Adversarial Discriminative Attention for Robust Anomaly Detection

Daiki Kimura
IBM Research AI
daiki@jp.ibm.com

Subhajit Chaudhury
IBM Research AI
subhajit@jp.ibm.com

Minori Narita
University of Massachusetts, Amherst
mnarita@umass.edu

Asim Munawar
IBM Research AI
asim@jp.ibm.com

Ryuki Tachibana
IBM Research AI
ryuki@jp.ibm.com

Abstract

Existing methods for visual anomaly detection predominantly rely on global level pixel comparisons for anomaly score computation without emphasizing on unique local features. However, images from real-world applications are susceptible to unwanted noise and distractions, that might jeopardize the robustness of such anomaly score. To alleviate this problem, we propose a self-supervised masking method that specifically focuses on discriminative parts of images to enable robust anomaly detection. Our experiments reveal that discriminator's class activation map in adversarial training evolves in three stages and finally fixates on the foreground location in the images. Using this property of the activation map, we construct a mask that suppresses spurious signals from the background thus enabling robust anomaly detection by focusing on local discriminative attributes. Additionally, our method can further improve the accuracy by learning a semi-supervised discriminative classifier in cases where a few samples from anomaly classes are available during the training. Experimental evaluations on four different types of datasets demonstrate that our method outperforms previous state-of-the-art methods for each condition and in all domains.

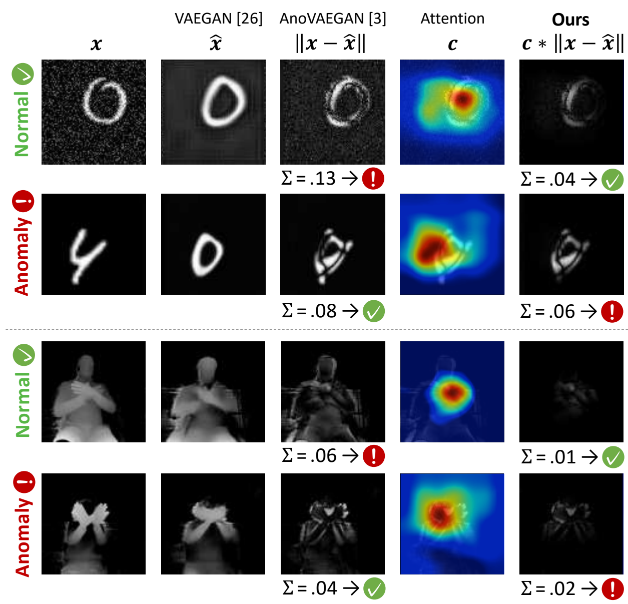


Figure 1: Proposed method is robust against noise by attentions from the network. Top: noisy MNIST dataset, Bottom: pigeon gesture [48] dataset. From left: input image, reconstructed image from VAEGAN [26], difference image for AnoVAEGAN [3], adversarial discriminative class activation map, and difference image for proposed method. The difference between the reconstructed image and the original image should be low for normal samples and high for anomaly samples. However, the prior method produces a higher score for the normal sample due to background noise ($0.13 > 0.08$ for MNIST dataset). The proposed method produces the correct classification ($0.04 < 0.06$).

1. Introduction

Anomaly detection [7] handles the task of recognizing whether an input sample lies close to the distribution of a given target normal class or anomaly class (outliers). Anomaly detection methods have applications in various fields such as medical imaging, industrial inspection, and fraud detection. The typical problem setting involves samples from the normal class distribution, and the goal is to detect whether or not an input test sample belongs to the normal class. This setting is also referred to as one-class classification [6, 33] or out-of-distribution detection [29].

A standard approach for this kind of problem setting is to learn a representation of the normal class distribu-

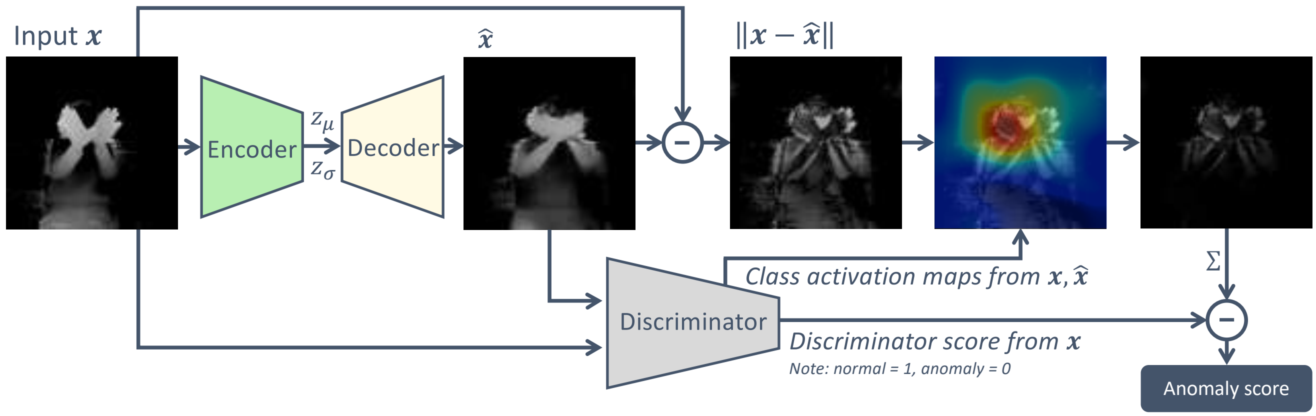


Figure 2: Overview of proposed method showing that a mask obtained from the class activation maps focuses on salient features, suppressing background noise, and thus improving robustness of anomaly score detection.

tion and define an anomaly score metric that provides a measurement of how different the test image is from the normal class distribution. Modern anomaly detection systems employ generative models to learn the manifold for the normal class. The anomaly score is usually computed as the difference in the Euclidean space between the reconstructed and original images. Auto-encoders, variational auto-encoders (VAEs), and generative adversarial networks (GANs) are commonly used for training such distributions [1–3, 10, 12, 39–41, 51].

