Anomaly Detection using Generative Adversarial Networks

Reviewing methodological progress and challenges

Fiete Lüer
eMundo GmbH, Gofore Oyj
Hofmannstr. 25-27
Munich, Germany
fiete.lueer@e-mundo.de

Christian Böhm
University of Vienna
Währinger Straße 29
Vienna, Austria
christian.boehm@univie.ac.at

ABSTRACT

The applications of Generative Adversarial Networks (GANs) are just as diverse as their architectures, problem settings as well as challenges. A key area of research on GANs is anomaly detection where they are most often utilized when only the data of one class is readily available.
In this work, we organize, summarize and compare key concepts and challenges of anomaly detection based on GANs. Common problems which have to be investigated to progress the applicability of GANs are identified and discussed. This includes stability and time requirements during training as well as inference, the restriction of the latent space to produce solely data from the normal class distribution, contaminated training data as well as the composition of the resulting anomaly detection score. We discuss the problems using existing work as well as possible (partial) solutions, including related work from similar areas of research such as related generative models or novelty detection. Our findings are also relevant for a variety of closely related generative modeling approaches, such as autoencoders, and are of interest for areas of research tangent to anomaly detection such as image inpainting or image translation.

Keywords

Adversarial Generative Models, Anomaly Detection, Generative Adversarial Network, Novelty Detection, Outlier Detection

1 Introduction

Anomalies are commonly described as patterns in data not conforming to expected behavior [1]. Detecting anomalies is a frequent problem occuring on various types of data where the "expected behavior" is usually represented by a set of *normal* instances. Anomaly detection can often be framed as a binary classification problem where the task is to distinguish solely between normal and abnormal instances. Detecting anomalies is crucial for an abundance of industries: It is important to detect intrusions into networks [2], abnormal data in (spatio-temporal) climate data [3], patterns which indicate human diseases [4] or to minimize the risk of fraud [5]. In many of these applications, prompt actions are required to diminish or avoid damage, or to enable novel applications. For many such applications, recent progress has shifted the focus to deep learning algorithms with a major application being complex high dimensional data, e.g. image data, where handcrafting features is prone to errors. Neural networks can be used to detect anomalies in various ways. It is possible to compare the anomalous input data in a discriminative way using a threshold output score or forecasted values. Discriminative models such as Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) have found frequent use but can suffer from a variety of problems. These issues include differences between the predicted probability estimate of a class label and the ground truth correctness likelihood, which can be improved by calibrating the output [6]. Other challenges cannot easily be solved by improvements of the processing or training itself. Requiring a balanced dataset to avoid a shift to classifying only the predominant class is one of the most impactful problems. This led to a surge in generative modeling approaches such as (variational) autoencoders (VAEs) or Generative Adversarial Networks (GANs): Through one-class (OC) learning of only the normal class distribution, we can attempt to learn a generative procedure which should only produce normal class data, leveraging the reconstruction error between real and synthetic data.
For either method, an anomaly score is often calculated, e.g. based on the output score of discriminative networks or the difference between real and forecasted or synthetic values. This score can be compared to a predefined threshold which may not be exceeded by normal class input data (e.g. autoregressive neural networks [7]). Since the concept is based on learning a distribution from which specific data is generated, it can be applied to related models such as (V)AEs whose applicability we further discuss in the following sections.
Existing reviews [8] [9] on GAN-based anomaly detection focus on applications and types of data as well as architectures and metrics. Our work especially targets the investigation of more general challenges of GAN-based anomaly detection, which are mostly independent of the application, type of data or network architecture. Existing work on solving these challenges is clustered into distinctive categories and connected to related work in similar domains such as AEs.
We introduce foundations on anomaly detection as well as GANs in Section 2. The basic idea of AnoGAN [4], a central approach used to detect anomalies using GANs, is presented in Section 3. Subsequent advances and challenges are used to discuss and extend this framework in Section 4, focusing on practical as well as theoretical challenges. This especially includes i) the speed and quality of the training process, ii)

restrictions to the latent space, iii) contaminated training data, iv) novel anomaly score components and compositions to better extract relevant information as well as v) inference accuracy and speed. Lastly, this work is summarized in a general discussion in Section 5.[1]

2 Preliminary

2.1 Anomaly detection

An outlier, also called anomaly, intuitively is an observation deviating so much from other observations as to arouse suspicions that it was generated by a different mechanism [10]. The detection of anomalies is especially important in the medical domain where recent advances, mostly through novel deep learning approaches, have significantly improved the performance on various tasks and types of medical imaging. This includes MRI segmentation [11], CT scan generation [12] or Alzheimers prediction using F-FDG PET scans [13]) but also extends to time series data, sometimes even surpassing the performance of human professionals (e.g. on ECG data [14]). Due to its relevance, anomaly detection has been researched and reviewed for several decades. This includes very broad reviews [1], as well as surveys which focus on more specific data structures, such as graphs [15], where recent advances on Graph Neural Networks [16] open up interesting future applications.

Anomaly detection is closely related to novelty detection, i.e. the problem of finding novel data points. While the problem formulation and methodology *can* differ, there is a variety of tasks where it does not differ. Abati et al. [17] introduce novelty detection as the discrimination of observations that do not conform to a learned model of regularity, which can be translated to the detection of data points with novel, additional informational value. Abati et al. further argue that the reconstruction error or discriminative in-distribution tests express how well we *remember* an event and the *surprisal* is modeled by low probabilities of events under an expected model, which might be a network conditioned on normal samples, or by lowering a variational free energy. But just as in human memory, the rememberance itself might not be sufficient. Most anomaly detection tasks have a non-trivial and non-symmetric distribution of samples which can lead to only a small allowed margin of reconstruction errors in some part of a resulting learned manifold and a larger margin in another part which has to be accounted for in most applications to create a realiable convex hull. This is a major obstacle for OC anomaly detection where the manifold of normal class data, which corresponds to such "a learned model of regularity", is approximated.

The broad variety of anomaly detection methods itself ranges from fixed threshold values over more traditional machine learning approaches such as density based methods like Kernel Density Estimators [18] to more recent deep learning approaches, which are often based on reconstruction losses, e.g. autoregressive predictions of future values [7] or GANs. One of the advantages of reconstruction-based anomaly detection is that it allows the localization of the anomalies within high dimensional data such as images [19] which enables the application of further processing techniques such as segmentation or inpainting tasks. Furthermore, the quality of the reconstruction, measured by the degree it differs from the test input, can allow the assessment of an anomaly score which does not only give information on whether some data is anomalous but also how stark the anomaly is in comparison to existing data.

One definition of the anomaly detection setting [20] assumes that there exists a probability density function p_n from which normal data instances are generated:

$$X_n \sim p_n(x) = p(x|y=0) \tag{1}$$

with y being a label signaling that some data belongs to the normal class ($y = 0$) or abnormal class ($y = 1$). A dataset is usually composed of normal instances from X_n as well as anomalous instances from dataset X_a, with the latter being accordingly distributed with the anomalous distribution $X_a \sim p_a(x) = p(x|y=1)$. This results in a joint distribution containing both, normal and anomalous data points:

$$X_{total} \sim p_{total}(x) = (1-\Lambda)p_n(x) + \Lambda p_a(x) \tag{2}$$

where $\Lambda \in [0, 1]$ encodes the relative amount of anomalous points $p(y = 1)$. The task of anomaly detection then is to assess if data is drawn from the normal or abnormal data distribution.

