**
OLED (Ours) s_{cont}	**0.6673**

Table 2. One-class novelty detection Average AUCROC results on CIFAR-10 image dataset following the protocol in [11].

CIFAR-10: OLED is evaluated on CIFAR-10 using the protocol defined in [11]. This protocol involves dividing the dataset into ten different anomaly detection datasets corresponding to the ten predefined classes in CIFAR-10. In each anomaly detection dataset, the inliers are sampled from 1 class, and the outliers are sampled from the remaining 9 classes. The predefined train and test splits are used to conduct the experiments. Testing data of all classes are used for testing. Following [11], AUCROC is the evaluation metric for this experiment.

OLED is compared to OCGAN [11] and other recently

Method	AUCROC	EER
TSC [43]	0.922	-
FRCN action [44]	0.922	-
AbnormalGAN [45]	0.935	0.13
MemAE [9]	0.941	-
GrowingGas [46]	0.941	-
FFP [47]	0.954	-
ConvAE+UNet [48]	0.962	-
STAN [49]	0.965	-
Object-centric [50]	0.978	-
Ravanbakhsh [51]	-	0.14
ALOCC [29]	-	0.13
Deep-cascade [52]	-	0.09
Old is gold [31]	0.981	0.07
OLED (Ours) s_{rec}	**0.9854**	**0.0646**
OLED (Ours) s_{mask}	**0.9853**	**0.0683**
OLED (Ours) s_{avg}	**0.9902**	**0.0540**
OLED (Ours) s_{cont}	**0.9866**	**0.0606**

Table 3. Frame-level AUCROC and EER comparison on UCSD dataset with state-of-the-art methods.

proposed methods for anomaly detection [10, 13, 1, 42]. The results are reported in Table 2. OLED outperforms the compared methods, including OCGAN, by a considerable margin. Particularly, s_{rec}, s_{mask}, s_{avg} and s_{cont} exceed all other identified approaches, recording an AUCROC of 0.662, 0.671, 0.6683 and 0.667, respectively. A visualization of OLED applied to both inlier and outlier samples for CIFAR-10 is available in Figure 2.

4.4. Video Novelty Detection

One of the common use cases of one-class classification is in the domain of novelty detection for surveillance purposes [9, 27, 29]. Nonetheless, this task is more difficult in the video domain because of the variations of mobile objects across the frames. In this experiment, each frame of the dataset is divided into patches of size 30×30 pixels following [29]. Training patches only include scenes of walking pedestrians, while in the testing phase, patches are extracted from outlier frames that contain abnormal as well as normal objects. Frame-level AUROC and EER are the two metrics used to compare our method with state-of-the-art methods in recent years. As depicted in Table 3, our method outperforms recent state-of-the-art models in the video novelty detection task. More specifically, our approach achieves an AUCROC performance of 99.02% and an EER of 5.4%. The visualization in Figure 4 demonstrates the separability of anomaly scores for inliers and outliers.

4.5. Mask Module Evaluation

The results from the experiments in Section 4.3 and Section 4.4 are a clear indication that $OLED$ is a strong method for anomaly detection. In every case, anomaly scores that leveraged masking, and by extension MM, yielded the highest performance. Visual results in Figure 2 and 3 support the initial hypothesis that MM generates masks that cover important structures in the input image. Furthermore, this is the case for both inlier and outlier images. The following section seeks to solidify these observations more formally.

To quantitatively assess the effectiveness of MM in masking important parts of images, MM is re-purposed to perform a binary segmentation task that involves identifying whether or not each pixel in the input image is important. Specifically, the activation maps A generated by MM serve as the predicted semantic maps for images. A is used instead of $MM(x)$ to avoid the threshold constraint imposed by T. Using A and the ground truth semantic maps, the pixelwise AUCROC score is computed for both inlier and outlier images.

The aforementioned analysis is realized by evaluating the MM trained on digit class 8 from the MNIST experiments in Section 4.3 on the corresponding test set. MNIST is well suited for this experiment because we are able to make the assumption that nonzero pixels are part of the digit and thus important. The ground truth semantic maps for the test set are obtained by setting non zero activations to 1 otherwise 0. The former signals the pixel corresponds to part of the written digit, and the latter signals the pixel is part of the background.

The results for the experiment are displayed in Table 5. MM is able to segment important pixels in both inlier and outlier images with a high degree of accuracy with no modifications to the original architecture. This is a testament to the usefulness of MM in the anomaly detection task and hints at broader use cases in computer vision.

4.6. Ablation Study

In order to further assess the value of the proposed learned masking approach, OLED is compared to the baseline method context autoencoders (CAE). As CAEs employ random masking during training, the following section seeks to compare the learned masking proposed by OLED with random masking utilized in CAEs. To realize this comparison, a CAE was implemented and evaluated on the MNIST dataset using the protocol outlined in Section 4.3. The CAE shared the same architecture as R. The CAE is given input images with a random 10 x 10 region cropped out during training, keeping the number of masked pixels relatively consistent with R.