However, such methods are liable to make erroneous detections due to background noise and noise on objects. For example, in industrial applications, using anomaly detection for robotic inspection might give a false positive output if there is an unexpected change in the background of the image. In contrast, humans can identify relationships among objects and focus only on the parts of the image that are relevant for visually discerning whether the image belongs to the normal class or not.

A straightforward approach to remedying this problem is to calculate detections or segmentations from saliency labels [17, 36]. However, the preparation of labeled data causes laborious real-world problems, and human biases can be misleading. We need an unsupervised method for obtaining saliency. The recently proposed gradient-weighted class activation mapping (Grad-CAM) [43] computes such saliency from gradients of convolutional neural networks (CNNs) without region labels.

In this work, we use visual saliency calculated by Grad-CAM [43] from an adversarial network to improve the noise robustness of anomaly detection. Figure 1 shows the benefits of introducing the attentions generated from the class activation maps, and Fig. 2 gives an overview of the proposed method. The visual saliency is calculated from the discriminator network in a variational auto-encoder with generative adversarial network [26], which is trained for reconstructing normal images. We use the class activation maps from the discriminator networks to focus on spatially relevant areas of the image to reduce the effect of unwanted background distractions and noise. After getting the activation map, we use it as a mask for calculating an anomaly score on the basis of error in the reconstruction of a given input. This technique is self-supervised method with adversarial network for anomaly detection, and does not require preparation of labeled segmentation data. We argue that, for discerning normal and anomaly class distribution, the discriminators attention map would align with the location of the primary object in the image and hence would mask unwanted noisy perturbations outside the main discriminative area. In this paper, we compare our method with the state-of-the-art methods on four different kinds of anomaly detection problems which have a wide variety of samples. We especially focus on realistic dataset settings that contain complex noise and extremely few training images [31]. Experiments are conducted to validate that our method can discard the effect of noise in the images, and thereby improve the AUROC scores for anomaly detection.

In addition, when we look at real-world anomaly detection problems, we find that in many cases, a small number of anomaly examples are known beforehand. For example, it is common for factory workers to know some typical anomaly modes of manufactured products while not knowing all possible anomaly modes. Yamaguchi *et al*. reported several typical anomaly patterns for a screening test for dementia [48, 49]. However, even in such cases, we should not simply train a classifier that discriminates normal and anomaly classes because the response from truly “unseen” anomaly samples will be intrinsically unexpected. Hence, we propose a semi-supervised discriminative attention based anomaly detection method to utilize the few labeled samples to improve the visual saliency model.

Our major contribution are:

1. We introduce the concept of utilizing discriminator's class activation map (CAM) for constructing a self-supervised method for suppressing spurious signals from image background for robust anomaly detection. We also extend it for semi-supervised learning case with some known anomaly samples.
2. We empirically demonstrate that discriminator attention in adversarial learning evolves over three stages which forms the basis of our method.
3. Experimental validations show that we outperform the state-of-the-art methods on four different datasets including domains with real-world settings have small training samples with complex noise distribution.

2. Related Work

2.1. Anomaly detection

Anomaly detection is also known in prior works as "novelty detection", "outlier detection", and "out-of-distribution detection". In an earlier work [42], the support-vector machine [11] framework was used to predict a positive score for samples drawn from the normal subspace and a negative score from the anomaly subspace. This study also provided a theoretical justification for generalization error bounds. Following the above kernel-based methods, auto-encoders were used for non-linear feature extraction from data [40]. This method outperformed linear principal component analysis (PCA) and kernel PCA [22].

Recently, deep learning has been extensively incorporated into anomaly detection methods because of notable improvements in deep learning's ability to assist other tasks such as image classification [18, 25]. Hendrycks *et al*. provide a baseline in outlier detection using CNNs that use probability from softmax distributions for normal and outlier images [19]. A common approach consists of learning a generative model on normal images and finding the error between the reconstructed image and input. Auto-encoder, VAE [24], and GAN [15] are the most popular methods for learning the distribution of a certain class. Simple reconstruction-error-based anomaly detection with a VAE was introduced by An *et al*. [2]. Chen *et al*. introduced noise in auto-encoder structure and showed improved performance in outlier detection [10].

Recent anomaly detection studies are using generative adversarial methods for generating reconstructed images. Schlegl *et al*. proposed a GAN-based anomaly detection method, called AnoGAN, for optical coherence tomography images of the retina [41]. In their work, the 'normal' class manifold is implicitly learned by GAN. For unseen test images, the distance between the original image and its projection on the learned non-linear manifold is used to obtain the anomaly score. Their method has an optimization step during test phase to get an image that is similar to the original. Therefore, AnoGAN has problems of having many local minima in optimization and is computationally expensive. Zenati *et al*. improved the speed of the above architecture by introducing a learned encoder for latent space re-projection instead of optimization [51]. Akcay *et al*. proposed an encoder architecture along with the generator for explicitly learning the distance function between the target and the anomaly class [1]. Recently, Sabokrou *et al*. introduced a new detection method using generator- and discriminator-based adversarial learning along with another network that helps the discriminator to learn the distribution of the normal class [38]. More recently, Baur *et al*. proposed another anomaly detection method [3] called AnoVAEGAN, which introduces VAEGAN [26] as the generative model. They applied AnoVAEGAN to detection in brain MR images. This method does not require the optimization steps, which helps to prevent it from dropping the local minimum. However, they only report one simple domain example with one condition, and this study did not introduce an attention mechanism.

