Unsupervised Anomaly Detection with Adversarial Mirrored AutoEncoders

Gowthami Somepalli[1] **Yexin Wu**[2] **Yogesh Balaji**[1] **Bhanukiran Vinzamuri**[3] **Soheil Feizi**[1]

[1]{gowthami,yogesh,soheil}@cs.umd.edu, University of Maryland, College Park,
[2]yw2423@cornell.edu, Cornell University,
[3] bhanu.vinzamuri@ibm.com, IBM Research

Abstract

Detecting out-of-distribution (OOD) samples is of paramount importance in all machine learning applications. Deep generative modeling has emerged as a dominant paradigm to model complex data distributions without labels. However, prior work has shown that generative models tend to assign higher likelihoods to OOD samples compared to the data distribution on which they were trained. First, we propose Adversarial Mirrored Autoencoder (AMA), a variant of Adversarial Auto-Encoder, which uses a mirrored Wasserstein loss in the discriminator to enforce better semantic-level reconstruction. We also propose a latent space regularization to learn a compact manifold for in-distribution samples. The use of AMA produces better feature representations that improve anomaly detection performance. Second, we put forward an alternative measure of anomaly score to replace the reconstruction-based metric which has been traditionally used in generative model-based anomaly detection methods. Our method outperforms the current state-of-the-art methods for anomaly detection on several OOD detection benchmarks.

1 INTRODUCTION

When deploying machine learning models in the real world, we need to ensure safety and reliability along with the performance. The models which perform well on the training data can be easily fooled when deployed in the wild[Nguyen et al., 2014, Szegedy et al., 2013]. Recognizing novel or anomalous samples in the landscape of constantly changing data is considered an important problem in AI safety [Amodei et al., 2016]. Flagging anomalies is of utmost importance in many real-life applications of machine learning such as self-driving and medical diagnosis. The task of identifying such novel or anomalous samples has been formalized as Anomaly Detection (AD). This problem has been studied for several years under various names, as thoroughly discussed in [Hodge and Austin, 2004, Chandola et al., 2009, Chalapathy and Chawla, 2019, Ruff et al., 2020].

If the training data has the class labels available, several approaches have been proposed for OOD detection with a neural network classifier [Liang et al., 2017, Vyas et al., 2018, Hendrycks et al., 2018, Lee et al., 2018, Hsu et al., 2020]. While these methods perform exceptionally well, they cannot be used in unsupervised or one class classification scenarios where labels are missing or not available for most of the classes. For instance, in credit card fraud recognition task, we are presented with a lot of normal transactions but no additional label available for transaction type. A rather obvious choice in such cases is to learn the underlying distribution of the data using generative models. Within deep generative models, two styles of approaches are popular, (1) likelihood based techniques, in which we train likelihood models such as flows [Kingma and Dhariwal, 2018] or Autoregressive models [Salimans et al., 2017], and use the likelihood scores from the trained models to detect outliers,(2) Auto-Encoder (AE) style approaches where reconstruction error of a given input is used to recognize the anomalies. While likelihood based approaches allow computation of exact likelihood for a given sample, they are found to assign high likelihood score to out-of-distribution samples as noted in the recent literature [Choi et al., 2018, Nalisnick et al., 2018, Ren et al., 2019]. The goal of AE based approaches is to learn a good latent representation of data by either performing reconstruction or adversarial training with a discriminator [Schlegl et al., 2017, Zenati et al., 2018, Akçay et al., 2019, Ngo et al., 2019].

