24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

Unsupervised Adversarial Learning of Anomaly Detection in the Wild

Amanda Berg[1,2] and **Michael Felsberg**[2] and **Jörgen Ahlberg**[1,2]

Abstract. Unsupervised learning of anomaly detection in high-dimensional data, such as images, is a challenging problem recently subject to intense research. Through careful modelling of the data distribution of normal samples, it is possible to detect deviant samples, so called anomalies. Generative Adversarial Networks (GANs) can model the highly complex, high-dimensional data distribution of normal image samples, and have shown to be a suitable approach to the problem. Previously published GAN-based anomaly detection methods often assume that anomaly-free data is available for training. However, this assumption is not valid in most real-life scenarios, a.k.a. in the wild. In this work, we evaluate the effects of anomaly contaminations in the training data on state-of-the-art GAN-based anomaly detection methods. As expected, detection performance deteriorates. To address this performance drop, we propose to add an additional encoder network already at training time and show that joint generator-encoder training stratifies the latent space, mitigating the problem with contaminated data. We show experimentally that the norm of a query image in this stratified latent space becomes a highly significant cue to discriminate anomalies from normal data. The proposed method achieves state-of-the-art performance on CIFAR-10 as well as on a large, previously untested dataset with cell images.

1 Introduction

Anomaly detection is the identification of *rare* samples, objects, or events that are regarded as anomalous compared to what is considered to be normal. Anomalies are sometimes also referred to as outliers [21]. Due to the quite general problem formulation, anomaly detection is applicable to a wide range of different fields, such as e.g. agriculture [10], medicine [33, 32], and finance [1, 2]. In the context of machine learning, anomaly detection can be *supervised*, *semi-supervised*, or *unsupervised*. This paper addresses *unsupervised* anomaly detection.

The objective of unsupervised anomaly detection is to detect previously unseen rare objects or events without any prior knowledge about these. The only information available is that the percentage of anomalies in the dataset is small, usually less than 1%. Since anomalies are rare and unknown to the user at training time, anomaly detection in most cases boils down to the problem of modelling the normal data distribution and defining a measurement in this space in order to classify samples as anomalous or normal. In high-dimensional data such as images, distances in the original space quickly lose descriptive power (curse of dimensionality) and a mapping to some more suitable space is required. Due to their latent space, Generative Adversarial Networks (GANs) [19] can model complex, high-dimensional data distributions [11] and are, therefore, suitable for anomaly detection in images. GAN-based methods also provide the ability to localize anomalies within images in contrast to many classical anomaly detection methods [32, 33]. Although partly addressed in recent works [3, 4, 12, 28, 32, 33, 37, 38], unsupervised anomaly detection still remains a challenging problem.

The main limitation of these previously published unsupervised GAN-based methods is their assumption that anomaly-free data is available for training. For this reason, we argue that they are not *truly* unsupervised, since completely anomaly-free data requires weak labelling. Anomaly contamination of GAN training data is expected to reduce detection performance [7]. In this work, we show that this is indeed the case for a recent, state-of-the-art GAN based anomaly detection method f-AnoGAN [32] and its variations.

Further, we show using t-SNE visualization [35] that anomalous and normal validation samples are scattered in latent space such that the GANs expressiveness with respect to classification is limited. To mitigate this problem, an image-to-latent-space encoder trained *jointly* with the generator is proposed. The joint training coupled with an image distance encoder loss enforces similar images to lie close to each other also in latent space. In this stratified latent space, latent vectors of anomalous samples prove to have shorter norms than those of normal samples. We show this empirically in a number of experiments on two datasets, based on CIFAR-10 and on a large cell-image dataset. Our approach achieves state-of-the-art performance in both cases.

Contributions

- We conduct an empirical study varying the amount of anomalies in the training data and measure the degradation of the anomaly detection in existing methods.
- We propose an approach to truly unsupervised anomaly detection based on simultaneous encoder training that improves results even when the training data is contaminated with anomalies.

2 Related work

Anomaly detection is an important problem relevant to a vast number of fields, e.g. malware intrusion detection [24], retinal damage detection [32, 33], and detection of anomalous events in surveillance videos [34]. A complete review of anomaly detection methods is beyond the scope of this paper, the interested reader is referred to [8, 9]. In the particular case of *unsupervised* anomaly detection, labels are unknown at training time. This paper is focused on unsupervised deep learning based anomaly detection of/in high-dimensional, non-sequential data with spatial coherence, i.e., images.

[1] Termisk Systemteknik AB, Sweden
Email: {amanda.,jorgen.ahl}berg@termisk.se
[2] Computer Vision Laboratory, Linköping University, Sweden
Email: {amanda.,jorgen.ahl,michael.fels}berg@liu.se

Classical methods for unsupervised anomaly detection include probabilistic methods that model the data distribution, e.g., by using a non-parametric Kernel Density Estimator (KDE) [29] as in [13] where it is applied to intrusion detection. Samples in low density areas are treated as anomalies. Another example of a probabilistic, parametric method is the RX anomaly detector [31]. Due to the *curse of dimensionality*, probabilistic methods are, however, not suitable for high-dimensional data such as images. Also, they typically do not provide the ability to *localize* anomalies in images.