The application of anomaly detection methods [6, 7, 47] are well reported in some literatures, especially the medical domain. Hauskrecht *et al*. proposed an alerting system that monitors a patient by using an anomaly detection method [16]. Prastawa *et al*. proposed a framework for using detection to automatically segment brain tumors in MR images [35]. Applications in other fields include anomaly detection in crowded scenes [28,37] and defect detection [5, 32]. In addition, Sultani *et al*. proposed anomaly detection in videos [44]. Further, Brown *et al*. and Wu *et al*. proposed introducing a recurrent neural-network-based attention for anomaly detection [4, 46]. However, previous methods [2,3,41], related to the our method, did not propose explicit measures for handling noise and distractions in images. In this paper, we propose using visual saliency from adversarial training to focus on discriminative attributes.

2.2. Attention from class activation maps

There are several techniques that try to explain the decisions of CNNs by looking at the gradient flow in the networks. Zhou *et al*. proposed class activation mapping (CAM) to visualize the region of interest (ROI) and performed object localization without training on any bounding box annotation [52]. This method tries to determine which feature maps are more important for a particular class by computing global average pooling (GAP) of the last convolution layer. A linear combination of the gradient with respect to a target class is then computed as the activation map for that class. However, this requires re-training of the base model for the classification task. Therefore, Selvaraju *et al*. proposed Grad-CAM, which does not require such re-training [43]. This method back-propagates gradients for a particular class by rectifying the gradients using ReLU activation to look at the pixels that have a positive effect on the class decision. An extension to this work was recently proposed as well [8].

3. Proposed Method

Our method for anomaly detection consists of a VAE-based generator network $\mathcal{G}$ and discriminator network $\mathcal{D}$. The generator network has a variational bottleneck, and its task is to reconstruct the image given as input. The discriminator network tries to learn a classifier between the real and reconstructed image. The goal is to learn a manifold of the normal class using the generative process. If the networks are trained well, the distance between reconstructed and input images will be small for normal input and high for anomaly input. However, the reconstruction process might produce some unwanted artifacts, especially in the presence of noise and distractions in the input image. Even if the training images also have a large amount images with such noise, the generative model generally does not learn the un-structured distributions. Our proposed method uses attention information from the discriminator network for self-attention to focus on the most salient feature to prevent the effect of noise. The details of each component are given in the following sections.

3.1. Training

Previous approaches to anomaly detection predominantly use an encoder-decoder model [1–3, 10, 39, 41] for learning the distribution of the normal class. The previous methods are using adversarial training for improved class-distribution learning. Although AnoGAN [41] has good detection results, it suffers from the problem of having local minimums based on the initial value during optimization in the test phase. AnoVAEGAN [3] can solve this problem by removing the optimization in the test phase; however, only one example is reported and it did not use attention. Since AnoVAEGAN is a baseline method for our proposed method, we explain the training of the VAE-based generative model in this section.

To implement the generator network, we use CNNs for both encoder-decoder networks to map samples from the normal image to itself. Thus, the generator network's principle task is to maintain an identity mapping for the distribution of the normal class. When the test image is an anomaly sample, the generator removes the salient features from the anomaly distributions by mapping them to the closest projection in the normal space.

Following the work of Kingma *et al*. [24], we injected noise in the latent space from a unit Gaussian distribution. Given an input sample $\boldsymbol{x}$, we obtain the latent distribution parameters from the encoder as $\boldsymbol{z_\mu}, \boldsymbol{z_\sigma} = \mathcal{G}_{enc}(\boldsymbol{x})$, and the reconstructed image is created from the decoder as $\hat{\boldsymbol{x}} = \mathcal{G}_{dec}(\boldsymbol{z_\mu} + \mathcal{N}(\mathbf{0}, \boldsymbol{I}) * \boldsymbol{z_\sigma})$. For generative modeling using VAE, we minimize the log likelihood of the model, $-\sum_i \log p_{\boldsymbol{\theta}}(\boldsymbol{x_i})$. The VAE approximates the true latent distribution $p(\boldsymbol{z}|\boldsymbol{x_i})$ with variational approximation and a proxy distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x_i})$. The parameters for this distribution are learned to match a prior distribution $p(\boldsymbol{z})$, which we consider as unit normal $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$. This loss is given as,

$$\mathcal{L}_{prior} = D_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x_i})|p(\boldsymbol{z})), \tag{1}$$

where $D_{KL}(.|.)$ is the Kullback-Leibler (KL) divergence between two distributions. Traditionally, the above $\mathcal{L}_{prior}$ and the l_2 between the input and reconstructed image (assuming Gaussian decoder distribution) are jointly minimized to train the network. However, the reconstruction quality of simple VAEs tends to be blurry and unrealistic in nature [13], which might affect the reconstructed image by giving it a high anomaly score even in the case of normal class inputs. Therefore, we use adversarial training that minimizes the Jensen-Shannon divergence [15] between the generated and true distributions to produce sharper images.

We designed the discriminator as a CNN whose task is to determine whether or not an input image belongs to the real images of the normal class distribution. The output is a probability distribution $p(y_c = 1|\boldsymbol{x})$ signifying the likelihood of the input image to be from the normal class. We train the discriminator in two possible settings: (1) self-supervised setting which is a case without prior anomaly data, and (2) semi-supervised setting which is a case with a few known anomaly samples.

3.1.1 Self-supervised setting

This is a standard problem setting for anomaly detection, in which we have access to only the normal class, and anomaly samples are not seen during training. We train the generator and discriminator with the following loss function,

$$\begin{aligned}\mathcal{L}_{GAN} = \; & \mathbb{E}_{\boldsymbol{x}\sim p_{data}} \log[\mathcal{D}(\boldsymbol{x})] \\ & + \mathbb{E}_{\boldsymbol{z_\mu}, \boldsymbol{z_\sigma}\sim p_z} \log[1 - \mathcal{D}(\mathcal{G}_{dec}(\boldsymbol{z_\mu}, \boldsymbol{z_\sigma}))] \\ & + \mathbb{E}_{\boldsymbol{x}\sim p_{data}} \log[1 - \mathcal{D}(\mathcal{G}_{dec}(\mathcal{G}_{enc}(\boldsymbol{x})))],\end{aligned} \tag{2}$$

where the $\mathcal{D}$ is the discriminator network. The first loss encourages the discriminator to classify the real data as positive (normal) samples, and the other two loss terms classify the images reconstructed from those sampled from latent space and input images as negative (anomaly) samples.