2.2 Generative Adversarial Networks

Goodfellow et al. [21] introduce GANs as a framework consisting of two players, generator $\mathcal{G}$ and discriminator $\mathcal{D}$, following a minimax game with value function $V(\mathcal{G}, \mathcal{D})$. We call this the adversarial loss function $\mathcal{L}$:

$$\begin{aligned}\mathcal{L} &= \min_{\mathcal{G}} \max_{\mathcal{D}} V(\mathcal{D}, \mathcal{G}) \\ &= \mathbb{E}_{x\sim p_{data}(x)} \left[\log \mathcal{D}(x)\right] + \mathbb{E}_{z\sim p_z(z)} \left[\log(1 - \mathcal{D}(\mathcal{G}(z)))\right].\end{aligned} \tag{3}$$

In this adversarial game, the generator tries to create synthetic data similar to real data points $x \sim p_{data}, x \in \mathbb{R}^{d_x}$ with x consisting of n samples during training. In case of OC anomaly detection, all training samples belong to the normal class. The discriminator tries to distinguish data from the real training set from data which the generator $\mathcal{G}(z)$ synthetically generates using the latent noise vector $z \sim p_z$ as input. p_z often follows a Gaussian (commonly $z \sim \mathcal{N}(0, \mathrm{I})$) or uniform distribution with $z \in \mathbb{R}^{d_z}$ and $d_z \ll d_x$. The goal is to learn the parameters of $\mathcal{G}_\theta(z) \sim p_\theta$ s.t. p_θ is a potential candidate to represent p_{data}.

GANs have been especially useful on image data, and with CNNs being very effective on this domain, deep convolutional GANs (DCGANs) have been proposed by Radford et al. [22]. A central empirical result critical for many areas of research that have since evolved and which is most crucial for the detection of anomalies, is the existence of smooth transitions in latent space. This means that sampling similar noise z and using it as input for the generator also leads to the generation of similar images given sufficient training of the network.

This property has been observed on a variety of data structures: after proposed by DCGAN, this property was used for many tasks related to imaging which are based on this behavior, such as image generation and style transfer [23]. Furthermore, investigations on time series data using RNNs suggest similar behavior [24], [25] and Bojchevski et al. [26]

[1] We further publish an open source framework to evaluate various GAN-based anomaly detection approaches at `https://github.com/emundo/ecgan`.

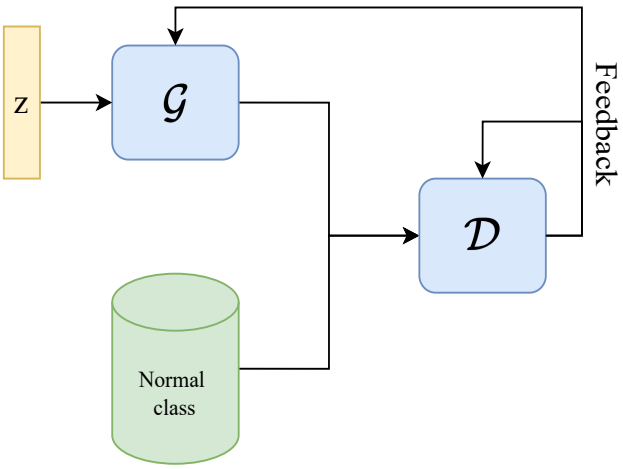

Figure 1: General training pipeline. The discriminator assesses the quality of generated and real data. This information is used to iteratively improve the discriminator as well as generator.

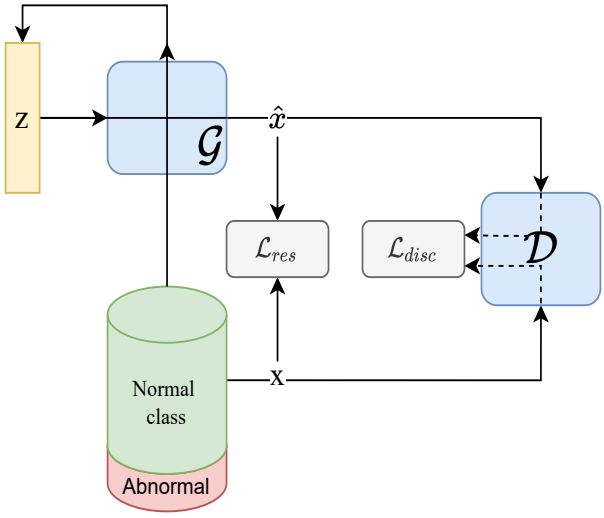

Figure 2: General anomaly detection pipeline. A given datum x is judged based on the discriminator as well as reconstruction loss between x and an arbitrarily similar sample $\hat{x}$. $\hat{x}$ can be retrieved by an explicit latent optimization minimizing the dissimilarity or implicitly by retrieving an inverse mapping using an encoding network.

visualize these transitions using attributes of graphs.

2.3 Anomaly Detection using Generative Adversarial Networks

The most widespread approach to GAN-based anomaly detection is based on learning the manifold of data of the normal class, meaning that the training data usually consists of only normal class data, i.e. $p_{data} \approx p_n$, such that $\Lambda = 0$ (Eq. 2) for the training data. The abnormal data is only used for validation and testing in the OC setting. It should be considered to include abnormal data during training as "synthetic" data in practical applications to improve the resulting learned generative procedure (e.g. using minimum likelihood regularization [27]) and to improve the separating hyperplane of the discriminator. Data points can be rated as normal or abnormal based on reconstructing the data x using $\mathcal{G}(z)$ (e.g. [28], [29]) and calculating a residual loss $\mathcal{L}_{res}$ using a well-defined and domain-dependent distance metric. Additionally, the discriminator can be used to assess if some datum belongs to the distribution represented by the generator by asserting a likelihood (e.g. [27]). A major approach in this area of research, AnoGAN, utilizes a combination of both components. The general training pipeline of GANs as well as the basic anomaly detection components have been visualized in Fig. 1 and Fig. 2. Before discussing AnoGAN and subsequently the most important developments in this domain and its applicability on a variety of data structures as well as practical or theoretical advances on this domain, we will briefly discuss existing applications and reviews of GAN-based anomaly detection.

Previous surveys on anomaly detection frequently focus on a broad array of methods (e.g. [30], [1]) or the general use of deep learning in anomaly detection and very general introductions to models or applications, e.g. Chalapathy et al. [31] for general deep learning based anomaly detection or Brophy et al. [32] for a more general use of GANs on time series data, which includes the use of anomaly detection. More recently, GAN-based anomaly detection has received a significant amount of attention and concurrent work exists which specifically reviews the usage of GANs for anomaly detection. Di Mattia et al. [9] perform a practical comparison of three major approaches in GAN-based anomaly detection. Sabuhi et al. [8] perform an exhaustive review of existing literature, investigating the domain of application, model architecture, datasets and evaluation metrics used by GAN-based anomaly detection. It shows a broad range of applications for GAN-based anomaly detection ranging from medical imaging [4] over the detection of deceptive reviews [33] to time series data [24] or videos [34]. While this work is an excellent resource for an overview of the currently used components, existing work rarely discusses the actual challenges of GAN-based anomaly detection apart from short remarks on training stability. Our work focuses on investigating these practical as well as theoretical obstacles which - to the best of our knowledge - have not been discussed in a systematic manner before.

3 The Fundamental Approach of AnoGAN

While GANs can be used to detect anomalies in a variety of ways, the procedure of AnoGAN by Schlegl et al. [4] can be considered central to most of these approaches and subsequent developments presented in this work are derived as adaptions from this approach. AnoGAN utilizes both, the difference in the features using the discriminator and its loss $\mathcal{L}_{disc}$, as well as the difference in data space, using the generator to calculate the absolute error as the residual loss $\mathcal{L}_{res}$. This is achieved by fully training a GAN on normal class data to learn the generator mapping $\mathcal{G} : z \mapsto x$. Dur-

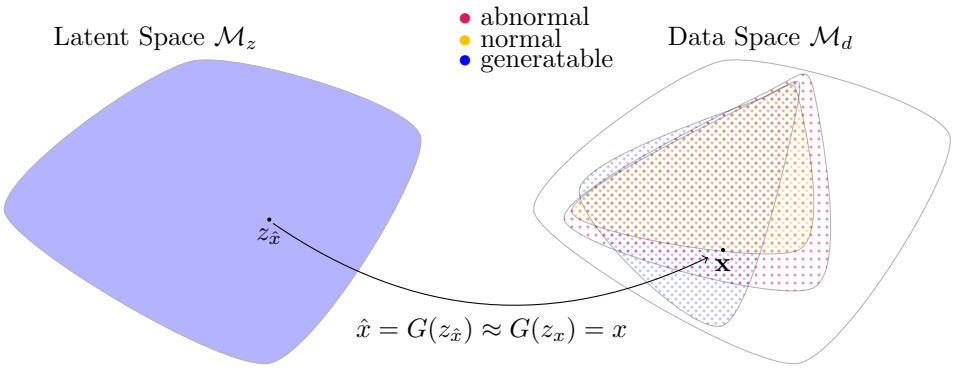

Figure 3: Simplified mapping between latent and data space. Each point in the lower dimensional latent space maps to a point in data space, resulting in a subspace of the data space which spans the data generatable by the GAN (blue area in $\mathcal{M}_d$). By training on only normal class data (yellow area in $\mathcal{M}_d$), it is attempted to restrict the generated data. In practice, the generatable data also includes abnormal data (red area in $\mathcal{M}_d$) and data which does not belong to the problem domain (e.g. if the task is to investigate CT images a valid image is generated but it is no realistic CT). Similarly, given a predefined threshold value for the discriminator error, one can separate a subset aiming to encompass normal class data as good as possible. By using a combination of both error components, we aim to minimize the amount of falsely classified samples.