In this work, we focus on the latter, i.e., the AE style methods and resolve two specific problems associated with them. First, the ℓ_p loss used for reconstruction in Auto-Encoder (AE) methods compares only pixel-level errors

Accepted for the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021).

but does not capture the high-level structure in the image. [Munjal et al., 2020, Rosca et al., 2017] proposed to alleviate this problem by introducing an adversarial loss [Goodfellow et al., 2014]. While adversarial loss fixes the problem of blurry reconstructions in low-diversity settings such as CelebA [Liu et al., 2018] faces, quality of reconstruction remains poor for more diverse datasets such as CIFAR [Krizhevsky et al., 2009] with many unrelated subclasses like cats, and airplanes [Munjal et al., 2020]. We posit that this issue arises because the loss function in [Munjal et al., 2020] compares distributions for a batch of samples but not the individual samples themselves. Hence a cat image reconstructed as an airplane is still a feasible solution since both airplane and cat belong to the same unlabeled input distribution. To address this problem, we propose Mirrored Wasserstein loss, where for a given sample $\mathbf{x}$ and its reconstruction $\hat{\mathbf{x}}$, a discriminator measures the Wasserstein distance between the joint distribution $(\mathbf{x}, \mathbf{x})$ and $(\mathbf{x}, \hat{\mathbf{x}})$. Stacking the image with its reconstruction allows discriminator to not only minimize the distance between distributions of images and reconstructions as before but also ensures that each reconstruction is pushed closer to its ground truth. In § 3.1, we give an intuition on how the mirrored Wasserstein loss improves the reconstructions quality as compared to the Wasserstein loss.

The second problem associated with AE methods is the regularization of latent space. In absence of explicit regularization, the model ends up over-fitting the training distribution. Several regularization approaches have been proposed in the past [Kingma and Welling, 2013, Makhzani et al., 2015], typically with a goal of sampling from the latent distribution. In our work, we consider regularizing the latent space of the model from the perspective of anomaly detection. Ideally, we want the latent space to be smooth and compact for the samples within the distribution, while simultaneously pushing away out-of-distribution samples. To this end, we perform a simplex interpolation between latent representations of multiple samples in the training data to ensure that decoder reconstructions of these latents are also realistic [Berthelot et al., 2018]. For the training purposes, we generate synthetic negative samples by sampling from atypical set in latent space [Cover, 1999]. Our latent space regularizer ensures high quality reconstructions for in-distribution latent codes, thus improving the Anomaly Detection performance as demonstrated quantitatively in Section 4.

In summary, our main contributions are:

- We propose **Adversarial Mirrored AutoEncoder (AMA)**, an Auto-Encoder Discriminator style network that uses Mirrored Wasserstein loss in the discriminator to enforce better reconstructions on diverse datasets.
- We propose latent space regularization during training by performing **Simplex Interpolation** of normal samples in the latent space and by sampling *synthetic negatives* by **Atypical Selection** and optimizing the latent space to be away from them.
- We propose an anomaly score metric that generates likelihood-like estimate for a given sample with respect to the distribution of reconstruction scores of training data.

Please find the AMA implementation at `https://github.com/somepago/AMA`

2 RELATED WORK

The problem we are trying to solve is OOD detection in datasets with no class labels. This problem is studied under various names in the literature such as One-class classification, Novelty detection, and so on.

Likelihood based approaches: Since generative modeling techniques such as Glow [Kingma and Dhariwal, 2018], PixelRNN [Oord et al., 2016], or PixelCNN++ [Salimans et al., 2017] allow us to compute exact likelihood of data samples, several anomaly detection methods are built on the top of the likelihood estimates provided by these models. LLR [Ren et al., 2019] proposes to train two models, one on the background statistics of the training data by random sampling of pixels and second model on the training data itself. Given an image, anomaly score is given by the ratio of likelihoods predicted by these two models. WAIC [Choi et al., 2018] suggests to use Watanabe Akaike Information Criteria calculated over ensembles of generative model as anomaly scoring metric. Serrà et al. [2019] proposes an $\mathcal{S}$-criterion, which is calculated by subtracting complexity estimate of the image from the negative log-likelihood predicted by a PixelCNN++ or a Glow model. Typicality test is used for OOD detection in Nalisnick et al. [2019] by employing a Monte-Carlo estimate of the empirical entropy. A limitation of this method is that it needs multiple images at the same time for evaluation.