In contrast, reconstruction-based methods provide the possibility to localize anomalies within images. The aim of these methods is to find a lower-dimensional latent space from which normal samples can be reconstructed. A query image is then projected onto this latent space and the reconstructed image is compared to the query image by some image distance measurement in order to discriminate anomalous cases. The latent space can be modelled using, e.g., Auto Encoders [36], Variational Auto Encoders [5], or Generative Adversarial Networks (GANs) [4, 12, 32, 33, 37, 38]. In the context of unsupervised anomaly detection, GANs were first introduced by Schlegl et. al. [33] (AnoGAN). They proposed to use a combination of the l_2-norm and a discrimination loss between a query image and its closest reconstruction match as an anomaly score. Based on this approach, Deecke et. al. [12] proposed a similar method (ADGAN) that improved the results slightly. In contrast to AnoGAN, ADGAN initialized the search in latent space for the closest match at multiple locations. Recently, and concurrent to this work, Schlegl et. al. [32] proposed f-AnoGAN, improving their method (AnoGAN) by replacing the Deep Convolutional GAN (DCGAN) [30] with a Wasserstein GAN (WGAN-GP) [20] and they also introduced an encoder that was trained separately for image to latent space mapping. The usage of an encoder instead of an iterative optimization procedure in order to speed up image to latent space mapping has also been explored by Zenati et. al. [37, 38] who employed an architecture similar to a Bidirectional GAN (BiGAN) [14] with pairs of (X, z) as input to the discriminator. We argue that the novelty of the proposed method compared to [37, 38] is the discussion of the impact of such an encoder on the structure of the latent space, and also the problem of training data contamination.

Ngo et al. [28] make the observation that the usual GAN objective encourages the distribution of generated samples to overlap with the real data, which may not be optimal in the case of anomaly detection. They further propose an *encirclement* loss that places generated samples at the boundary of the distribution and can then use the discriminator directly to discriminate anomalous samples.

Golan and El-Yaniv [18] proposed another type of method trained to map input images to a set of geometric transformations. In contrast to the reconstruction-based methods, it can not provide anomaly localization in images.

Some of the methods mentioned above [4, 12, 32, 33] claim to be unsupervised while at the same time assuming anomaly-free data for training. The acquisition of anomaly-free data requires labelling of data as normal. However, anomalous objects and/or events are rare and difficult to label in most real-world scenarios.

Beggel et. al. [7] conclude that the anomaly detection performance is reduced when the training set is contaminated with anomalies. They use an Adversarial Auto Encoder [26] to mitigate the problem by rejecting potential anomalies already during training. The proposed method improves detection results in the case of anomalies present in the training data in a different way. Instead of rejecting, we propose to use an encoder trained jointly with the GAN. As we show in our experiments, the anomalies need not to be rejected at training time, but mapped closer to the origin.

3 Method

The architecture of the proposed method is a combination of the progressive growing GAN (pGAN) [22] and ClusterGAN [27] but without class labels. An overview of the architecture at both training and testing time is presented in Figure 1. The generator and discriminator are equal to the ones in pGAN [22], while the encoder was inspired by ClusterGAN [27]. The architecture and objective function is further described below. At test time, the discriminator is discarded and the parameters of the generator and encoder are fixed. A query image Q is considered to be anomalous or not based on an anomaly score a.

3.1 Network architecture

One of the major drawbacks of AnoGAN [33] is its reliance on accurate reconstruction by a DCGAN [30]. DCGANs are, among other things, known to suffer from mode collapse [6]. For that reason, the inventors of AnoGAN replaced the DCGAN with a WGAN-GP [20] in f-AnoGAN. We instead propose to employ a progressive growing GAN (pGAN) [22]. pGAN also employs the WGAN-GP loss but incrementally adds new layers to the generator and discriminator while training. This approach has proven to increase the stability and robustness of a GAN, especially in the case of high-resolution images. The generator $\mathcal{G}(z : \theta_G)$ $\mathcal{G} : z \mapsto X_g$ and discriminator $\mathcal{D}(X : \theta_D)$ $\mathcal{D} : X \mapsto Y$ of the proposed method are equal to the ones used in pGAN. The prior, $z \sim \mathcal{N}(0, \mathrm{I}) \in \mathbb{R}^{N_z}$ is drawn from a Gaussian distribution.

Another update in f-AnoGAN compared to AnoGAN was the introduction of an encoder instead of the iterative search, which greatly improved detection speed. The encoder $\mathcal{E}(X : \theta_E)$ maps images to latent space $\mathcal{E} : X \mapsto \hat{z}$. In contrast to f-AnoGAN, the proposed method suggests to train the encoder $\mathcal{E}$ together with $\mathcal{G}$ and $\mathcal{D}$ in the same progressive manner as $\mathcal{G}$ and θ_G and θ_E are updated jointly. Various training strategies to learn an encoder have been explored by Dumoulin et. al. [15], although on different problems, and they emphasized the importance of learning $\mathcal{G}$ and $\mathcal{E}$ jointly. We make the same observation in our experiments.

Deecke et. al. [12] concluded that the discriminator is unsuitable for anomaly detection. While trained to separate real from generated images, thus forcing the two probability distributions to overlap, it is not trained to handle anomalous samples drawn from a different distribution. At test time, see Figure 1b, $\mathcal{D}$ is discarded and the parameters of $\mathcal{G}$ and $\mathcal{E}$, θ_G and θ_E, are fixed.

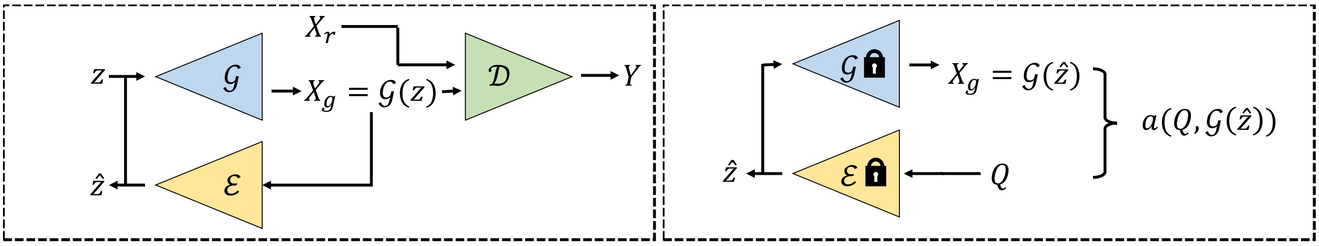
3.2 Objective function

Similar to f-AnoGAN and pGAN, we employ the WGAN-GP loss [20]. However, $\mathcal{E}$ is trained jointly with $\mathcal{G}$, not in a subsequent step as in f-AnoGAN. The GAN objective for the proposed method takes the following form:

$$\begin{aligned} &\min_{\theta_G, \theta_E} \max_{\theta_D} \mathop{\mathbb{E}}_{X \sim p_{\mathrm{data}}} q(\mathcal{D}(X)) + \mathop{\mathbb{E}}_{z \sim p_z} q(1 - \mathcal{D}(\mathcal{G}(z))) + \\ &\mathop{\mathbb{E}}_{z \sim p_z} \|(\mathcal{G}(z) - \mathcal{G}(\mathcal{E}(\mathcal{G}(z))))\|_1 \end{aligned} \tag{1}$$

where $q(x) = x$ since we use a Wasserstein loss [27]. The third term, $\mathop{\mathbb{E}}_{z \sim p_z} \|\mathcal{G}(z) - \mathcal{G}(\mathcal{E}(\mathcal{G}(z)))\|_1$, is new compared to previous works [20, 22, 32].

In contrast to BiGAN and ALI [15], the proposed architecture allows $\mathcal{G}$ and $\mathcal{E}$ to interact with each other during training, similar to the



(a) Training

(b) Testing

Figure 1: An overview of the proposed architecture at (a) training and (b) testing time. The encoder $\mathcal{E}$ is trained jointly with the generator $\mathcal{G}$. At test time, the discriminator $\mathcal{D}$ is discarded and the parameters of $\mathcal{G}$ and $\mathcal{E}$ are fixed. A query image Q is encoded and compared to its reconstruction $\mathcal{G}(\mathcal{E}(Q))$ in order to find an anomaly score a.

encoder used in ClusterGAN. However, while ClusterGAN computes the encoder loss in the latent space $z - \mathcal{E}(\mathcal{G}(z))$, we instead choose to compute the encoder loss in image space $\mathcal{G}(z) - \mathcal{G}(\mathcal{E}(\mathcal{G}(z)))$. The by $\mathcal{G}$ reconstructed query image Q should be the closest match in image space to Q rather than the closest match in latent space, since the anomaly score a, see next section, is partly based on a distance measure in image space. Also, the image space loss structures the latent space in a different way than the latent space loss, separating normal and anomalous samples, see the evaluation section.

3.3 Anomaly detection

We propose to use an anomaly score consisting of two terms, a normalized *residual* and an *origin distance* loss. The *residual* loss $\mathcal{L}_n$ for the query image $Q \in [0,1]^{W \times H \times D}$ is defined as the ℓ_2-norm between Q and its closest match $\mathcal{G}(\hat{z})$:

$$\mathcal{L}_n(Q, \mathcal{G}(\hat{z})) = \frac{1}{N_X} \left\| w(Q) - w(\mathcal{G}(\hat{z})) \right\|_2 \tag{2}$$

where $\hat{z} = \mathcal{E}(Q)$ is the encoded latent vector for image Q. In order to minimize the impact of the image contrast to the residual loss, we, unlike f-AnoGAN, propose to apply a minmax normalization $w(x)$ of images. The normalization $w(X) : [\min(X), \max(X)]^{W \times H \times D} \mapsto [0,1]^{W \times H \times D}$ where $\min(X)$ and $\max(X)$ finds the minimum and maximum elements of X, is defined as

$$w(X) = \frac{X - \min(X)}{\max(X) - \min(X)}, \tag{3}$$

where the division is element-wise and $N_X = W \cdot H \cdot D$. Without minmax normalization, low contrast samples yield low residual losses and vice versa.

Based on our observations regarding joint encoder and generator training and how that affects the structure of the latent space, we define an *origin distance* loss $\mathcal{L}_o$ as the distance in latent space from encoded vector $\hat{z}$ to the origin:

$$\mathcal{L}_o(\hat{z}) = -\frac{1}{\sqrt{N_z}} \left\| \hat{z} \right\|_2 . \tag{4}$$

The anomaly score is then defined as the convex combination between $\mathcal{L}_n$ and $\mathcal{L}_o$ as

$$a(Q, G(\hat{z})) = \lambda \mathcal{L}_n(Q, G(\hat{z})) + (1 - \lambda)(\mathcal{L}_o(\hat{z})), \tag{5}$$

where $\lambda \in [0,1]$. Samples are classified as anomalies if $a(Q, G(\hat{z})) > \alpha$.

In [32], f-AnoGAN used a convex combination of a residual loss and a *discrimination* loss as anomaly score. The discrimination loss depends on the difference between the discriminator output and the average discriminator output. In our experiments, adding the discriminator loss did not improve detection results.

4 Evaluation and results

4.1 Datasets

Two different datasets were used for evaluation in this work. The fully annotated KTH-Cellvideos dataset [17, 25], depicting different cells, and the CIFAR-10 dataset [23]. All training images were normalized to lie within range $[-1, 1]$.

4.1.1 CIFAR-10

The CIFAR-10 dataset [23] consists of 50000 $32 \times 32 \times 3$ training images in 10 classes (5000 images per class) and 10000 test images (1000 images per class). In this work, a subset of the dataset, denoted as $\mathrm{CIFAR}_{\mathrm{CAR}}$, was used. Images from the car class were treated as normal samples and images from all other classes as anomalous samples. The test set consisted of the 1000 normal test samples (car) and 1000 randomly chosen anomalous test samples from all other classes.

4.1.2 KTH-Cellvideos

The KTH-Cellvideos dataset [17, 25] consists of grayscale medical images featuring living cells in microscopy image sequences. About 50% of the labelled objects in the dataset is debris, e.g. bubbles, and they are labelled as such. Events such as mitosis (cell division) and apoptosis (cell death) are also labelled and segmentation masks are available for all cells. In this work, debris is treated as anomalies and cells as normal samples.