ing inference, we try to determine if some novel datum x should be labeled as anomalous. To achieve this, it is required to find the noise vector z which is mapped as close as possible to x using the generator by interpolating through latent space: Initial noise z_1 is sampled randomly to generate $\mathcal{G}(z_1)$ in data space. The dissimilarity between x and $\mathcal{G}(z_1)$ is calculated. AnoGAN utilizes the absolute error but other common, well-defined distances have frequently been applied as proxies for the similarity of data points, including the euclidan distance. For time series data, RBF kernels have been a common choice and more time series specific measures such as dynamic time warping can easily be utilized. This distance is used to define a loss function to provide gradients that allow to move to some z_2 where $\mathcal{G}(z_2)$ is more similar to x than $\mathcal{G}(z_1)$. This is repeated Γ times to find the most similar image $\mathcal{G}(z_\Gamma)$ which can be constructed using the normal class manifold learned by the generator. Γ can either be a fixed value or be dynamically determined by a target similarity ϵ. A mixture of both strategies is commonly used where we try to find an ϵ-similar datum and interrupt optimization after a maximum of n_{max} steps in case the input is too dissimilar for the target similarity to be reached and to guarantee a maximum runtime. As soon as $\hat{x} = \mathcal{G}(z_\Gamma)$ is determined, it is compared to x using the similarity in data space by calculating the residual loss $\mathcal{L}_{res}$. In the case of images, and very similarly in the case of other regular data such as time series, $\mathcal{L}_{res}$ can be calculated by pointwise comparison of x and $\mathcal{G}(z_\Gamma)$:

$$\mathcal{L}_{res}(x, z_\Gamma) = |x - \mathcal{G}(z_\Gamma)|, \tag{4}$$

or another distance measure, depending on the target domain. It does not necessarily need to be the distance minimized during the latent optimization procedure, even though they do not differ in most applications. Afterwards, the discriminator is used to calculate the discriminative loss which enforces $\mathcal{G}(z_\Gamma)$ to lie on the learned manifold. Just as we try to force the generator to only produce healthy data given any valid z, the discriminator should only assign high confidence values to healthy data. The resulting discriminator loss is used by feeding $\mathcal{G}(z_\Gamma)$ to the discriminator, resulting in the following loss:

$$\mathcal{L}_{disc}(z_\Gamma, x) = \sigma(\mathcal{D}(\mathcal{G}(z_\Gamma), \alpha) \tag{5}$$

with σ being the sigmoid cross entropy which is used to describe the discriminator loss during training with logits $\mathcal{D}(\mathcal{G}(z_\Gamma))$ and targets $\alpha = 1$. The exact calculation of $\mathcal{L}_{disc}$ and $\mathcal{L}_{res}$ can differ: Schlegl et al. [4] further propose another $\mathcal{L}_{disc}$ based on feature matching

$$\mathcal{L}_{disc}(z_\Gamma, x) = |f(x) - f(G(z_\Gamma))|, \tag{6}$$

which has since been frequently applied [35], [29], [36]. The generator and discriminator are jointly used to calculate a combined loss which is a weighted sum of both components:

$$\mathcal{L}_{total}(z_\Gamma, x) = (1 - \lambda)\mathcal{L}_{res}(z_\Gamma) + \lambda\mathcal{L}_D(z_\Gamma). \tag{7}$$

Here, $\mathcal{L}_{total}(z_\Gamma, x)$ can be used directly to calculate an anomaly score. The anomaly score can be thresholded by some predefined or optimized τ to determine a label corresponding to x using $\mathcal{H} : \mathcal{L}_{total}(z_\Gamma, x) \mapsto \{0, 1\}$ with $\mathcal{H} = 0$ corresponding to normal samples and $\mathcal{H} = 1$ corresponding to abnormal samples respectively:

$$\mathcal{H}(\mathcal{L}_{total}(z_\Gamma, x), \tau) = \begin{cases} 0 \text{ if } \mathcal{L}_{total}(z_\Gamma, x) \leq \tau \\ 1 \text{ if } \mathcal{L}_{total}(z_\Gamma, x) > \tau \end{cases}. \tag{8}$$

Parts of the general anomaly detection procedure have been visualized in Fig. 3, the AnoGAN interpolation in Fig. 4. In practice, the sets are not convex and not easily interpolateable due to a complex loss surface when minimizing the dissimilarity.

4 Advances in GAN-based Anomaly Detection

We focus on five important challenges that should be considered when performing anomaly detection with GANs: speed and quality of the training process, restricting the latent

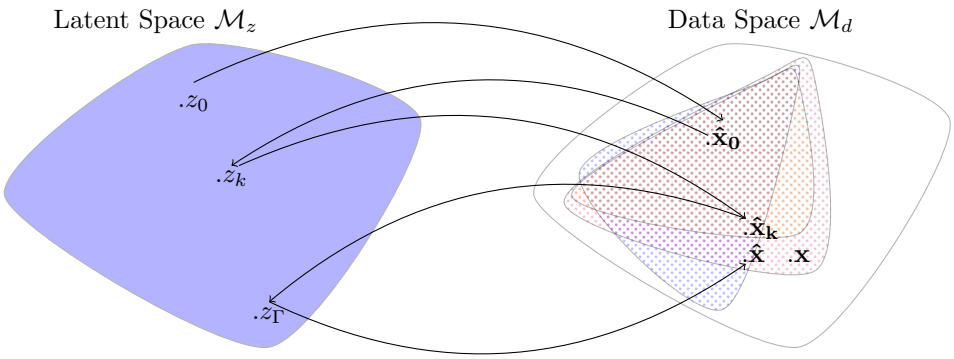

Figure 4: AnoGAN visualized. To approximate an abnormal sample x, initial latent vector z_0 is sampled and used to generate x_0. By minimizing the dissimilarity in data space, the latent vector z_Γ is retrieved, resulting in the best reconstruction $\hat{x}$ to x, i.e. minimizing the reconstruction error. The residual loss corresponds to the remaining difference in data space.

space, contaminated training data, the compositional choice of the anomaly score and inference performance. Some of these challenges introduce large and abstract areas of research and are related to each other. They can be split up further if desired, but shall act as a first reference point when considering anomaly detection with GANs or related generative models.

4.1 Speed and quality of the training process

Training GANs and the selection of suitable hyperparameters generally remains a non-trivial task: Choosing a sufficient latent distribution and dimensionality d_z ("the degree of compression"[37]) is not only crucial for inference time. It also has significant implications on the reconstruction of anomalous samples (e.g. [38]) and thus the reconstruction error and the anomaly detection performance in general. A small d_z can lead to non-convergence and insufficient preservation of information by the network because not all relevant features can be learned. If d_z is too large, training (and subsequently inference) are slowed down which is especially relevant if interpolation through latent space up until ϵ-similarity is chosen. Furthermore, too many irrelevant features might be learned, which hinders anomaly detection performance.