Joint VAE-and-discriminator training is performed by minimizing the two losses and maximizing the one loss together in a minimax way,

$$\min_{\mathcal{G}_{enc}, \mathcal{G}_{dec}} \max_{\mathcal{D}} \; (\; \mathcal{L}_{GAN} + \mathcal{L}_{prior} + \mathcal{L}_{image} \;) \; , \tag{3}$$

where $\mathcal{L}_{image}$ is the distance between the original and reconstructed images.

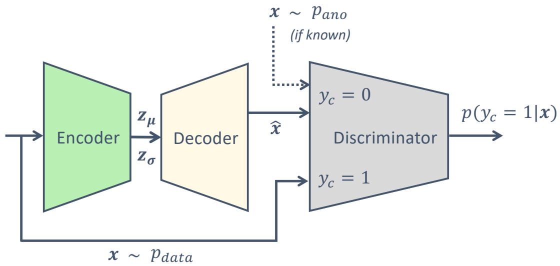


Figure 3: Training of the model. If some samples from a few anomaly modes are known *a priori*, they are used as negative samples for the discriminator.

3.1.2 Semi-supervised setting

Here, we assume that, when anomaly condition consists of multiple classes/modes, the system knows one anomaly class during training; or when an anomaly exists in one class, the system knows a small number of samples. We incorporate this known data into our discriminator training framework as shown in Fig. 3. The anomaly data is treated as negative samples in the discriminator along with the reconstructed images from the generator. The adversarial loss in this case is modified as follows,

$$\mathcal{L}_{GAN}^{ano} = \mathcal{L}_{GAN} + \mathbb{E}_{\boldsymbol{x} \sim p_{ano}}(\log[1 - \mathcal{D}(\boldsymbol{x})]), \quad (4)$$

where the p_{ano} is the known anomaly distribution.

3.2. Anomaly detection with attention

Although the above generative model is trained for normal class distributions, in cases where real-world noise appears in the test images, the reconstructed image might significantly vary from the normal samples. Anomaly detection by humans is very efficient in such cases; we can learn to ignore irrelevant disturbances by focusing on the salient objects in the scene. We propose the attention mechanism for self-supervised setting which has no prior anomaly samples. Here, we explain an assumption for evolution of discriminative attention. Also, we assume a small number of anomaly samples during training will enable faster focusing on the detailed parts. We explain about the semi-supervised setting which is with a few anomaly samples.

3.2.1 Self-supervised setting

To obtain the attention information, we apply a Grad-CAM [43] method to the discriminator network for extracting a highlight map of the important regions as a class activation map without any modification to the network and training on boundary box information. Our idea is to use the discriminator's attention information to identify salient features in the given images. Only using normal image samples for training a GAN makes the discriminator learns distinctive features between normal and reconstructed images while the generator tries to generate "normal"-like images. Hence, we believe that the evolution of discriminator's attention would follow these stages:

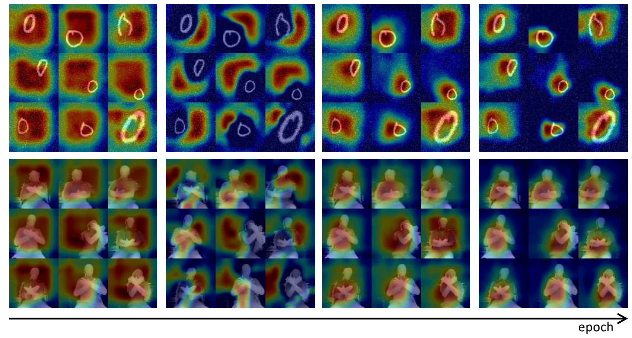


Figure 4: Class activation maps for discriminator with increasing epochs. The top images show maps for MNIST with noise (normal class is '0'); the bottom images show maps for the pigeon gesture. The maps gradually focus on detailed aspects, such as digit region or hand position.

1. In the early training phase, the discriminator would focus on almost the entire area of the images due to limited learning.

2. After the generator can produce convincing foreground structure, the discriminator would try to use features from the background for discriminating the real and fake samples. However, since the background predominantly consists of non-discriminative features, the discriminator would eventually focus on the foreground for attempting discriminative classification.

3. In the last stage, the discriminator will gradually focus on detailed areas of the foreground region which can be used for fine-grained classification.

We validated our hypothesis about the discriminator's attention evolution by experiments as shown in Fig. 4. The fact that discriminator's final activation map is focused on the foreground region enables us to use it for robust anomaly detection in this paper.

The anomaly detection is processed, determining "normal" or "anomaly" by a threshold and anomaly score in each test image. In this paper, since we will evaluate the ROC curve, we will not discuss setting the threshold as an exact number. The anomaly score is computed by multiplying the class activation map and the error from the image reconstructed by the generator and the given test image. While multiplying the attention, we add the normalization process for attention values, which is for normalizing the area size of attention. Also, we found combining the attention maps generated from not only given input $\boldsymbol{x}$ but