The labelled objects in the dataset were split into a training and a test set. All labelled objects (normal/debris) were cropped in a 64 by 64 neighbourhood. In addition, training samples were rotated three times by randomly generated angles. That is, each labelled object (except for the ones reserved for the test set) in the original dataset gave rise to four samples in the training dataset. In total, there were $N = N_n + N_a$ training patches where $N_n = 525657$ is the number of normal training patches and $N_a = \frac{\gamma N_n}{1-\gamma}$ the number of anomalous training patches. $\gamma \in [0,1]$ is the user-defined percentage of anomalies in the training data. The test set consisted of 256 normal test images and 256 anomaly test images.

4.2 Experiments

In order to evaluate the proposed method, a series of experiments was conducted. Code and detailed descriptions of network architectures and training configurations are available at `https://github.com/amandaberg/GANanomalyDetection`. For all experiments, $N_z = 512$ and $\lambda = 0.05$. Training of the proposed method was performed on an NVIDIA GTX1080 GPU, the batch size started at 128 and ended at 32 for KTH-Cellvideos and 64 for CIFAR-10. KTH-Cellvideos networks were trained for 48 epochs (6 epochs on full resolution) and CIFAR-10 networks were trained for 32 epochs (4 epochs on full resolution). Training time was about 36 hours for KTH-Cellvideos and about 12 hours for CIFAR-10.

All f-AnoGAN networks were trained with default parameters, batch size 16 and the dimension of z was 128. The KTH-Cellvideos networks were trained for 7 epochs. The training time was about 16 hours for the generator and about 1 hour for the encoder.

The default implementation of f-AnoGAN accepts images of dimension $64 \times 64 \times 1$ as input. Images in $\text{CIFAR}_{\text{CAR}}$ have dimension $32 \times 32 \times 3$. The default implementation was adapted by increasing the number of channels to 3 and removing one residual block in the discriminator, generator, and encoder respectively.

For dataset $\text{CIFAR}_{\text{CAR}}$, the f-AnoGAN generator was not able to generate visually pleasing images after seven epochs due to the low number of training samples (5000). Even training the network for as much as 70 epochs did not improve the detection performance. Therefore, since more iterations did not improve detection performance, f-AnoGAN was only trained for seven epochs for $\text{CIFAR}_{\text{CAR}}$.

Anomaly detection results are measured as the Area Under the Receiver Operating Characteristics (ROC) Curve (AUC) [16].

4.2.1 Encoder

Training jointly vs. training separately In the AnoGAN (note: not f-AnoGAN) paper [33], an iterative search was used to find the closest match to the query image Q in latent space. The drawbacks with this approach are that a) the optimization can get stuck in local minima, and b) evaluation was time-consuming. Here, we show that when training our method without an encoder and using an iterative search similar to the one in [33], encoded validation samples lie scattered all over the latent space, see Figure 2a. There is no separation between normal and anomalous samples.

In contrast, the introduction of an encoder stratifies the latent space. For f-AnoGAN, where the encoder is trained separately, the separation of samples (according to t-SNE) appears to be somewhat worse, Figure 2b, than for the proposed method, Figure 2c. AUC scores confirming this for the two methods are presented in the anomaly score section below. We believe that the joint encoder training enforces similar images to lie close to each other also in latent space. For the t-SNE plots, a perplexity value of 30 was used and the visualizations were consistent across multiple runs.

In Figure 3, the histograms of the coefficients of the encoded latent vectors for the validation samples from the KTH-Cellvideo dataset can be found. The networks were trained with 0% anomalies in the training data. It is clear that the proposed joint encoder training spreads the coefficients more evenly across the latent space, Figure 3c. These plots also explain why the norm of the latent vector, or the distance to origin, is not a discriminative loss in the case of f-AnoGAN. For f-AnoGAN, the samples end up on a hypercube, Figure 3a-b. In contrast, the density of coefficients is higher for anomalies close to the origin for the proposed method, Figure 3c.

In what follows, we give a possible explanation why the norms of latent variables representing anomalies are empirically smaller than those of normal images. Recall $z \sim \mathcal{N}(0, \mathrm{I}) \in \mathbb{R}^{N_z}$. In the implementation of pGAN, the prior z is normalised to unit length before being processed. A normalized *random* vector $z \in \mathbb{R}^{N_z}$ drawn from $\mathcal{N}(0, \mathrm{I})$ will have small coefficients. GAN training moves data clusters in the latent space away from the origin, otherwise the discriminator would not be able to separate them from the prior distribution, i.e. the noise. The encoder maps normal samples to clusters. Assuming high intra-class variability among anomalies, anomalies will be mapped away from the clusters and end up closer to the origin, i.e. the noise, and thus have smaller coefficients similar to a *random* vector.

When the training data is contaminated with anomalies, see Figure 2d and 2e, the confusion between normal and anomalous samples increases for f-AnoGAN. This is also confirmed in Table 2, (method d) where the norm-based loss $\mathcal{L}_o$ decreases AUC for f-AnoGAN. In contrast, the proposed method maintains the separability between samples (Figure 2f) even though the training data is contaminated with as much as 2% anomalies (method h).