Due to the inherent instabilities of the originally proposed GAN architecture, which commonly causes mode collapse and further problems that hinder stable training, many researchers have resorted to practical workarounds. These workarounds include minibatch-discrimination [39] or using ensemble learning [27]. Since such methods often imply a significant overhead, novel objective functions such as the Wasserstein objective function (WGAN [40]) have found practical use. WGAN requires 1-Lipschitz continuity to stabilize training, which is enforced by weight clipping and has subsequently been replaced by ensuring that the norm of the gradient of the discriminator is penalized to stay (approximately) 1 in WGAN-GP [41]. This has also been successfully applied during training for anomaly detection, see e.g. [28], [35].

Depending on the underlying structure of data and networks used, different approaches to stabilize training can further be applied. These include spectral normalization [42], where a Lipschitz constraint to regularize the training is applied to the discriminator by normalizing all weights with the largest singular value, and a Lipschitz constraint to guarantee similar scores for data which is close in data space [43]. Progressive growing of GANs [44] attempts to stabilize and facilitate training by incrementally increasing the resolution of the data to iteratively increase the difficulty of the learned problem, which has since been used in the medical domain [45] as well as anomaly detection [46].

4.2 Restricting the latent space

While the previously presented approaches focus on the stability of the training process itself, the restriction of producing only normal class samples is less frequently discussed. In most settings, GANs are used to generalize and there is not always a need to completely restrict the latent space. While producing out-of-distribution data is often not only a byproduct but the explicit goal of image synthesis, it is difficult to restrict the generalizability of GANs to a subset of the generatable data. We initially assumed that the normal class data is a subset of $\mathbb{R}^{d_x}$ and $\mathcal{G} : z \mapsto x$ with $d_z \ll d_x$. Optimally, the latent space is a lower dimensional representation or approximation of the space of normal class data. However, it cannot be guaranteed that the generator is unable to produce abnormal class samples in the most common setting: Given $z \in \mathcal{N}(0, \mathrm{I})^{d_z}$, the sampled values during training will lie in a finite interval, likely with values only deviating up to a small number of standard deviations from zero in each dimension. Even though sampling large latent values is highly unlikely with $\lim_{z_i \to \infty} P(z_i) = 0 \; \forall \; \mathrm{i} \; \in \{1, .., d_z\}$, it is in principle possible to draw them since each value has a non-zero probability to be drawn from the (standard) Gaussian distribution [36]. Since AnoGAN only utilizes similarity based gradient information and disregards how likely the reconstruction is, it is possible to reconstruct highly unlikely data which has been observed by a variety of existing work [36], [47]. Tong et al. [43] discuss the sensitivity of the

reconstruction error in low density regions of normal class data and abnormal data close to the convex hull using autoencoders. They form the hypothesis that the model interpolates "too well for anomaly detection". To avoid this, we can restrict the possible values a priori, e.g. by using a uniform or truncated normal distribution, which introduces a tradeoff between variety and fidelity [48]. This has to the best of our knowledge not been applied to anomaly detection and requires a theoretical or empirical estimation of the cardinality of the problem domain. A different approach is to restrict the "accepted" reconstructions: given a multivariate standard Gaussian distributed latent space, it is possible to punish unlikely interpolations. This can be evaluated using the χ distribution where only certain deviations of the euclidean distance from origin of the latent vector, its latent norm, from its mode can be allowed by using predefined accepted standard deviations. This can also implicitly be learned if the latent norm is utilized as error component extending Eq. 7 using e.g. a grid search or SVM [36].

The ultimate goal of OC classification - and generation in this context - is to obtain a latent space where all instance from the latent space represent a datum from the given class [38]. With the approximation of the shape of the boundary separating in-class from out-of-class samples being the main goal, it becomes even more important to handle the data which is close to such boundary, i.e. the data close to the convex hull of normal data where at least $\mathcal{L}_{res}$ is low for anomalous points as well. Furthermore, $\mathcal{L}_{disc}$ is also low if not accounted for points close to the convex hull. This is increasingly important if we only interpolate through latent space: Since the generator might in general also produce abnormal data, moreso if d_z is chosen poorly, a bad initial random sampling might lead to some z_Γ producing $\mathcal{G}(z_\Gamma)$ which is far - the maximal allowed dissimilarity measured by ϵ - away from x, meaning that a normal point might be reconstructed with a high $\mathcal{L}_{res}$ or an abnormal point which could be reconstructed with a low $\mathcal{L}_{res}$ if ϵ is chosen poorly, leading to inaccurate classifications. This gets more important if the data is higher dimensional: Perera et al. [38] argue that especially complex data has a comparably weak novelty detection performance because the model automatically learns to represent some out-of-class objects if the shape is complex, leading to a low reconstruction error. This can further be problematic if the normality depends on the context, e.g. in the medical domain. This leads to a fuzzy and non-deterministic boundary since more information is required for a reliable detection. A stricter restriction of the manifold is thus required for many use cases. Sabokrou et al. [49] try to achieve a stricter decision boundary by enhancing inlier samples and distorting outliers, utilizing the addition of normal distributed noise. Koizumi et al. [50] generate a more suitable boundary by simulating non-normal data based on the Neyman-Pearson lemma, increasing the true positive rate by using an arbitrarily low false positive rate. They directly use an objective function which aims to increase the anomaly detection performance and utilize rejection sampling to simulate anomalous data and improve the resulting convex hull. Such a more distinctive boundary created by explicitly punishing the generation of abnormal data during training has received more attention recently (e.g. [46], [51]). Since the difficulties to establish the boundaries get larger the higher the dimensionality of the training data is, Liu et al. [52] utilize generative adversarial active learning to only generate informative potential outliers. Their approach shows to be especially effective on datasets with clusters of different shape as well as high ratios of irrelevant variables.

Further related approaches to divide anomalous from non-anomalous datasamples include Deep SVDDs [37] which have been used to learn a minimum-volume hypersphere in which the representations of, as close as possible to, all non-anomalous samples shall lie, regularizing the compactness of the learnt representation. Similarly, OCGAN [38] tackles the problem of restricting the latent representations by using a denoising autoencoder with a latent space which is forced to have bounded support enforced by the encoders output layer. Instead of only rewarding the generation of similar samples which might also be close to the convex hull, they add an adversarially trained discriminator in latent space to ensure that the representations of in-class examples resemble uniform random samples drawn from the same bounded space. The latent space is explored to produce potential out-of-class and high quality *informative-negative* examples which are fed into the network to steer the training to produce only in-class examples. Another approach which leverages the disciminator is to have a generator with the aim to produce weak anomalies to create such a distinct boundary [27]. Similarly, the use of a generator that does not attempt to mirror the true data distribution but improves low density areas to improve the generalization has been proposed [53]. Combining such approaches is also of interest includes using two generators with one generator learning the distribution of normal class data p_g^{normal} and one generator working as a *bad* generator learning p_g^{bad}. The discriminator receives data from p_{data}, p_g^{bad}, $p_g^{healthy}$, where data from p_g^{bad} would lead to an improved enforcement of the boundary between normal and abnormal data. Such a distinct boundary does not only have direct applications in common anomaly detection tasks: Neal et al. [54] use a GAN to generate counterfactual examples for unknown classes in image classification, reducing the influence of incorrect high-confidence predictions.

In a similar spirit to the previously mentioned investigation of the χ distribution mentioned before, Berg et al. [19] attempt to structure the latent space in such a way that the anomalous and normal samples are separated. The origin distance is used to measure the distance from the encoded image to the origin in the latent space, assuming that anomalous samples are farther away in latent space. It is possible that using the distance from a latent space centered around normality might allow a more distinctive mapping of points close to the convex hull than using only the reconstruction error.

4.3 Contaminated training data

Previously we have assumed that the data set contains correctly labeled instances. We would like to highlight an often overlooked, but in practice extremely important, problem: the contamination of data, meaning that most data sets rarely contain only correctly labeled data. The discussed generative models remain sensitive to outliers in training data and few anomalous points might contaminate the training set in OC classification [43]. One way to account for this is by assuming that a given sample (labeled as normal) can also come from an anomalous distribution [43] to allow a

margin of error.
Only very recently, researchers have begun to systematically investigate the impact of contamination, focusing on image data [19], [55], [46]. Some approaches are to reject potential anomalies during training [55] or to jointly train an encoder with the GAN in a progressive manner [19]. Salehi et al. [56] mention that the latent space may also primarily capture features that are shared by both, normal and anomalous data. Their work focuses on autoencoders but this holds true for GANs as well. They attempt to force their network to encapture features unique to the normal class using adversarial examples.