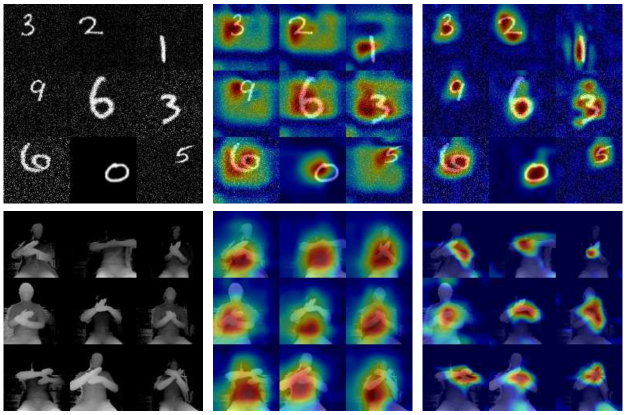


Figure 5: Class activation maps from different discriminators. From left: input, map from $\mathcal{D}_{\mathcal{G}}$, and map from $\mathcal{D}_{ano}$. The map from $\mathcal{D}_{ano}$ can generate better attentions.

also from the generated image $\hat{\boldsymbol{x}}$ has better results. This is because the generated image (normal class) might have a wider range than the input image (e.g., the normal class is "0", and the test input is "1"). In addition, we subtract the probability of the normal distribution from the discriminator, which is an output of $\mathcal{D}$. Hence, the anomaly score calculation of our proposed method is,

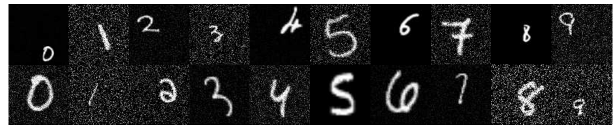
$$\begin{aligned} \boldsymbol{c} &= CAM_{\mathcal{D}}(\boldsymbol{x}) + CAM_{\mathcal{D}}(\hat{\boldsymbol{x}}), \\ score &= \frac{\|\boldsymbol{c} * (\boldsymbol{x} - \hat{\boldsymbol{x}})\|_2^2}{\|\boldsymbol{c}\|_2^2} - \beta \mathcal{D}(\boldsymbol{x}), \end{aligned} \tag{5}$$

where $CAM_{\mathcal{D}}(\boldsymbol{x}) = \mathrm{ReLU}(\sum_k \alpha_{\boldsymbol{x}}^k A^k)$ is a class activation map from the discriminator network for a given image $\boldsymbol{x}$, $\hat{\boldsymbol{x}} = \mathcal{G}_{dec}(\mathcal{G}_{enc}(\boldsymbol{x}))$ is the reconstructed image, and β is a parameter that weighs the discriminator score.

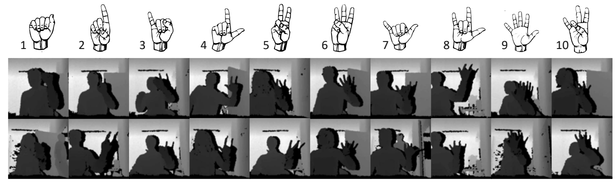
3.2.2 Semi-supervised setting

In this scenario, we found that training an ensemble of discriminators on the generated images and given anomaly samples separately produces a better result for merging than training one discriminator due to the differences for each training goal. We combine the class activation maps from two discriminators: $\mathcal{D}_{\mathcal{G}}$ trained with generated images, and $\mathcal{D}_{ano}$ trained with a few anomaly samples. Figure 5 shows the attention outputs for each discriminator. Hence,

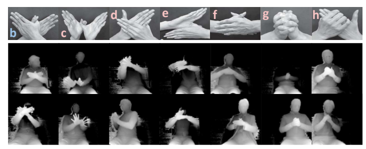
$$\begin{aligned} \boldsymbol{c} = CAM_{\mathcal{D}_{\mathcal{G}}}(\boldsymbol{x}) &+ CAM_{\mathcal{D}_{\mathcal{G}}}(\hat{\boldsymbol{x}}) \\ &+ CAM_{\mathcal{D}_{ano}}(\boldsymbol{x}) + CAM_{\mathcal{D}_{ano}}(\hat{\boldsymbol{x}}), \\ score = \frac{\|\boldsymbol{c} * (\boldsymbol{x} - \hat{\boldsymbol{x}})\|_2^2}{\|\boldsymbol{c}\|_2^2} &- \beta(\mathcal{D}_{\mathcal{G}}(\boldsymbol{x}) + \mathcal{D}_{ano}(\boldsymbol{x})). \end{aligned} \tag{6}$$



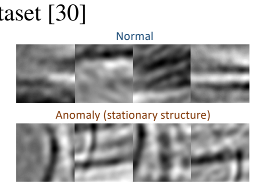
(a) MNIST with noise



(b) Hand gesture dataset [30]



(c) Pigeon gesture for [48]



(d) Venus [50]

Figure 6: Sample images for datasets

4. Experiments

4.1. Datasets

In this paper, we evaluated the proposed method on four datasets from different domains. For all datasets, we assume real-world environment settings for noise and the number of training images. We use high-variance and real-world noise, the normal class is just one mode, and the anomaly class contains several modes (e.g., 9 modes). This means, the training uses only one mode; however, the evaluation uses all modes from normal and anomaly classes. Further, each mode contains a small number of images because obtaining real images is generally difficult. Some sample images for all datasets are shown in Fig. 6.

- MNIST with noise

We modified the original MNIST [27] dataset by randomly placing digits in an 84×84 image with arbitrary digit size. Furthermore, we added noise $\mathcal{N}(0, \sigma^2)$, where σ is also sampled from $\mathcal{N}(30, 30^2)$ for each image.

- Hand gesture

We evaluated on a public hand gesture dataset [30, 45] to extend following real hand-gesture anomaly detection. It consists of depth images with 10 hand pose from 14 persons. It only provides 10 images for each gesture and each person. Note that we cropped and resized the images to be 84×84 images.