Distance in image space vs. distance in latent space The proposed loss for the encoder is the third term in (1), hereby denoted by d_I:

$$d_I = \|\mathcal{G}(z) - \mathcal{G}(\mathcal{E}(\mathcal{G}(z)))\|_1 . \tag{6}$$

Generated images $\mathcal{G}(z)$ are compared with their reconstructed images $\mathcal{G}(\mathcal{E}(\mathcal{G}(z))))$ in image space. Another option would be to compare the distance between the latent vector z and the reconstructed latent vector $\hat{z} = \mathcal{E}(\mathcal{G}(z))$ in the latent space:

$$d_z = \|z - \mathcal{E}(\mathcal{G}(z))\|_1 . \tag{7}$$

Results for the proposed method using d_I and d_z are provided in Table 1 and t-SNE visualizations [35] of latent space projections are shown in Figure 4. The network was trained on the KTH-Cellvideos dataset with 0% anomalies in the training data. Comparing the distance in image space (d_I) is clearly preferable when it comes to separation of the validation samples in latent space. A good d_I implies a good d_z but the opposite is not true. We believe this is because d_I enforces similar images (in image space) to lie close to each other also in latent space. Small variations in z and $\hat{z}$ during reconstruction are forced to yield similar images.

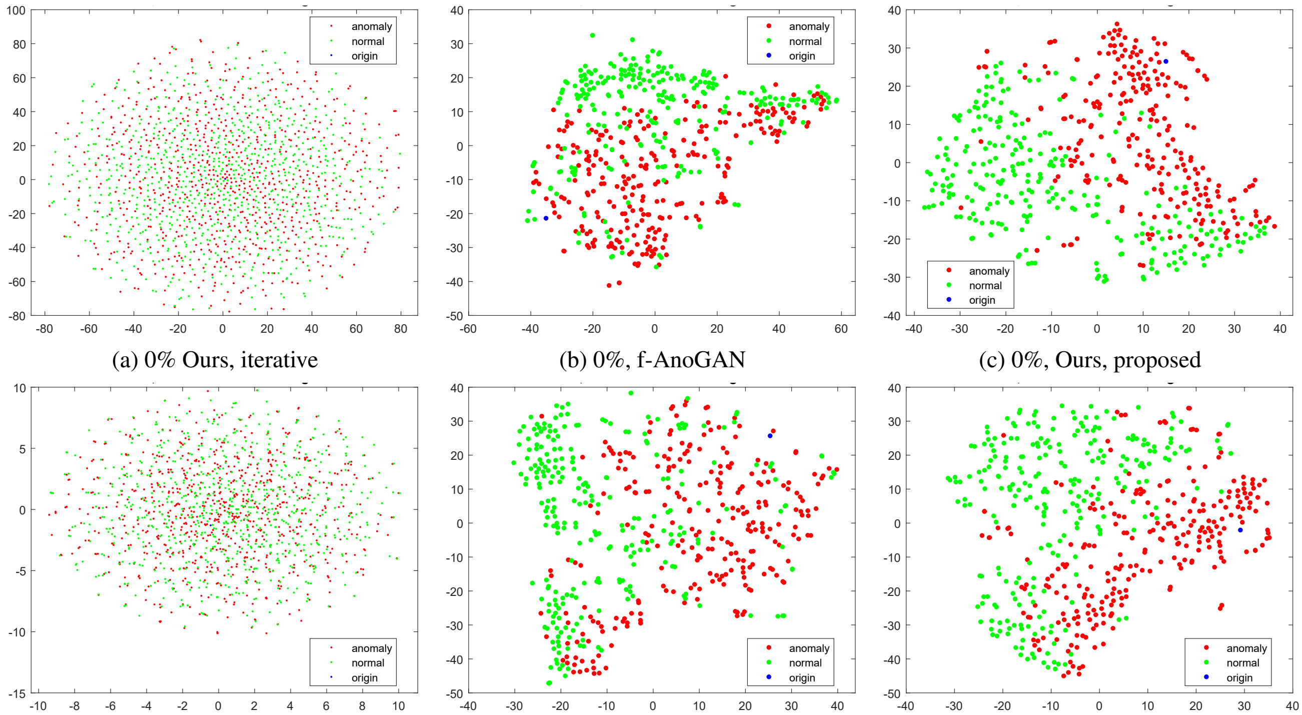
Table 1: AUC results for the proposed method with different encoder losses, d_z and the proposed d_I, for the KTH-Cellvideos dataset.

Encoder loss	$\mathcal{L}_n$	$\mathcal{L}_o$	$\mathcal{L}_n + \mathcal{L}_o$
d_I (proposed)	0.78	0.89	**0.90**
d_z	0.66	0.69	0.66

4.2.2 Anomaly score

As previously described, we propose to use a convex combination of a normalized residual loss $\mathcal{L}_n$ and a norm-based loss $\mathcal{L}_o$. In Table 2, AUC results for different combinations of these losses for both f-AnoGAN and our method can be seen. The networks were trained on two different datasets with two different percentages of anomalies in the training data. A is the anomaly score proposed in [32] and $\mathcal{L}_r$ is the residual loss, also from [32], without the proposed minmax normalization. Hence,

$$\mathcal{L}_r(Q, \mathcal{G}(\hat{z})) = \|Q - \mathcal{G}(\hat{z})\|_2 . \tag{8}$$



(a) 0% Ours, iterative (b) 0%, f-AnoGAN (c) 0%, Ours, proposed

(d) 2% Ours, iterative (e) 2%, f-AnoGAN (f) 2%, Ours, proposed

Figure 2: t-SNE visualization of validation samples projected to latent space for our method trained (a,d) without an encoder and iterative search for closest match, (c,f) with an encoder with latent space projection to find the closest match, and for (b,e) f-AnoGAN. The networks were tained on KTH-Cellvideos with (a-c) 0% and (d-f) 2% anomalies in the training data.

Table 2: AUC results for different anomaly losses for the proposed method and f-AnoGAN trained on three different datasets with 0% and 2% anomalies.