4.4 The compositional choice of the anomaly score

The multi-component anomaly score is one of the reasons why GANs are of interest for anomaly detection: Utilizing both, the reconstructive and the discriminative capabilities seems to increase the anomaly detection performance, most likely because different, complementary information is learned by the networks. In theory, either component should be sufficient to decide if some point is anomalous, with the reconstruction error allowing some limited explainability through visualization and it would be desirable to retrieve easy-to-interpret probabilities from the discriminator. In practice, it is usually impossible to train GANs to (near-) optimality and since the underlying approach is not unsupervised, acquiring and labeling the normal class data is seldom possible in the real world, especially without the before mentioned contamination. Exploring the practical differences and (dis-)advantages is important and current work rarely directly compares the performance of both: A variety of work argues that either the discriminator or the generator is partially unfit to deal with anomalous data without offering any experimental evidence. Most often, this is not specific to GANs, but focuses on the up- or downsides of the novelty likelihood, portrayed by the discrimination error in the GAN framework, or the residual error: On one hand, Deecke et al. [28] argue that the discrimination error is not equipped to deal with samples completely unlike the training data and only utilize the reconstruction error. On the other hand, it is argued that the reconstruction error does not always work very well in practice and suffers from a variety of problems such as intrinsic biases or instability if samples shall be reconstructed outside of the learned manifold [43], [57], [58], [51]. Pidhorskyi et al. [59] further argue that the reconstruction error only affects the noise portion of the model and does not include the signal portion.
Schlegl et al. [35] argue that only minimizing the reconstruction error is a subpar measure in regions of the latent space which are only sampled sparsely during training, but that the reconstruction error itself leads to better localization properties. Restricting the generator during training is an important body of future work (e.g. [56], [60], [61]).
Many problems presented here are shared with related frameworks such as (V)AEs, with An et al. [62] argueing that statistical anomaly detection methods, such as their proposed reconstruction probability, are a more objective, intuitive and robust anomaly score than the reconstruction error for VAEs. And while it is stated that they do "not require model specific thresholds for judging anomalies" using such probability based metrics, this is not necessarily true, even more so for GANs: using a fixed $\mathcal{G}$, the optimal discriminator is described by $\mathcal{D}^*_{\mathcal{G}} = \frac{p_{data}(x)}{p_{data}(x)+p_g(x)}$ and in case of training up to optimality means $p_g = p_{data}$ (having a Jensen-Shannon divergence of zero) where the optimal discriminator will always return $\frac{1}{2}$ for all training samples as well as produced samples from the normal class manifold [21]. This has to be taken into account since it implies that the discrimination error might increase over time for both, normal class and good synthetic samples if a target value of 1 is used (Eq. 5). During optimization of the anomaly score, this has to be accounted for, e.g. using non-linear punishments of low discriminator predictions and frequent reparameterization of the anomaly score. Schlegl et al. [35] average the discriminator output across a high number of normal training samples and substract the discriminator output of some test data from the average discriminator score. While the problem solved by this methodology differs, it is one possible solution to the aforementioned problem. Some work tries to work around this by not using the direct discriminator score but the feature matching loss [4], [63], [36].
Choi et al. [64] further investigate the susceptiblity to out-of-distribution errors, argueing that likelihood models are in fact very susceptible to out-of-distribution samples, assigning large likelihoods to such samples (see also [65]). Even if the computation of the likelihood would be exact, a one-tailed test to check if some data has a low likelihood does not hold for high-dimensional data: They consider an isotropic high dimensional gaussian distribution, where a datum at the origin has maximum likelihood but is considered highly atypical because most of the probability mass lies in an annulus of radius $\sqrt{d_z}$. While likelihoods can determine whether a point lies in the support of a distribution, they do not reveal where the probability mass is concentrated. This also motivates the previously mentioned restriction of the latent space based on the χ distribution. Although the density estimation of GANs should not be able to account for probability mass, the generative ensemble presented in Choi et al. [64] demonstrate anomaly detection capabilities by combining density estimation and uncertainty estimation.
Berg et al. [19] compare the use of the distance in image space and the distance in latent space as a discriminative factor of reconstructed data. They find that the distance in image space (the reconstruction error as introduced above) is clearly preferable when it comes to a separation of the validation samples in latent space, and that a good distance in data space implies a good distance in latent space in practice but not vice versa.
Summarizing, the arguments supporting or opposing either error component can differ, but commonly named restrictions are the domain-specificity of the reconstruction error and the lack of its general interpretability or the black-box of the discriminator. Since both perspectives commonly argue with the instability of the opposite component for out-of-distribution samples, a systematic evaluation would be of great use. For now, it is likely and reasonable that both methods struggle with such data and the performance often depends on the problem domain. Restrictions on the latent space named in the previous section can help to reduce this impact and utilizing information of the latent space can be relevant for anomaly detection as well. Currently, the optimal weighting of the respective components should be evaluated for each experiment. Lüer et al. [36] find that the optimal weighting parameter between the error components further changes over time during training. They improve de-

tection performance by considering non-linear relationships between the error components using a non-linear SVM instead of a linear weighting of the anomaly score.

4.5 Inference speed and accuracy

One of the most significant practical limitation of the AnoGAN approach is that inference requires high amounts of computational resources and time. The required resources depend on a variety of variables: Deecke et al. [28] sample from multiple initial random z_0 to reduce the influence of erroneous local minima caused by initial sampling in unsuitable regions of the latent space. In general, this should increase the detection performance but also significantly increases the computational costs. The interpolation further depends on a variety of parameters, including the optimizer and its learning rate schedule, the similarity measure and resulting target similarity ϵ as well as the maximum allowed iterations n_{max} for the latent space interpolation. The computational costs are not necessarily always a restriction, e.g. inference time is often less relevant if medical images are investigated. But many real-world use cases involve constant monitoring and thus evaluations of data which is changing at a fast pace. An example is time series analysis in intrusion detection or medical monitoring, where the non-trivial and non-convex optimization is too expensive. Additional to the inference speed, the optimizer, ϵ and the allowed interpolation iterations are strongly dependent of each other and strongly influence the anomaly detection performance: Each use case requires the selection of such parameters and a low ϵ or high amount of interpolation iterations might lead to the generation of anomalous samples while a high ϵ or low amount of allowed interpolation iterations leads to dissimilar samples which do not correspond to the true anomalousness. To avoid the costly and error prone parameter selection and optimization, one can learn an additional, inverse mapping to $\mathcal{G}$, $E(x) = \mathcal{G}^{-1}(x) = z$. Following prior work, especially adversarial learned inference [66] and adversarial feature learning [67], Zenati et al. [68] utilize a bidirectional GAN to learn such a mapping for the task of anomaly detection. The discriminator does not only use $\mathcal{G}(z)$ or x as input but $(E(x), x)$ or $(z, \mathcal{G}(z))$. The mapping of the encoder can either be learned jointly during training [68], [69] but also after training [35]. Berg et al. [19] report that the joint training led to a better separation of normal and anomalous samples in latent space.

The work is closely related to CycleGAN [70], which utilize a cycle consistency loss to obtain such an inverse mapping which is furthermore cycle consistent: While a mapping from a bidirectional GAN learns *some* mapping $E : x \to z$, the inverse mapping of CycleGAN enforces that $\mathcal{G}$ and E are inverses or inverse approximations *of each other*. This means that both mappings are bijections, which is also desirable for the task of consistent anomaly detection, i.e. $E(\mathcal{G}(z)) \approx z$ and $\mathcal{G}(E(x)) \approx x$ corresponding to the forward and backward cycle-consistency respectively. This has since been adopted for anomaly detection and leads to a significant speedup during inference [63], [71], [72].