- Pigeon gesture

This is for a medical screening test of dementia. We captured 84×84 depth images of a "pigeon"-pose screening test [48] from 18 people. While capturing the images, we

MNIST with noise	normal digit										average	Pigeon [48]	Venus [50]
	0	1	2	3	4	5	6	7	8	9			
AnoVAE [2]	.63±.2	.83±.1	.54±.1	.57±.1	.62±.1	.67±.0	.61±.1	.69±.1	.55±.1	.70±.0	.640±.12	.643±.04	.800±.08
AnoGAN [41]	.45±.0	.75±.0	.48±.0	.49±.0	.58±.0	.51±.0	.51±.0	.59±.0	.47±.0	.56±.0	.540±.09	.489±.01	.842±.05
AnoVAEGAN [3]	.71±.0	.86±.0	.60±.0	.65±.0	.68±.0	.66±.0	.70±.0	.73±.0	.64±.0	.70±.0	.692±.07	.682±.03	.881±.04
Ours ($\mathcal{D}$ only)	.55±.1	.70±.1	.54±.0	.52±.0	.53±.1	.51±.0	.57±.0	.59±.0	.57±.0	.56±.0	.563±.06	.597±.05	.555±.10
Ours (CAM only)	**.74±.0**	**.92±.0**	**.62±.0**	**.68±.0**	**.69±.0**	**.69±.0**	**.73±.0**	**.78±.0**	.66±.0	**.74±.0**	**.725±.08**	**.766±.05**	**.892±.04**
Ours (CAM+$\mathcal{D}$)	**.74±.0**	**.92±.0**	**.62±.0**	**.68±.0**	.68±.0	**.69±.0**	**.73±.0**	.77±.1	**.67±.0**	**.74±.0**	**.725±.08**	**.766±.05**	**.892±.04**

Table 1: AUROC value for MNIST with noise, pigeon gesture dataset for [48], and Venus image dataset [50]. For MNIST, the results for each column have different normal class conditions, e.g., the left most result has only "0" as the normal class, while other digits are treated as anomalies. During training, only the normal class is given. The values are the average results for five different random seeds; each seed generates a different image set. **Bold** results are the best for each condition.

Hand gesture [30]	normal hand gesture										average
	1	2	3	4	5	6	7	8	9	10	
AnoVAE [2]	.79±.01	.69±.02	.69±.03	.72±.01	.69±.01	.64±.02	.66±.02	.66±.02	.64±.02	.62±.03	.682±.05
AnoGAN [41]	.66±.07	.62±.03	.62±.05	.55±.05	.59±.03	.58±.02	.63±.04	.57±.06	.50±.05	.54±.03	.587±.06
AnoVAEGAN [3]	.84±.06	.82±.03	.81±.02	.81±.05	.79±.05	.81±.04	.86±.03	.82±.02	.69±.03	.79±.03	.804±.06
Ours ($\mathcal{D}$ only)	.78±.06	.78±.05	.75±.03	.77±.08	.75±.04	.75±.04	.78±.03	.81±.04	.67±.06	.80±.04	.763±.06
Ours (CAM only)	.85±.06	.84±.03	.82±.03	.81±.05	.79±.06	.80±.05	.87±.02	.80±.02	.70±.06	.81±.02	.809±.06
Ours (CAM+$\mathcal{D}$)	**.87±.05**	**.87±.03**	**.84±.04**	**.84±.03**	**.82±.05**	**.83±.03**	**.89±.02**	**.85±.02**	**.74±.07**	**.86±.01**	**.841±.05**

Table 2: AUROC value for hand gesture dataset [30]. The results for each column have different normal class conditions. The values are the average results for five different random seeds; each seed generates a different image set.

asked participants to randomize their hand position and angle, the closed or open state of their fingers, their sitting positions, and the angle of their posture. We randomly sampled 10 images for each pose and each person, i.e. it is same difficult conditions with the previous dataset. For this dataset, only the "b"-pose is the normal class, and "c", "d", and "e" are candidates for the known anomaly class because Yamaguchi *et al*. reported these as typical patterns for the subjects [48].

- Venus images

We used a public Venus image dataset [21, 50] to detect stationary structures. Detecting stationary features automatically is helpful for Venus researches [14]. Yamazaki *et al*. provide images from several wavelengths, and we chose ultraviolet 283 nm images. We did a pre-process similar to Fukuhara *et al*. [14]. We sampled 500 normal patches and 225 stationary patches as anomaly samples. We selected 5 large stationary structures as candidates for the known anomaly.

4.2. Quantitative evaluation

We perform quantitative evaluation with all the above datasets using the AUROC curve.

4.2.1 Self-supervised setting

We use the discriminator's attention information to focus on relevant areas in the image. We compare our results with the reconstruction-based methods of VAE [2] (AnoVAE), GAN [41] (AnoGAN), and VAEGAN [3] (AnoVAEGAN). The network structure for all networks is inspired from the VAEGAN [26] paper. The $\mathcal{G}_{enc}$ has three convolutional layers with down-sampling, leaky ReLU, and batch normalization [20]. The dimension of the latent variable is 128. The $\mathcal{G}_{dec}$ has five convolutional layers with up-sampling, leaky ReLU (the final is tanh), and batch normalization. The $\mathcal{D}$ has five convolutional layers with down-sampling, leaky ReLU (the final is sigmoid), and batch normalization. We use RMSProp optimizer (the learning ratio is 0.0003) for all networks. We prepare two baseline methods for comparison: one uses only the output of the discriminator ($\mathcal{D}$ only), and the other uses only the focused error (CAM only). The β in Eq. (5) is 0.2 for MNIST and pigeon, 5 for hand gesture, and 0.05 for Venus (we tried some values $\{0.05, 0.1, 0.2, 1, 5\}$).