Some recent studies [Choi et al., 2018, Nalisnick et al., 2018, Ren et al., 2019] suggest that deep generative models trained on a dataset (say CIFAR-10) could assign higher likelihoods to some out-of-distribution (OOD) images (e.g. SVHN). This behaviour is persistent in a wide range of generative models such as VAE, Glow, PixelRNN, and PixelCNN++ and raises the question whether the likelihood provided by these approaches can be reliably used for detecting anomalies.

Auto-Encoders or GANs based methods: A number of methods proposed recently use a different kind of metric for scoring anomalies. In DeepSVDD [Ruff et al., 2018], an Encoder-Decoder network is used to learn the latent representations of the data while minimizing the volume of a lower-dimensional hypersphere that encloses them. They

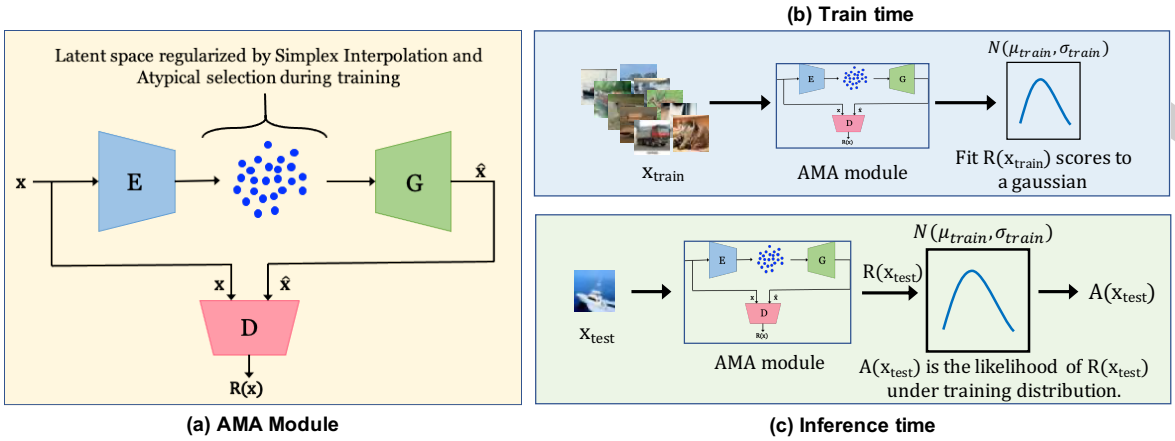

Figure 1: **AMA pipeline:** Our model consists of an Encoder **E**, a Generator **G** and a Discriminator **D**. (a) First we train the model on all the training samples by optimizing the min-max objective from Eq. 3 with latent space regularization discussed in § 3.2. (b) Next, we take the trained AMA module and pass the complete train data to generate R-scores using Eq. 5 and fit them to a Gaussian distribution. (c) During inference, given an image x_{test}, we first calculate $R(x_{test})$ by passing it through frozen-AMA module, and then $A(x_{test})$ using Eq. 6, which is essentially the likelihood of $R(x_{test})$ under the Gaussian curve we generated in (b). Lower the $A(x_{test})$, more the likelihood of the given test sample being anomalous.

hypothesize that anomalous data is likely to fall outside the sphere, and normal data is likely to fall inside the sphere. This technique is inspired by traditional SVDD (Support Vector Data Description) [Tax and Duin, 2004] where a hypersphere is used to separate normal samples from anomalies. Ano-GAN [Schlegl et al., 2017] is one of the first works that uses Generative Adversarial Nets (GANs) [Goodfellow et al., 2014] for anomaly detection. In this work, a GAN is trained only on normal samples. Since a GAN model is not invertible, an additional optimization is performed to find the closest latent representation for a given test sample. The anomaly score is computed as a combination of reconstruction loss and discriminator loss. FGAN [Ngo et al., 2019] trains a GAN on the normal samples and uses a combination of adversarial loss and dispersion loss (distance based loss in latent space) to discover anomalies. Akçay et al. [2019] use a series of Encoder, Decoder and Discriminator networks to optimize the reconstructions as well as distance between the representations. ALAD [Zenati et al., 2018] uses BiGAN [Donahue et al., 2016] to improve the latent representations of the data. Each of these methods use a combination of discriminator-based score and reconstruction error for detecting anomalies.