Similarly to the BiGAN approach, a significant body of work has evolved around the use of adversarially trained autoencoders. While using a (variational) autoencoder deviates from the likelihood-free principle of traditional GANs, training and mode coverage are significantly improved. The extension of using the discriminator additionally to the autoencoder has shown to improve training as well as detection [29], [36], [73]. Here, the reconstruction error (Eq. 4) as well as the discriminator error, frequently using the feature matching error from Eq. 6, remain core components. Additionally, the latent error has been utilized to leverage properties of the latent space [36] or more generally to ensure learning to correctly encode normal class data. One notable example is GANomaly [69] which utilizes a second encoder to retrieve a latent representation of the generated image and learns to minimize the difference of the parametrizations of both encoders. Using such mappings to latent space, the model is not as dependent on random initial noise anymore, but subsequently depends even more on a sufficient inverse mapping, which can be difficult to learn by itself. Inverse mappings have since been widely adopted, not showing any systematic deficits in the performance and reporting a significant speedup during inference, see Table 1 (evaluated on the beatwise preprocessed MITBIH dataset [74] using a β-VAEGAN[36]). Latent optimization is based on the similarity measured by an RBF-Kernel optimized using Adam [75] with an adapting learning. The (runtime as well as anomaly detection) performance largely depends on the previously mentioned parameters, further including the batch size used during optimization since similarity is frequently calculated batch-wise. However, this averaged similarity can distort the results and should be avoided by evaluating the similarity per sample. Comparability across datasets and across variations of these parameters is very limited. Due to only utilizing the forward pass of the encoder-based architecture, higher batch sizes can lead to significantly faster inference while retaining equally good detection performance, improving inference time from 1.9 ms on GPU (2.4 ms on CPU) using a batch size of 1 to 0.08 ms on GPU using a batch size of 512.

Table 1: Inference times (consumer GPU) for a β-VAEGAN model using different anomaly detection approaches on 22427 test samples of the beatwise preprocessed MITBIH dataset.

	ϵ	n_{max}	batch size	**Runtime (ms)**
`AnoGAN`	0.005	500	1	1139.7 ± 483.3
`AnoGAN`	0.005	500	64	38.27 ± 2.7
`AnoGAN`	0.005	1000	1	2254.4 ± 997.6
`AnoGAN`	0.05	500	1	325.8 ± 560.5
`AnoGAN`	0.05	100	1	77.9 ± 117.7
`Encoder`	-	-	1	$\mathbf{1.9 \pm 0.3}$

5 Discussion

We identify five major, intertwined obstacles GAN-based anomaly detection needs to tackle: speed and quality of the training process, restrictions to the latent space, contaminated training data, novel anomaly score components and compositions as well as inference accuracy and speed. The **speed and quality of the training process** is a fundamental requirement to make GAN-based anomaly detection suitable for a wide array of applications. Adversarial generative models generalize well, which is not always a desired property during anomaly detection and requires **restrictions of the latent space** to produce only normal class

data. In general, losses are usually weighted

$$\mathcal{L}_{total} = \sum_{i=1}^{k} \mathcal{L}_i \cdot \lambda_i, \quad \sum_{i=1}^{k} \lambda_i = 1, \tag{9}$$

with $k = 1$ [28] or $k = 2$ [4] being common choices. But in general, **novel anomaly score components and compositions** such as explicit information about the latent space and non-linear weigthings can be incorporated. The performance and speed of AnoGAN depends of various parameters which are relevant for the search of $\mathcal{G}(z_\Gamma)$ - the optimizer, target dissimilarity ϵ and the maximum amount of iterations, as well as the component weighting λ and the anomaly score threshold τ. While increasing the amount of components can improve the performance, it also increases the time for optimizing λ_i and τ, making a large search space computational infeasible. The optimization process can be performed by a variety of mechanisms, including a naive grid search or by optimizing an SVM [36]. Using fixed weighting parameters and anomaly score thresholds can hinder performance and should not be utilized. The optimization is usually still cheap in terms of computational resources required in comparison to the GAN training, but one should be aware of the biases the optimization might introduce.

To avoid tuning interpolation parameters and **speed up inference and accuracy**, inverse mappings have been utilized in a variety of settings. Lastly, OC training still requires a definition and selection of normal class data, which can suffer from **contaminated data**. Additionally, more extensive research for non-image data will be required: Time series are usually split up into subsequences during training as well as evaluation. MAD-GAN [76] report more false positives for larger subsequence lengths. A possible explanation is that more training data is required to model larger time series due to the curse of dimensionality. Furthermore, the subsequence length likely influences the (in)stability of the training process.

Although a variety of challenges remain open and while a significant amount of progress will still be the crucial requirement for some practical applications, the recent success and interest in GANs on the domain of anomaly detection sketches their potential. Existing work focuses on the medical domain [8], which especially benefits from the GAN-based procedure: By only requiring the definition of a normal class, it is possible to detect unknown anomalies/pathologies which can be investigated in-depth. However, guarantees regarding the detection performance are crucial for many medical use cases and the problem of generating abnormal data in the one-class setting has not yet received sufficient attention.

More investigations into the comparison of the learned structure between GANs and similarly used representation learning networks (especially autoencoders) are of interest to improve understanding and ultimately performance of existing work. First attempts on comparing generative models with traditional models (e.g. [77]) commonly do not cover the pecularities of the respective models sufficiently, especially the sensitivity to hyperparameters. Automated machine learning approaches might allow a more robust and fair comparison.

GANs are most commonly used if the underlying dataset is heavily imbalanced, i.e. if anomalies are scarce and obtaining extensive data of the normal class is feasible. Most existing work focuses on the comparison of a narrow set of metrics, most commonly F_β and AUC. The F_β score is especially susceptible to imbalanced data and since the test data often also includes only few anomalies, the F_β score is less meaningful with the resulting score usually being very high and misrepresenting the actual capabilities. Using more balanced measures, such as be the phi coefficient, might improve the insights that can be gained.

This work has focused on GAN-based approaches that are explicitly used to classify data, usually as a feature extractor or via reconstructions, not implicitly e.g. through generating data to augment other anomaly detection methods or related tasks such as segmentation [78] or data imputation [79]. Learning the normal class manifold and using GANs to derive or detect changes from it is not limited to anomaly detection in a binary classification setting. It can also be extended to multiclass anomaly detection and sufficient learning of the class boundaries also allows to generate more specific data. In medicine, such data can be used for augmentation [80], [78], [81], [82], [83], [80] or practical training of medical professionals [84]. This has since allowed to improve anomaly detection performance on tasks with only few available training samples. More extensive evaluations on the implications and possible fallacies are still required. Approaches to protect the privacy of the patients (such as differential privacy, e.g. [85]), can be of high importance in this case and need to be incorporated in advance. Augmentations can also be of relevance during anomaly detection since it might be useful to use generative models to oversample infrequent normal samples (e.g. [86]) which are often close to the convex hull and are commonly misclassified as false positives.

GAN-based anomaly detection relies on significant amounts of normal class data. Since no negative information is incorporated in most settings, the detection of anomalies has proven to be difficult if the similarity in data space is high. If only the total anomaly score is available, it is more difficult to interpret the results, even though visualizations have started to give significant insights into the structure. Furthermore, training GANs is a significantly more ambitious, complex and time-consuming approach than solely utilizing a discriminative model. The training procedure requires a significant amount of domain knowledge and has a higher-dimensional hyperparameter space in comparison to many more traditional methods, not necessarily allowing the fast training of a reasonably good baseline. However, the use of implicit generative models can help improve generalizability on complex data manifolds, partially reducing the assumptions that have to be posed to define and detect anomalies. Implicit generative modeling thrives if it is difficult to explicitly describe anomalies, which is especially common for high dimensional data. GANs provide a general framework which can be applied to many different data structures and established components can be easily incorporated (e.g. the use of CNNs for image data). Another advantage of more recent methods is the speed of inference in comparison to many related OC anomaly detection methods. An additional benefit of using GANs is the reduced amount of data required in a semi-supervised learning process in comparison to similar fully supervised approaches which is especially relevant in domains where labeling is expensive [87].

Independent of the actual task, many of the listed challenges are applicable to other generative models. This in-

cludes (also variational or adversarial) autoencoders which are used in a similar way to detect anomalies, e.g. [88] or [51]. In-depth comparisons are left for future investigation. The similarities of their structure and the compression of data into a latent representation of a lower dimension than the data space and the use of the encoder, which is commonly used to calculate a residual/reconstruction error, are of high relevance. VAEs themselves suffer from the problem of blurriness in image space. This most likely results from the diffuse probability mass distribution over the data space due to the combination of the conditional independence assumption with the maximum likelihood training paradigm [66], [89]. Due to the comparably new state of inverse mapping, possible disadvantages in quality or learned structure have yet to be explored on a larger scale.