Table 1 and Table 2 show the AUROC values for all the above datasets. For noisy MNIST and the hand gesture dataset, we tried all combinations of normal class. Note that

Noisy MNIST normal digit: 0	known anomaly digit 1	3	5	7	9	average	Pigeon [48], normal: b c	d	e	average	Venus [50]
Ours (CAM$_\mathcal{G}$+$\mathcal{D}_\mathcal{G}$)	.74±.01	.74±.01	.74±.01	.74±.01	.74±.01	.738±.01	.77±.05	.77±.05	.77±.05	.766±.05	.892±.04
CNN regression	.90±.01	**.94±.01**	**.95±.01**	.93±.01	.94±.00	.932±.02	.84±.03	.90±.01	.89±.01	.878±.03	.827±.07
Ours (CAM$_{ano}$)	.78±.02	.77±.02	.75±.01	.77±.02	.77±.02	.768±.02	.74±.06	.81±.03	.80±.03	.784±.05	.888±.04
Ours (CAM+$\mathcal{D}_\mathcal{G}$+$\mathcal{D}_{ano}$)	**.93±.01**	**.94±.01**	**.95±.01**	**.95±.00**	**.95±.01**	**.945±.01**	**.87±.05**	**.94±.01**	**.92±.01**	**.910±.04**	**.924±.04**

Table 3: AUROC value for noisy MNIST, pigeon gesture for [48], and Venus dataset [50]. In this table, one of the anomaly classes is known during training. Each column shows different anomaly class conditions. For the Venus dataset, there is only one known anomaly candidate. The value is an average of five random seeds.

Hand gesture [30] normal: 1	known anomaly hand gesture 2	3	4	5	6	7	8	9	10	average
Ours (CAM$_\mathcal{G}$+$\mathcal{D}_\mathcal{G}$)	.87±.05	.87±.05	.87±.05	.87±.05	.87±.05	.87±.05	.87±.05	.87±.05	.87±.05	.869±.05
CNN regression	.91±.02	.93±.02	.91±.03	.95±.02	**.95±.02**	.93±.02	.95±.01	**.94±.03**	**.95±.01**	.935±.02
Ours (CAM$_{ano}$)	.84±.07	.85±.05	.85±.06	.85±.05	.85±.06	.85±.06	.85±.06	.85±.05	.86±.06	.849±.05
Ours (CAM+$\mathcal{D}_\mathcal{G}$+$\mathcal{D}_{ano}$)	**.94±.03**	**.95±.01**	**.93±.02**	**.96±.01**	**.95±.01**	**.94±.02**	**.96±.01**	**.94±.02**	**.95±.01**	**.948±.02**

Table 4: AUROC value for hand gesture dataset [30]. In this table, one of the anomaly classes is known during training. Each column shows different anomaly class conditions. The value is an average of five seeds.

the normal class contains only one mode, and the evaluation is done with normal and anomaly classes which consists of many modes; hence, the number of evaluation images is much larger than that of training images. The results show that AnoVAE, AnoGAN, and AnoVAEGAN do not produce high performance, while our proposed method (CAM+$\mathcal{D}$) has the highest performance in all different datasets. The baseline method that uses only the discriminator value ($\mathcal{D}$ only) sometimes provides poor performance. This is because the generated images do not converge to an image similar to the anomaly samples in some conditions. However, for the hand gesture dataset, the output of the discriminator helps to increase the accuracy. Note that the performance of noisy MNIST has a lower value due to the high variance value of the artificial noise. When we set a lower σ, the AUROC is higher; the reason for high value is to prevent the ceiling effect for the experiment. The calculation time for prodocuing an attention map is around 10 milliseconds per image on Nvidia Titan X.

4.2.2 Semi-supervised setting

We investigate the performance for a case in which a single anomaly mode is known beforehand. This means we give only one mode for all evaluations; hence, we do some evaluations for each different known-anomaly condition on the noisy MNIST, hand gesture, and pigeon gesture dataset. As a baseline, we prepare a supervised regression-based CNN method (CNN regression) that has four convolutional layers with down-sampling, leaky ReLU, and batch normalization. We also prepare another baseline method that uses only weighted error by combined class activation map with known anomaly samples (CAM$_{ano}$).

Table 3 and Table 4 show the results. A few anomaly samples helped to improve the performance by a considerable margin. Our final hybrid method (CAM+$\mathcal{D}_\mathcal{G}$+$\mathcal{D}_{ano}$) has the best score in all datasets.

5. Conclusion

We presented an anomaly detection method that uses a masking method using discriminator's class activation maps for improving robustness against background noise. The core idea behind such an application is based on the study of discriminator's CAM evolution in an adversarial learning setting. Experimental evaluations showed that our method outperforms other previous anomaly detection methods in both self-supervised and semi-supervised case. As a limitation of our method, in special scenarios where anomaly signal lies in the background, our method would mask out such signals thus failing to detect abnormality in the input. However, such cases are rare since anomaly signals predominantly lie on the image foreground for most practical purposes. Possible future extensions include jointly learning the activation maps with the generator and discriminator training. There is also the potential for extending the concept of discriminator attention to a reinforcement learning setting, especially imitation learning methods that generate rewards from videos [9, 23, 34] for improved reward function estimation.

References

[1] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. *arXiv preprint arXiv:1805.06725*, 2018.

[2] J. An and S. Cho. Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE*, 2:1–18, 2015.

[3] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. *arXiv preprint arXiv:1804.04488*, 2018.

[4] A. Brown, A. Tuor, B. Hutchinson, and N. Nichols. Recurrent neural network attention mechanisms for interpretable system log anomaly detection. *arXiv preprint arXiv:1803.04967*, 2018.

[5] D. Carrera, F. Manganini, G. Boracchi, and E. Lanzarone. Defect detection in sem images of nanofibrous materials. *IEEE Transactions on Industrial Informatics*, 13(2):551–561, 2017.

[6] R. Chalapathy, A. K. Menon, and S. Chawla. Anomaly detection using one-class neural networks. *arXiv preprint arXiv:1802.06360*, 2018.

[7] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM computing surveys*, 2009.

[8] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847. IEEE, 2018.

[9] S. Chaudhury, D. Kimura, T.-H. Pham, A. Munawar, and R. Tachibana. Video imitation gan: Learning control policies by imitating raw videos using generative adversarial reward estimation. *arXiv preprint arXiv:1810.01108*, 2018.