The results from the above experiment are displayed in 4. Similar to OLED, s_{rec}, s_{mask}, s_{avg} and s_{cont} are reported for the CAE. OLED is able to substantially outperform CAE, despite having identical architectures for the base reconstruction module. This is a clear indication that

Method	Score Type	AUCROC
CAE	s_{rec}	0.9209
CAE	s_{mask}	0.8936
CAE	s_{cont}	0.6869
CAE	s_{avg}	0.8768
OLED (Ours)	s_{rec}	**0.9772**
OLED (Ours)	s_{mask}	**0.9851**
OLED (Ours)	s_{cont}	**0.9650**
OLED (Ours)	s_{avg}	**0.9845**

Table 4. Comparison between our method (OLED) vs. Context Autoencoder (CAE) on MNIST image dataset.

Data	AUCROC
Inlier	0.8499
Outlier	0.8472

Table 5. Segmentaion performance of mask generator M on MNIST dataset.

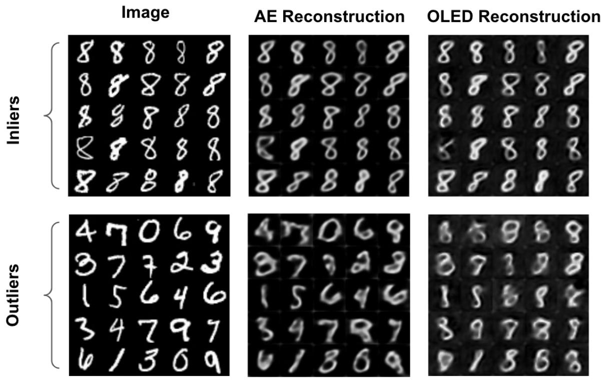

Figure 3. AE vs OLED Reconstructions for the MNIST dataset.

the learned masking approach proposed in OLED outperforms random masking for the anomaly detection task. Additionally, masking at test time enhances the performance of OLED but substantially decreases the performance of the CAE. This supports our intuition that the wrong placement of the masks by CAEs leads to suboptimal representations and introduce unwanted variations in the reconstruction error of samples that is detrimental to novelty detection performance.

5. Discussion

The results presented in Section 4 are a clear indication of the effectiveness of OLED for the anomaly detection task. In all three anomaly detection experiments on MNIST, CIFAR-10 and UCSD, OLED outperformed state-of-the-art methods by a large margin. Additional experiments evaluating the performance of MM demonstrated strong performance in segmenting the most important parts of samples for both the inlier and outlier class.

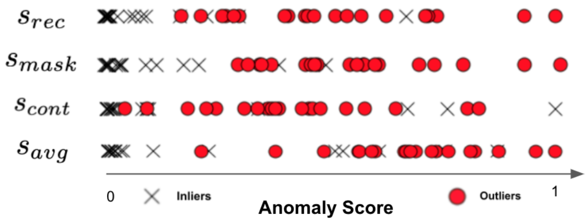

Figure 4. Sample of anomaly scores for both the inlier and outlier class for the UCSD dataset.

As initially hypothesized, OLED is able to reconstruct samples from the inlier class with ease but struggles to reconstruct samples from the outlier class. This addresses one of the fundamental problems AE face when applied to the anomaly detection task; reconstructing outliers too well. OLED accomplishes this by offering representations that are optimized for reconstructing important parts of the inlier samples through the adversarial training of R and MM. Beyond this, anomaly detection is enhanced through the use of masking at test time.

OLED also presents the benefit of being trained end-to-end, resulting in a less cumbersome training procedure than some of the identified methods. In this way, MM can be included seamlessly into existing AE-based anomaly detection methods. There are also no constraints that prevent OLED from being applied to other modalities of data. Furthermore, the core innovation proposed in this research, learned optimal masking, has the potential to be applied to other tasks in computer vision and beyond.

6. Conclusion

In this paper, we proposed an adversarial framework for novelty detection in both images and videos. More specifically, our method includes a Mask Module and a Reconstructor; the Mask Module is a convolutional autoencoder that learns to cover the most important parts of images, and the Reconstructor is a convolutional encoder-decoder that strives to reconstruct the masked images. The mask module will learn to mask the parts of input in a way to increase the reconstruction loss while the Reconstructor tries to minimize it. The proposed approach allows semantically rich representations and improves novelty detection at test time by covering the most important parts of the context. We have applied our method to a variety of tasks, including outlier and anomaly detection in images and videos. The results illustrate the superiority of OLED in identifying samples related to the outlier class compared to recent state-of-the-art models.

References

[1] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017.

[2] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 341–349, 2017.

[3] Arthur Zimek, Erich Schubert, and Hans-Peter Kriegel. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(5):363–387, 2012.

[4] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[5] Yoshua Bengio, Pascal Lamblin, Dan Popovici, Hugo Larochelle, et al. Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, 19:153, 2007.

[6] Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019.

[7] Yan Xia, Xudong Cao, Fang Wen, Gang Hua, and Jian Sun. Learning discriminative reconstructions for unsupervised outlier removal. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1511–1519, 2015.