6 Concluding Remarks

Pairing generative models and their reconstructive capabilities with adversarial feedback has shown state-of-the-art performance on many complex and high dimensional problems, especially if data is heavily imbalanced towards one class. The generative procedure still suffers from unstable training, frequently leading to deviations from implicit generative modeling by combining autoencoders with an adversarial component. The addition of an inverse mapping from data space to latent space has further significantly improved inference times. While the achieved performances are remarkable, sensitive applications such as medical usecases - currently the predominant domain GAN-based anomaly detection is applied to [8] - still have to be treated with caution when utilized in practice. Apart from improving the efficiency and effectiveness of models themselves, future work especially needs to account for properties of the latent space and the possibility to generate data which is either abnormal or not belonging to the problem domain at all. This also includes work on utilizing negative information to strengthen the decision boundary. The influence of GAN-generated augmentation data on arbitrary anomaly detection approaches requires further investigations, especially regarding their theoretical limitations. Weightings of the individual anomaly detection components should be performed for each experiment and can significantly differ between models and datasets. Even though the applicability largely depends on the availability of normal class data, GANs can be very versatile and the trained GAN model can be used to perform a large variety of auxiliary tasks, making it a very interesting modelling approach. Many of our findings are also applicable more generally on other classification tasks using different generative models, such as diffusion models.

Acknowledgements

This work is supported by the Bavarian Research Foundation under grant AZ-1419-20.

7 REFERENCES

[1] Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM computing surveys (CSUR) **41**(3) (2009) 1–58

[2] Denning, D.E.: An intrusion-detection model. IEEE Transactions on software engineering (2) (1987) 222–232

[3] Das, M., Parthasarathy, S.: Anomaly detection and spatio-temporal analysis of global climate system. In: Proceedings of the third international workshop on knowledge discovery from sensor data. (2009) 142–150

[4] Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G.: Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: International conference on information processing in medical imaging, Springer (2017) 146–157

[5] Abdallah, A., Maarof, M.A., Zainal, A.: Fraud detection system: A survey. Journal of Network and Computer Applications **68** (2016) 90–113

[6] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning, PMLR (2017) 1321–1330

[7] Malhotra, P., Vig, L., Shroff, G., Agarwal, P.: Long short term memory networks for anomaly detection in time series. In: Proceedings. Volume 89., Presses universitaires de Louvain (2015)

[8] Sabuhi, M., Zhou, M., Bezemer, C.P., Musilek, P.: Applications of generative adversarial networks in anomaly detection: A systematic literature review. IEEE Access (2021)

[9] Di Mattia, F., Galeone, P., De Simoni, M., Ghelfi, E.: A survey on gans for anomaly detection. arXiv preprint arXiv:1906.11632 (2019)

[10] Hawkins, D.M.: Identification of outliers. Volume 11. Springer (1980)

[11] Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D.L., Erickson, B.J.: Deep learning for brain mri segmentation: state of the art and future directions. Journal of digital imaging **30**(4) (2017) 449–459

[12] Liu, F., Jang, H., Kijowski, R., Bradshaw, T., McMillan, A.B.: Deep learning mr imaging–based attenuation correction for pet/mr imaging. Radiology **286**(2) (2018) 676–684

[13] Ding, Y., Sohn, J.H., Kawczynski, M.G., Trivedi, H., Harnish, R., Jenkins, N.W., Lituiev, D., Copeland, T.P., Aboian, M.S., Mari Aparici, C., et al.: A deep learning model to predict a diagnosis of alzheimer disease by using 18f-fdg pet of the brain. Radiology **290**(2) (2019) 456–464

[14] Hannun, A.Y., Rajpurkar, P., Haghpanahi, M., Tison, G.H., Bourn, C., Turakhia, M.P., Ng, A.Y.: Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature medicine **25**(1) (2019) 65

[15] Akoglu, L., Tong, H., Koutra, D.: Graph based anomaly detection and description: a survey. Data mining and knowledge discovery **29**(3) (2015) 626–688

[16] Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)

[17] Abati, D., Porrello, A., Calderara, S., Cucchiara, R.: Latent space autoregression for novelty detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 481–490

[18] Yeung, D.Y., Chow, C.: Parzen-window network intrusion detectors. In: Object recognition supported by user interaction for service robots. Volume 4., IEEE (2002) 385–388

[19] Berg, A., Ahlberg, J., Felsberg, M.: Unsupervised learning of anomaly detection from contaminated image data using simultaneous encoder training. arXiv preprint arXiv:1905.11034 (2019)

[20] Pimentel, T., Monteiro, M., Viana, J., Veloso, A., Ziviani, N.: A generalized active learning approach for unsupervised anomaly detection. stat **1050** (2018) 23

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

[22] Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)

[23] Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 4401–4410

[24] Li, D., Chen, D., Jin, B., Shi, L., Goh, J., Ng, S.K.: Mad-gan: Multivariate anomaly detection for time series data with generative adversarial networks. In: International Conference on Artificial Neural Networks, Springer (2019) 703–716

[25] Lüer, F., Mautz, D., Böhm, C.: Anomaly detection in time series using generative adversarial networks. In: 2019 International Conference on Data Mining Workshops (ICDMW), IEEE (2019) 1047–1048

[26] Bojchevski, A., Shchur, O., Zügner, D., Günnemann, S.: Netgan: Generating graphs via random walks. arXiv preprint arXiv:1803.00816 (2018)

[27] Wang, C., Zhang, Y.M., Liu, C.L.: Anomaly detection via minimum likelihood generative adversarial networks. In: 2018 24th International Conference on Pattern Recognition (ICPR), IEEE (2018) 1121–1126

[28] Deecke, L., Vandermeulen, R., Ruff, L., Mandt, S., Kloft, M.: Image anomaly detection with generative adversarial networks. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer (2018) 3–17

[29] Zhou, B., Liu, S., Hooi, B., Cheng, X., Ye, J.: Beatgan: Anomalous rhythm detection using adversarially generated time series. In: IJCAI. (2019) 4433–4439

[30] Pimentel, M.A., Clifton, D.A., Clifton, L., Tarassenko, L.: A review of novelty detection. Signal Processing **99** (2014) 215–249

[31] Chalapathy, R., Chawla, S.: Deep learning for anomaly detection: A survey. arXiv preprint arXiv:1901.03407 (2019)

[32] Brophy, E., Wang, Z., She, Q., Ward, T.: Generative adversarial networks in time series: A survey and taxonomy. arXiv preprint arXiv:2107.11098 (2021)

[33] Aghakhani, H., Machiry, A., Nilizadeh, S., Kruegel, C., Vigna, G.: Detecting deceptive reviews using generative adversarial networks. In: 2018 IEEE Security and Privacy Workshops (SPW), IEEE (2018) 89–95

[34] Ravanbakhsh, M., Nabi, M., Sangineto, E., Marcenaro, L., Regazzoni, C., Sebe, N.: Abnormal event detection in videos using generative adversarial nets. In: 2017 IEEE International Conference on Image Processing (ICIP), IEEE (2017) 1577–1581

[35] Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U.: f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. Medical image analysis **54** (2019) 30–44

[36] Lüer, F., Dolgich, M., Weber, T., Böhm, C.: Adversarial anomaly detection using gaussian priors and nonlinear anomaly scores. In: 2023 International Conference on Data Mining Workshops (ICDMW), IEEE (2023)

[37] Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S.A., Binder, A., Müller, E., Kloft, M.: Deep one-class classification. In: International conference on machine learning. (2018) 4393–4402

[38] Perera, P., Nallapati, R., Xiang, B.: Ocgan: One-class novelty detection using gans with constrained latent representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 2898–2906

[39] Salimans, T., Goodfellow, I.J., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In Lee, D.D., Sugiyama, M., von Luxburg, U., Guyon, I., Garnett, R., eds.: Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain. (2016) 2226–2234