A recent survey by [Chalapathy and Chawla, 2019, Ruff et al., 2020] does a comprehensive study of anomaly detection approaches.

Negative Selection Algorithms (NSA): NSA is one of the early biologically inspired algorithms to solve one-class classification problem, first proposed by [Forrest et al., 1994] to detect data manipulation caused by computer viruses. The core idea is to generate synthetic negative samples which do not match normal samples using a search algorithm and use them to train a downstream, supervised anomaly classifier [Dasgupta and Majumdar, 2002]. Since the search space for negative samples for high dimension data can grow exponentially very large, it can be computationally very expensive to sample synthetic negatives [Jinyin and Dongyong, 2011]. Recent work by Sipple [2020] proposes a simpler approach to perform negative selection by using uniform sampling and building a binary classifier with positives and *synthetic negatives* to perform anomaly detection task.

3 ADVERSARIAL MIRRORED AUTOENCODER (AMA)

As discussed earlier, AMA consists of 2 major improvements over the conventional Auto-Encoder architectures: (i) Mirrored Wasserstein Loss, and (ii) Latent space regularization. These improvements help us outperform several state-of-the-art likelihood, as well as reconstruction-based anomaly detection methods. Fig. 1 shows an overview of our overall anomaly detection pipeline using AMA. In the following sub-sections, we discuss each of the components of our anomaly detection framework in detail.

3.1 MIRRORED WASSERSTEIN LOSS

For training auto-encoders, ℓ_1 or ℓ_2 reconstruction loss between the original image and its reconstruction, defined as $\|\mathbf{x} - \mathbf{x}_{rec}\|_p$, is typically used. Reconstruction losses based on ℓ_p distances result in blurred decodings, thus producing poor generative models. Also, the use of ℓ_p reconstruction

losses as anomaly scores, which is the standard technique used in Auto-Encoder based anomaly detection, has several limitations: (1) ℓ_p distances do not measure the perceptual similarity between images, which makes it hard to detect outliers that are semantically different, (2) A large ℓ_p reconstruction loss between input and its decoding can be an outcome of poor generative modeling and not because the image is an outlier.

Motivated by the success of Generative Adversarial Networks (GANs) in obtaining improved generations, a number of approaches replace the ℓ_p reconstruction losses in Auto-Encoders with an adversarial loss that captures high-level details in the image. While this loss is good enough to get good reconstructions in low-diversity datasets like MNIST, CelebA, but it is not enough to reconstruct diverse datasets like CIFAR-10 or Imagenet [Munjal et al., 2020].

A regular Wasserstein loss function only ensures the input and its generated sample both belong to the same distribution, but doesn't necessarily make input and its reconstruction look alike.

To resolve this problem, for a given sample $\mathbf{x} \sim \mathbb{P}_X$ and its reconstruction $\hat{\mathbf{x}} \sim \mathbb{P}_{\hat{X}}$, we perform a Wasserstein minimization between the joint distributions $\mathbb{P}_{X,X}$ and $\mathbb{P}_{X,\hat{X}}$. The discriminator now takes in stacked pairs of input images $(\mathbf{x}, \mathbf{x})$ and $(\mathbf{x}, \hat{\mathbf{x}})$. This clearly avoids the problems discussed in the previous part as the distribution $(\mathbf{x}, \mathbf{x})$ always has pairs of samples that are similar looking. If a car image is reconstructed as an airplane, the generated distribution will contain a (car, airplane) sample, which is never found in the input distribution $(\mathbf{x}, \mathbf{x})$. Hence, the model will aim to generate samples sharing the same semantics. Figure 2 shows the difference in image reconstructions using AMA with regular Wasserstein loss versus AMA with Mirrored Wasserstein loss. While both the models perform well in terms of image quality, we can see that for the first image, the ground truth is the number 30, and regular Wasserstein loss model is fitting number 9, though very unlike the ground truth, but still from the same distribution, while AMA with Mirrored Wasserstein loss is faithful to the ground truth and reconstructed a very similar looking 30.