		CIFAR$_{\text{CAR}}$		KTH-Cellvideos	
	Method	0%	2%	0%	2%
a)	f-AnoGAN A	0.45	0.44	0.45	0.43
b)	f-AnoGAN $\mathcal{L}_r$	0.41	0.40	0.40	0.40
c)	f-AnoGAN $\mathcal{L}_n$	0.54	0.51	0.78	0.76
d)	f-AnoGAN $\mathcal{L}_o$	0.53	0.50	0.55	0.43
e)	Ours A	0.49	0.47	0.55	0.53
f)	Ours $\mathcal{L}_r$	0.42	0.41	0.51	0.51
g)	Ours $\mathcal{L}_n$	0.58	0.56	0.78	0.78
h)	Ours $\mathcal{L}_o$	0.70	0.63	0.89	0.87
i)	Ours, proposed	**0.72**	**0.64**	**0.90**	**0.89**

f-AnoGAN fails to separate normal from anomalous samples in both CIFAR$_{\text{CAR}}$ and KTH-Cellvideos (method a and b). Method a) is the default f-AnoGAN implementation. The AUC drastically improves for KTH-Cellvideos when we add the minmax normalization to the residual loss (method c). However, the norm-based loss $\mathcal{L}_o$ cannot discriminate between normal and anomalous samples (method d).

For our method, AUC increases when we add the minmax normalization and the origin distance loss $\mathcal{L}_o$ (method g and h). The proposed method, method i), which uses a convex combination of the two achieves state-of-the art results on both KTH-Cellvideos and CIFAR$_{\text{CAR}}$.

Regarding training dataset contamination with anomalous samples, there is no degradation in AUC for the proposed method on the dataset KTH-Cellvideos, in contrast to f-AnoGAN. Some examples of closest matches for the proposed method versus f-AnoGAN can be seen in Figure 5.

5 Conclusion

In this paper, we provide an empirical study of training anomaly detectors using contaminated training data and conclude that detection performance can deteriorate. We also propose an approach to truly unsupervised anomaly detection that can maintain results even when the training data is contaminated with anomalies[3].

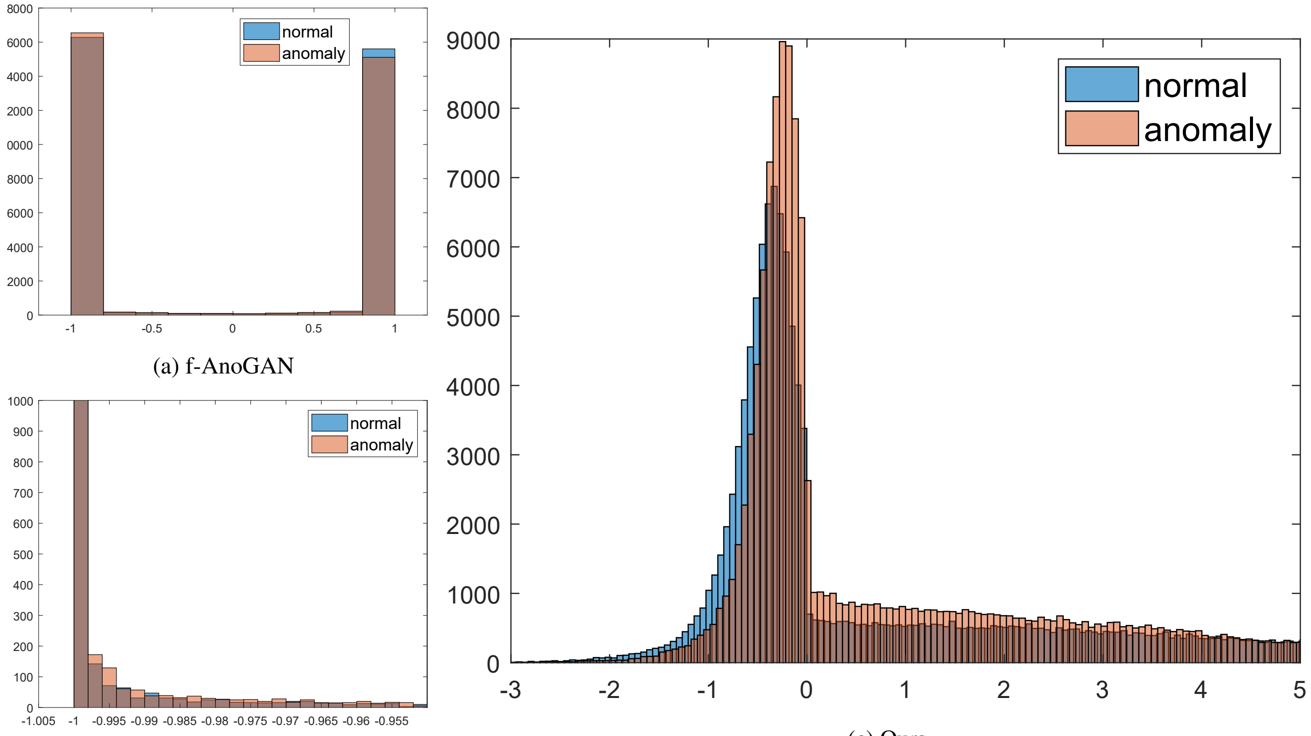
We conclude that *joint* generator and encoder training together with an encoder loss based on image distance is superior to training the encoder and generator separately. Joint generator and encoder training enforces similar images to lie close to each other and, thus, stratifies the latent space. At the same time, robustness to anomalies in the training data is improved.

Further work includes additional analysis of the structure of the latent space and how it is affected by different encoder losses as well as a more extensive study on the choice of the weight λ.

ACKNOWLEDGEMENTS

This research was funded by the Swedish Research Council through the project Learning Systems for Remote Thermography, grant no. D0570301, the project Energy Minimization for Computational Cameras (2014-6227), the project ELLIIT (the Strategic Area for ICT research, funded by the Swedish Government), as well as the European Union's Horizon 2020 reseach and innovation programme under grant agreement no. 783221, Aggregate FARming in the CLOUD (AFarCloud).