[40] Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan. arXiv preprint arXiv:1701.07875 (2017)

[41] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: Advances in neural information processing systems. (2017) 5767–5777

[42] Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings, OpenReview.net (2018)

[43] Tong, A., Wolf, G., Krishnaswamy, S.: A lipschitz-constrained anomaly discriminator framework. arXiv preprint arXiv:1905.10710 (2019)

[44] Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)

[45] Baur, C., Albarqouni, S., Navab, N.: Generating highly realistic images of skin lesions with gans. In: OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis. Springer (2018) 260–267

[46] Kimura, M., Yanagihara, T.: Anomaly detection using gans for visual inspection in noisy training data. In: Asian Conference on Computer Vision, Springer (2018) 373–385

[47] Yoon, S., Noh, Y.K., Park, F.C.: Autoencoding under normalization constraints. arXiv preprint arXiv:2105.05735 (2021)

[48] Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)

[49] Sabokrou, M., Khalooei, M., Fathy, M., Adeli, E.: Adversarially learned one-class classifier for novelty detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 3379–3388

[50] Koizumi, Y., Saito, S., Uematsu, H., Kawachi, Y., Harada, N.: Unsupervised detection of anomalous sound based on deep learning and the neyman–pearson lemma. IEEE/ACM Transactions on Audio, Speech, and Language Processing **27**(1) (2018) 212–224

[51] Kimura, D., Chaudhury, S., Narita, M., Munawar, A., Tachibana, R.: Adversarial discriminative attention for robust anomaly detection. In: The IEEE Winter Conference on Applications of Computer Vision. (2020) 2172–2181

[52] Liu, Y., Li, Z., Zhou, C., Jiang, Y., Sun, J., Wang, M., He, X.: Generative adversarial active learning for unsupervised outlier detection. IEEE Transactions on Knowledge and Data Engineering (2019)

[53] Dai, Z., Yang, Z., Yang, F., Cohen, W.W., Salakhutdinov, R.R.: Good semi-supervised learning that requires a bad gan. In: Advances in neural information processing systems. (2017) 6510–6520

[54] Neal, L., Olson, M., Fern, X., Wong, W.K., Li, F.: Open set learning with counterfactual images. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 613–628

[55] Beggel, L., Pfeiffer, M., Bischl, B.: Robust anomaly detection in images using adversarial autoencoders. arXiv preprint arXiv:1901.06355 (2019)

[56] Salehi, M., Arya, A., Pajoum, B., Otoofi, M., Shaeiri, A., Rohban, M.H., Rabiee, H.R.: Arae: Adversarially robust training of autoencoders improves novelty detection. Neural Networks **144** (2021) 726–736

[57] Šmídl, V., Bím, J., Pevnỳ, T.: Anomaly scores for generative models. arXiv preprint arXiv:1905.11890 (2019)

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

[59] Pidhorskyi, S., Almohsen, R., Doretto, G.: Generative probabilistic novelty detection with adversarial autoencoders. In: Advances in neural information processing systems. (2018) 6822–6833

[60] Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Adversarial examples are not bugs, they are features. In: Advances in Neural Information Processing Systems. (2019) 125–136

[61] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)

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

[63] Zenati, H., Romain, M., Foo, C.S., Lecouat, B., Chandrasekhar, V.: Adversarially learned anomaly detection. In: 2018 IEEE International Conference on Data Mining (ICDM), IEEE (2018) 727–736

[64] Choi, H., Jang, E., Alemi, A.A.: Waic, but why? generative ensembles for robust anomaly detection. arXiv preprint arXiv:1810.01392 (2018)

[65] Nalisnick, E., Matsukawa, A., Teh, Y.W., Gorur, D., Lakshminarayanan, B.: Do deep generative models know what they don't know? arXiv preprint arXiv:1810.09136 (2018)

[66] Dumoulin, V., Belghazi, I., Poole, B., Mastropietro, O., Lamb, A., Arjovsky, M., Courville, A.: Adversarially learned inference. arXiv preprint arXiv:1606.00704 (2016)

[67] Donahue, J., Krähenbühl, P., Darrell, T.: Adversarial feature learning. arXiv preprint arXiv:1605.09782 (2016)

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

[69] Akcay, S., Atapour-Abarghouei, A., Breckon, T.P.: Ganomaly: Semi-supervised anomaly detection via adversarial training. In: Asian conference on computer vision, Springer (2018) 622–637

[70] Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. (2017) 2223–2232

[71] Hirose, N., Sadeghian, A., Vázquez, M., Goebel, P., Savarese, S.: Gonet: A semi-supervised deep learning approach for traversability estimation. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE (2018) 3044–3051

[72] Armanious, K., Jiang, C., Abdulatif, S., Küstner, T., Gatidis, S., Yang, B.: Unsupervised medical image translation using cycle-medgan. In: 2019 27th European Signal Processing Conference (EUSIPCO), IEEE (2019) 1–5

[73] van Hespen, K.M., Zwanenburg, J.J., Dankbaar, J.W., Geerlings, M.I., Hendrikse, J., Kuijf, H.J.: An anomaly detection approach to identify chronic brain infarcts on mri. Scientific Reports **11**(1) (2021) 1–10

[74] Kachuee, M., Fazeli, S., Sarrafzadeh, M.: Ecg heartbeat classification: A deep transferable representation. In: 2018 IEEE international conference on healthcare informatics (ICHI), IEEE (2018) 443–444

[75] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

[76] Li, D., Chen, D., Jin, B., Shi, L., Goh, J., Ng, S.K.: Mad-gan: Multivariate anomaly detection for time series data with generative adversarial networks. In: International Conference on Artificial Neural Networks, Springer (2019) 703–716

[77] Škvára, V., Pevnỳ, T., Šmídl, V.: Are generative deep models for novelty detection truly better? arXiv preprint arXiv:1807.05027 (2018)

[78] Mahmood, F., Borders, D., Chen, R., McKay, G.N., Salimian, K.J., Baras, A., Durr, N.J.: Deep adversarial training for multi-organ nuclei segmentation in histopathology images. IEEE transactions on medical imaging (2019)

[79] Luo, Y., Cai, X., Zhang, Y., Xu, J., et al.: Multivariate time series imputation with generative adversarial networks. In: Advances in Neural Information Processing Systems. (2018) 1596–1607

[80] Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. Neurocomputing **321** (2018) 321–331

[81] Bowles, C., Chen, L., Guerrero, R., Bentley, P., Gunn, R., Hammers, A., Dickie, D.A., Hernández, M.V., Wardlaw, J., Rueckert, D.: Gan augmentation: Augmenting training data using generative adversarial networks. arXiv preprint arXiv:1810.10863 (2018)

[82] Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: Synthetic data augmentation using gan for improved liver lesion classification. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018), IEEE (2018) 289–293

[83] Han, C., Murao, K., Noguchi, T., Kawata, Y., Uchiyama, F., Rundo, L., Nakayama, H., Satoh, S.: Learning more with less: Conditional pggan-based data augmentation for brain metastases detection using highly-rough annotation on mr images. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. (2019) 119–127

[84] Han, C., Hayashi, H., Rundo, L., Araki, R., Shimoda, W., Muramatsu, S., Furukawa, Y., Mauri, G., Nakayama, H.: Gan-based synthetic brain mr image generation. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE (2018) 734–738

[85] Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. (2016) 308–318

[86] Lim, S.K., Loo, Y., Tran, N.T., Cheung, N.M., Roig, G., Elovici, Y.: Doping: Generative data augmentation for unsupervised anomaly detection with gan. In: 2018 IEEE International Conference on Data Mining (ICDM), IEEE (2018) 1122–1127

[87] Madani, A., Moradi, M., Karargyris, A., Syeda-Mahmood, T.: Semi-supervised learning with generative adversarial networks for chest x-ray classification with ability of data domain adaptation. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE (2018) 1038–1042

[88] Chen, X., Konukoglu, E.: Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders. arXiv preprint arXiv:1806.04972 (2018)

[89] Theis, L., Oord, A.v.d., Bethge, M.: A note on the evaluation of generative models. arXiv preprint arXiv:1511.01844 (2015)