Formally speaking, our model formulates a distribution of a set of samples $\mathbf{x} \sim \mathbb{P}_X$, using the Mirrored Wasserstein loss, as follows:

$$W(\mathbb{P}_{X,X}, \mathbb{P}_{X,\hat{X}}) = \max_{\mathbf{D} \in Lip-1} \mathop{\mathbb{E}}_{x \sim \mathbb{P}_X} [\mathbf{D}(\mathbf{x}, \mathbf{x}) - \mathbf{D}(\mathbf{x}, \hat{\mathbf{x}})] \tag{1}$$

where $\hat{\mathbf{x}} = \mathbf{G}(\mathbf{E}(\mathbf{x}))$ and *Lip-1* denotes the 1-Lipschitz constraint. Note that Eq. (1) is similar to the loss function of Wasserstein GAN [Martin Arjovsky and Bottou, 2017] with the only difference that discriminator $\mathbf{D}$ acts on the stacked images $(\mathbf{x}, \mathbf{x})$ and $(\mathbf{x}, \hat{\mathbf{x}})$. This is equivalent to minimizing the Wasserstein distance between conditional distributions $W(\mathbb{P}_{X|X}, \mathbb{P}_{\hat{X}|X})$. This model also shares similarities to discriminator architectures used in conditional image to image translations such as Pix2Pix [Isola et al., 2017].

Figure 2: Better reconstructions with Mirrored Wasserstein Loss. (a) Ground Truth (b) Reconstructions using AMA with regular Wasserstein loss (c) Reconstructions using AMA with Mirrored-Wasserstein loss. The quantitative comparisons are shown in Table 1

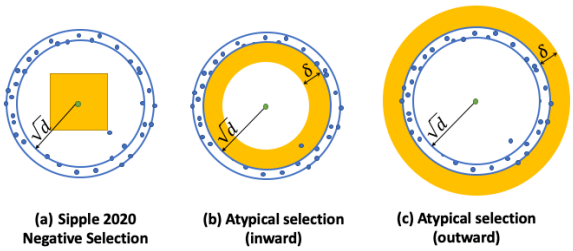

Figure 3: An illustration of the negative sampling from atypical set in the latent space. In each of the cases, typical set resides between the two blue d-dimensional spheres. Synthetic negative latents are drawn from yellow region (a) In [Sipple, 2020], a cube centered at the origin is used as the negative sampling region. Instead, we propose to sample the synthetic negatives closer to the typical set between the spheres (b) $\sqrt{d} - \delta$ and $\sqrt{d}$ or (c) $\sqrt{d}$ and $\sqrt{d} + \delta$.

Lemma 1 *If E and G are optimal encoder and generator networks, i.e.,* $\mathbb{P}_{X,\mathbf{G}(\mathbf{E}(X))} = \mathbb{P}_{X,X}$, *then* $\mathbf{x} = \mathbf{G}(\mathbf{E}(\mathbf{x}))$.

3.2 LATENT SPACE REGULARIZATION

The neural networks are universal approximators, and an Auto-Encoder trained without any constraints on the latent space will tend to overfit the training dataset. While several regularization schemes have been proposed, in this section, we develop our regularization framework adapted for the task for anomaly detection.