[3] Code is available at `https://github.com/amandaberg/GANanomalyDetection`

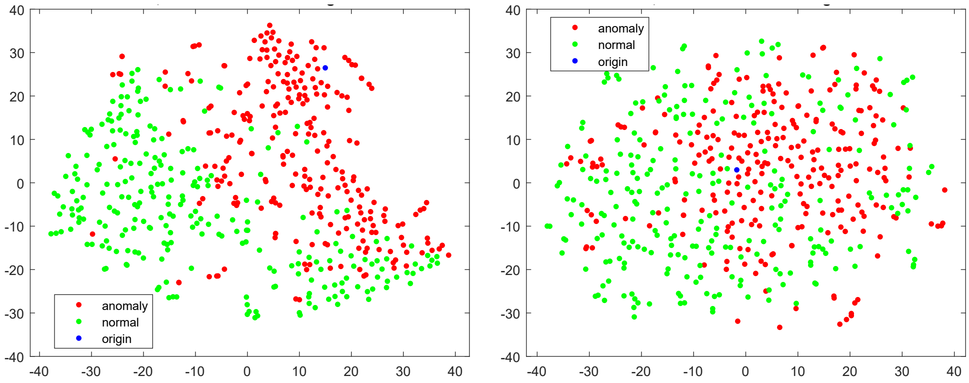


(a) f-AnoGAN

(b) f-AnoGAN

(c) Ours

Figure 3: Histogram plots for (a) f-AnoGAN (10 bins) and (c) the proposed method (600 bins) of the coefficients of the encoded latent vectors $\hat{z}$ for the validation samples of KTH-Cellvideos. (b) shows is a plot of the same data as (a), but with different axis limits and number of bins (1000 bins).



(a) d_I

(b) d_z

Figure 4: t-SNE visualization of validation samples projected to latent space when encoder training loss is based on the distance in (a) image space and (b) latent space.

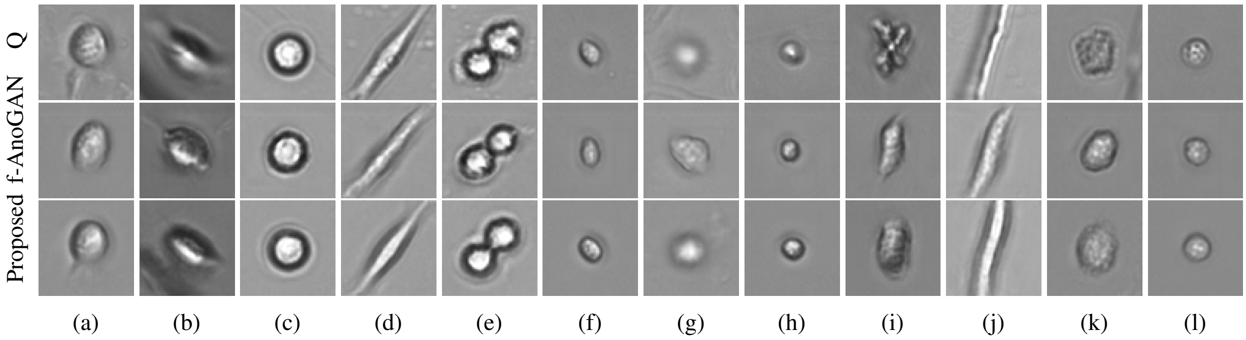


Figure 5: Closest matches for query image Q (row 1) by f-AnoGAN (row 2) and the proposed method (row 3). Columns (a)-(f) are examples of cells and columns (g)-(l) are examples of anomalies.

REFERENCES

[1] Aisha Abdallah, Mohd Aizaini Maarof, and Anazida Zainal, ‘Fraud Detection System: A Survey’, *Journal of Network and Computer Applications*, **68**, 90–113, (jun 2016).

[2] Mohiuddin Ahmed, Abdun Naser Mahmood, and Md. Rafiqul Islam, ‘A Survey of Anomaly Detection Techniques in Financial Domain’, *Future Generation Computer Systems*, **55**, 278–288, (feb 2016).

[3] Samet Akçay, Amir Atapour-Abarghouei, and Toby P. Breckon, ‘GANomaly: Semi-supervised Anomaly Detection via Adversarial Training’, in *2018 Asian Conference on Computer Vision (ACCV)*, eds., C. V. Jawahar, , Hongdong Li, , Greg Mori, , and Konrad Schindler, pp. 622–637. Springer International Publishing, (dec 2019).

[4] Samet Akçay, Amir Atapour-Abarghouei, and Toby P. Breckon, ‘Skip-GANomaly: Skip Connected and Adversarially Trained Encoder-Decoder Anomaly Detection’, in *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, (jan 2019).

[5] Jinwon An and Sungzoon Cho. Variational Autoencoder based Anomaly Detection using Reconstruction Probability, 2015.

[6] Martin Arjovsky, Soumith Chintala, and Léon Bottou, ‘Wasserstein GAN’, *CoRR*, (abs/1701.07875), (jan 2017).

[7] Laura Beggel, Michael Pfeiffer, and Bernd Bischl, ‘Robust Anomaly Detection in Images using Adversarial Autoencoders’, *CoRR*, (abs/1901.06355), (jan 2019).

[8] Raghavendra Chalapathy and Sanjay Chawla, ‘Deep Learning for Anomaly Detection: A Survey’, *CoRR*, (abs/1901.03407), (jan 2019).

[9] Varun Chandola, Arindam Banerjee, and Vipin Kumar, ‘Anomaly Detection: A Survey.’, *ACM Comput. Surv.*, **41**(3), 15:1–15:58, (2009).

[10] Peter Christiansen, Lars N Nielsen, Kim A Steen, Rasmus N Jørgensen, and Henrik Karstoft, ‘DeepAnomaly: Combining Background Subtraction and Deep Learning for Detecting Obstacles and Anomalies in an Agricultural Field.’, *Sensors (Basel, Switzerland)*, **16**(11), (nov 2016).