[8] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International Conference on Learning Representations*, 2018.

[9] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1705–1714, 2019.

[10] Davide Abati, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Latent space autoregression for novelty detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 481–490, 2019.

[11] Pramuditha Perera, Ramesh Nallapati, and Bing Xiang. Ocgan: One-class novelty detection using gans with constrained latent representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2898–2906, 2019.

[12] David Zimmerer, Simon AA Kohl, Jens Petersen, Fabian Isensee, and Klaus H Maier-Hein. Context-encoding variational autoencoder for unsupervised anomaly detection. *arXiv preprint arXiv:1812.05941*, 2018.

[13] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

[14] Guillaume Alain and Yoshua Bengio. What regularized autoencoders learn from the data-generating distribution. *The Journal of Machine Learning Research*, 15(1):3563–3593, 2014.

[15] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[16] Christoph Baur, Stefan Denner, Benedikt Wiestler, Nassir Navab, and Shadi Albarqouni. Autoencoders for unsupervised anomaly segmentation in brain mr images: A comparative study. *Medical Image Analysis*, page 101952, 2020.

[17] Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[18] Paul Hayton, Bernhard Schölkopf, Lionel Tarassenko, and Paul Anuzis. Support vector novelty detection applied to jet engine vibration spectra. In *NIPS*, pages 946–952. Citeseer, 2000.

[19] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[20] Heiko Hoffmann. Kernel pca for novelty detection. *Pattern recognition*, 40(3):863–874, 2007.

[21] Liang Xiong, Barnabás Póczos, and Jeff Schneider. Group anomaly detection using flexible genre models. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, 2011.

[22] Yang Cong, Junsong Yuan, and Ji Liu. Sparse reconstruction cost for abnormal event detection. In *CVPR 2011*, pages 3449–3456. IEEE, 2011.

[23] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. Learning deep representations of appearance and motion for anomalous event detection. *arXiv preprint arXiv:1510.01553*, 2015.

[24] Mohammad Sabokrou, Mahmood Fathy, and Mojtaba Hoseini. Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder. *Electronics Letters*, 52(13):1122–1124, 2016.

[25] Mayu Sakurada and Takehisa Yairi. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, pages 4–11, 2014.

[26] Shuangfei Zhai, Yu Cheng, Weining Lu, and Zhongfei Zhang. Deep structured energy based models for anomaly detection. In *International Conference on Machine Learning*, pages 1100–1109. PMLR, 2016.

[27] Chong Zhou and Randy C Paffenroth. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 665–674, 2017.

[28] Yong Shean Chong and Yong Haur Tay. Abnormal event detection in videos using spatiotemporal autoencoder. In *International symposium on neural networks*, pages 189–196. Springer, 2017.

[29] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. Adversarially learned one-class classifier for novelty detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3379–3388, 2018.

[30] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[31] Muhammad Zaigham Zaheer, Jin-ha Lee, Marcella Astrid, and Seung-Ik Lee. Old is gold: Redefining the adversarially learned one-class classifier training paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14183–14193, 2020.

[32] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H Rohban, and Hamid R Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14902–14912, 2021.

[33] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12742–12752, 2021.

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[35] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[36] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[37] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Technical Report*, 2009.

[38] Antoni Chan and Nuno Vasconcelos. Ucsd pedestrian dataset. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 30(5):909–926, 2008.

[39] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):18–32, 2013.

[40] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[41] Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with pixelcnn decoders. *arXiv preprint arXiv:1606.05328*, 2016.

[42] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018.

[43] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 341–349, 2017.

[44] Ryota Hinami, Tao Mei, and Shin'ichi Satoh. Joint detection and recounting of abnormal events by learning deep generic knowledge. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3619–3627, 2017.

[45] Mahdyar Ravanbakhsh, Moin Nabi, Enver Sangineto, Lucio Marcenaro, Carlo Regazzoni, and Nicu Sebe. Abnormal event detection in videos using generative adversarial nets. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1577–1581. IEEE, 2017.

[46] Qianru Sun, Hong Liu, and Tatsuya Harada. Online growing neural gas for anomaly detection in changing surveillance scenes. *Pattern Recognition*, 64:187–201, 2017.

[47] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6536–6545, 2018.

[48] Trong-Nguyen Nguyen and Jean Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[49] Sangmin Lee, Hak Gu Kim, and Yong Man Ro. Stan: Spatio-temporal adversarial networks for abnormal event detection. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1323–1327. IEEE, 2018.

[50] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7842–7851, 2019.

[51] Mahdyar Ravanbakhsh, Enver Sangineto, Moin Nabi, and Nicu Sebe. Training adversarial discriminators for cross-channel abnormal event detection in crowds. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1896–1904. IEEE, 2019.

[52] Mohammad Sabokrou, Mohsen Fayyaz, Mahmood Fathy, and Reinhard Klette. Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. *IEEE Transactions on Image Processing*, 26(4):1992–2004, 2017.