[10] J. Chen, S. Sathe, C. Aggarwal, and D. Turaga. Outlier detection with autoencoder ensembles. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 90–98. SIAM, 2017.

[11] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[12] L. Deecke, R. Vandermeulen, L. Ruff, S. Mandt, and M. Kloft. Image anomaly detection with generative adversarial networks.

[13] A. Dosovitskiy and T. Brox. Generating images with perceptual similarity metrics based on deep networks. In *Advances in Neural Information Processing Systems*, pages 658–666, 2016.

[14] T. Fukuhara, M. Futaguchi, G. L. Hashimoto, T. Horinouchi, T. Imamura, N. Iwagaimi, T. Kouyama, S.-y. Murakami, M. Nakamura, K. Ogohara, et al. Large stationary gravity wave in the atmosphere of venus. *Nature Geoscience*, 10(2):85–88, 2017.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[16] M. Hauskrecht, I. Batal, M. Valko, S. Visweswaran, G. F. Cooper, and G. Clermont. Outlier detection for patient monitoring and alerting. *Journal of biomedical informatics*, 46(1):47–55, 2013.

[17] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognitio. *CVPR*, 2016.

[19] D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[20] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[21] JAXA. Akatsuki science data archive, `http://darts.jaxa.jp/planet/project/akatsuki/index.html.en`, 2017.

[22] I. Jolliffe. Principal component analysis. In *International encyclopedia of statistical science*, pages 1094–1096. Springer, 2011.

[23] D. Kimura, S. Chaudhury, R. Tachibana, and S. Dasgupta. Internal model from observations for reward shaping. *arXiv preprint arXiv:1806.01267*, 2018.

[24] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *ICLR*, 2014.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[26] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1558–1566, New York, New York, USA, 20–22 Jun 2016. PMLR.

[27] Y. LeCun and C. Cortes. MNIST handwritten digit database. 2010.

[28] W. Li, V. Mahadevan, and N. Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):18–32, 2014.

[29] S. Liang, Y. Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

[30] G. Marin, F. Dominio, and P. Zanuttigh. Hand gesture recognition with leap motion and kinect devices. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 1565–1569, Oct 2014.

[31] A. Munawar, P. Vinayavekhin, and G. De Magistris. Limiting the reconstruction capability of generative neural network using negative learning. In *Machine Learning for Signal Processing (MLSP), 2017 IEEE 27th International Workshop on*, pages 1–6. IEEE, 2017.

[32] P. Napoletano, F. Piccoli, and R. Schettini. Anomaly detection in nanofibrous materials by cnn-based self-similarity. *Sensors*, 18(1):209, 2018.

[33] E. J. Pauwels and O. Ambekar. One class classification for anomaly detection: Support vector data description revisited. In *Proceedings of the 11th International Conference on Advances in Data Mining: Applications and Theoretical Aspects*, ICDM'11, pages 25–39, Berlin, Heidelberg, 2011. Springer-Verlag.

[34] X. B. Peng, A. Kanazawa, S. Toyer, P. Abbeel, and S. Levine. Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow. *arXiv preprint arXiv:1810.00821*, 2018.

[35] M. Prastawa, E. Bullitt, S. Ho, and G. Gerig. A brain tumor segmentation framework based on outlier detection. *Medical image analysis*, 8(3):275–283, 2004.

[36] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[37] M. Sabokrou, M. Fathy, M. Hoseini, and R. Klette. Real-time anomaly detection and localization in crowded scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 56–62, 2015.

[38] M. Sabokrou, M. Khalooei, M. Fathy, and E. 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.

[39] M. Sakurada and T. Yairi. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *ACM MLSDA*, 2014.

[40] M. Sakurada and T. Yairi. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, page 4. ACM, 2014.

[41] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*, pages 146–157. Springer, 2017.

[42] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

[43] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[44] W. Sultani, C. Chen, and M. Shah. Real-world anomaly detection in surveillance videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[45] University of Padova. Hand gesture datasets, `http://lttm.dei.unipd.it/downloads/gesture/`, 2014.

[46] C. Wu, Y. Guo, and Y. Ma. Adaptive anomalies detection with deep network. In *Proceeding of the Seventh International Conference on Adaptive and Self-Adaptive Systems and Applications*, 2015.

[47] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe. Learning deep representations of appearance and motion for anomalous event detection. *arXiv preprint arXiv:1510.01553*, 2015.

[48] H. Yamaguchi, Y. Maki, and T. Yamagami. Yamaguchi fox-pigeon imitation test: a rapid test for dementia. *Dementia and geriatric cognitive disorders*, 2010.

[49] H. Yamaguchi, S. Takahashi, K. Kosaka, K. Okamoto, T. Yamazaki, M. Ikeda, M. Osawa, M. Amari, Y. Harigaya, S. Awata, et al. Yamaguchi fox–pigeon imitation test (yf-pit) for dementia in clinical practice. *Psychogeriatrics*, 11(4):221–226, 2011.

[50] A. Yamazaki, M. Yamada, Y. J. Lee, S. Watanabe, T. Horinouchi, S.-y. Murakami, T. Kouyama, K. Ogohara, T. Imamura, T. M. Sato, Y. Yamamoto, T. Fukuhara, H. Ando, K.-i. Sugiyama, S. Takagi, H. Kashimura, S. Ohtsuki, N. Hirata, G. L. Hashimoto, M. Suzuki, C. Hirose, M. Ueno, T. Satoh, T. Abe, N. Ishii, and M. Nakamura. Ultraviolet imager on venus orbiter akatsuki and its initial results. *Earth, Planets and Space*, 70(1):23, Feb 2018.

[51] H. Zenati, C. S. Foo, B. Lecouat, G. Manek, and V. R. Chandrasekhar. Efficient gan-based anomaly detection. *arXiv preprint arXiv:1802.06222*, 2018.

[52] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.