[11] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath, ‘Generative Adversarial Networks: An Overview’, *CoRR*, (abs/1710.07035), (oct 2017).

[12] Lucas Deecke, Robert Vandermeulen, Lukas Ruff, Stephan Mandt, and Marius Kloft, ‘Image Anomaly Detection with Generative Adversarial Networks’, in *Machine Learning and Knowledge Discovery in Databases*, pp. 3—-17. Springer International Publishing, (2018).

[13] Dit-Yan Yeung and C. Chow, ‘Parzen-Window Network Intrusion Detectors’, in *Object Recognition Supported by User Interaction for Service Robots*, volume 4, pp. 385–388. IEEE Comput. Soc.

[14] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell, ‘Adversarial Feature Learning’, *CoRR*, (abs/1605.09782), (may 2016).

[15] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville, ‘Adversarially Learned Inference’, *CoRR*, (abs/1606.00704), (jun 2016).

[16] Tom Fawcett, ‘An Introduction to ROC Analysis’, *Pattern Recognition Letters*, **27**(8), 861–874, (jun 2006).

[17] P M Gilbert, K L Havenstrite, K E G Magnusson, A Sacco, N A Leonardi, P Kraft, N K Nguyen, S Thrun, M P Lutolf, and H M Blau, ‘Substrate Elasticity Regulates Skeletal Muscle Stem Cell Self-Renewal in Culture’, *Science (New York, N.Y.)*, **329**(5995), 1078–81, (aug 2010).

[18] Izhak Golan and Ran El-Yaniv, ‘Deep Anomaly Detection Using Geometric Transformations’, in *NIPS’18 Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 9758–9769, (2018).

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, ‘Generative Adversarial Nets’, in *Advances in Neural Information Processing Systems 27*, eds., Z Ghahramani, M Welling, C Cortes, N D Lawrence, and K Q Weinberger, 2672–2680, Curran Associates, Inc., (2014).

[20] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville, ‘Improved Training of Wasserstein GANs’, in *NIPS’17 Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 5769—-5779, (mar 2017).

[21] Victoria J. Hodge and Jim Austin, ‘A Survey of Outlier Detection Methodologies’, *Artificial Intelligence Review*, **22**(2), 85–126, (oct 2004).

[22] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, ‘Progressive Growing of GANs for Improved Quality, Stability, and Variation’, in *ICLR 2018*, (oct 2017).

[23] Alex Krizhevsky, ‘Learning Multiple Layers of Features from Tiny Images’, *University of Toronto*, (2012).

[24] Donghwoon Kwon, Hyunjoo Kim, Jinoh Kim, Sang C. Suh, Ikkyun Kim, and Kuinam J. Kim, ‘A Survey of Deep Learning-Based Network Anomaly Detection’, *Cluster Computing*, 1–13, (sep 2017).

[25] Klas E. G. Magnusson, Joakim Jalden, Penney M. Gilbert, and Helen M. Blau, ‘Global Linking of Cell Tracks Using the Viterbi Algorithm’, *IEEE Transactions on Medical Imaging*, **34**(4), 911–929, (apr 2015).

[26] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian Goodfellow, ‘Adversarial Autoencoders’, in *International Conference on Learning Representations*, (2016).

[27] Sudipto Mukherjee, Himanshu Asnani, Eugene Lin, and Sreeram Kannan, ‘ClusterGAN : Latent Space Clustering in Generative Adversarial Networks’, *CoRR*, (abs/1809.03627), (sep 2018).

[28] Cuong Phuc Ngo, Amadeus Aristo Winarto, Connie Kou Khor Li, Sojeong Park, Farhan Akram, and Hwee Kuan Lee, ‘Fence GAN: Towards Better Anomaly Detection’, *CoRR*, (abs/1904.01209), (apr 2019).

[29] Emanuel Parzen, ‘On Estimation of a Probability Density Function and Mode’, *The Annals of Mathematical Statistics*, **33**(3), pp. 1065–1076, (1962).

[30] Alec Radford, Luke Metz, and Soumith Chintala, ‘Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks’, *CoRR*, (abs/1511.06434), (nov 2015).

[31] I S Reed and X Yu, ‘Adaptive Multiple-Band CFAR Detection of an Optical Pattern with Unknown Spectral Distribution’, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **38**(10), 1760–1770, (1990).

[32] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Georg Langs, and Ursula Schmidt-Erfurth, ‘f-AnoGAN: Fast Unsupervised Anomaly Detection with Generative Adversarial Networks’, *Medical Image Analysis*, 1–24, (2019).

[33] 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 *Information Processing in Medical Imaging*, 146—-157, (mar 2017).

[34] Waqas Sultani, Chen Chen, and Mubarak Shah, ‘Real-world Anomaly Detection in Surveillance Videos’, *CoRR*, (abs/1801.04264), (jan 2018).

[35] Laurens van der Maaten and Geoffrey Hinton, ‘Visualizing Data using t-SNE’, *Journal of Machine Learning Research*, **9**(Nov), 2579–2605, (2008).

[36] Yan Xia, Xudong Cao, Fang Wen, Gang Hua, and Jian Sun, ‘Learning Discriminative Reconstructions for Unsupervised Outlier Removal’, in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1511–1519. IEEE, (dec 2015).

[37] Houssam Zenati, Chuan Sheng Foo, Bruno Lecouat, Gaurav Manek, and Vijay Ramaseshan Chandrasekhar, ‘Efficient GAN-Based Anomaly Detection’, *CoRR*, (abs/1802.06222), (feb 2018).

[38] Houssam Zenati, Manon Romain, Chuan Sheng Foo, Bruno Lecouat, and Vijay Ramaseshan Chandrasekhar, ‘Adversarially Learned Anomaly Detection’, in *2018 IEEE International Conference on Data Mining (ICDM)*, pp. 727–736, (dec 2018).