Simplex Interpolation in Latent space: Berthelot et al. [2018] showed that by forcing linear combination of latent codes of a pair of data points to look realistic after decoding, the encoder learns a better representation of data. This is demonstrated by improved performance on downstream tasks such as supervised learning and clustering. However, Sainburg et al. [2018] argues that pairwise interpolation

between samples of $\mathbf{x}$ proposed by Berthelot et al. [2018] does not reach all points within the latent distribution, and may not necessarily make the latent distribution compact. Hence, we propose to use simplex interpolation between i randomly selected points to make the manifold smoother and amenable.

Given k normal samples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ in a batch, we uniformly sample k scalars α_i from $[0, 0.5]$, we define an interpolated sample as:

$$\hat{\mathbf{x}}_{inter} = \mathbf{G}\left(\frac{1}{\sum \alpha_i}(\alpha_1 \mathbf{e}_1 + \alpha_2 \mathbf{e}_2 + \ldots \alpha_k \mathbf{e}_k)\right)$$
$$\mathbf{e}_i = \mathbf{E}(\mathbf{x}_i) \quad \forall i$$

Here, $\hat{\mathbf{x}}_{inter}$ denotes the interpolated latent reconstruction. A discriminator is then trained to distinguish between $(\mathbf{x}, \mathbf{x})$ pair and $(\mathbf{x}, \hat{\mathbf{x}}_{inter})$ pair, while the generator learns by trying to fool the discriminator. That is,

$$\min_{\mathbf{G}} \max_{\mathbf{D} \in Lip-1} \mathop{\mathbb{E}}_{x \sim \mathbb{P}_X} [\mathbf{D}(\mathbf{x}, \mathbf{x}) - \mathbf{D}(\mathbf{x}, \hat{\mathbf{x}}_{inter})]$$

where $\mathbf{x}$ is the input and $\hat{\mathbf{x}}_{inter}$ is the reconstruction of interpolation in latent space between $\mathbf{x}$ and other data points.

This ensures that the distribution of interpolated points follow the same distribution as the original data distribution, thereby improving the smoothness in the latent space. We use $k = 3$ in all our experiments. We empirically observe that larger values of k give marginal improvements.

Negative Sampling by Atypical Selection: In our experiments, we observed that regularization on the convex combination of latent codes of training samples works better if we also provide some negative examples, i.e., examples which should not look realistic. Since we are working in an unsupervised setting, we propose to generate synthetic negative samples in the by sampling from "atypical set" of the latent space distribution.

A typical set of a probability distribution is the set whose elements have information content close to that of the expected information. It is essentially the volume that not only covers most of mass of the distribution, but also reflects the properties of samples from the distribution. Due to the concentration of measure, a generative model will draw samples only from typical set [Cover, 1999]. Even though the typical set has the highest mass, it might not have the highest probability density. Recent works [Choi et al., 2018, Nalisnick et al., 2019] propose that normal samples reside in typical set while anomalies reside outside of typical set, sometimes even in high probability density region. Hence we propose to sample outside the typical set in the latent space to generate synthetic negatives.

The Gaussian Annulus Theorem [Blum et al., 2016, Vershynin, 2018] states that in a d-dimensional space, a typical set resides with high probability at a distance of $\sqrt{d}$ from the origin. In the absence of true negatives, we can obtain synthetic negatives by sampling the latents just outside and closer to the typical set than the origin and then use the generator for reconstruction. Although, our latent space is not inherently Gaussian, we observe that due to the ℓ_2 regularization placed on the latent encodings, most of the training samples' encodings are close to $\sqrt{d}$ in magnitude. We sample atypical points uniformly between spheres with radii $\sqrt{d}$ and $\sqrt{d} \pm \delta$ as illustrated in Fig. 3 (b) (c). We call this procedure **Atypical Selection**. The δ and the direction of the selection, *inward* or *outward* are hyperparameters which are chosen based on the true anomaly samples available during the validation time.

Sipple [2020] proposed a similar technique where *synthetic negatives* are sampled around the origin as shown in Fig. 3(a). We show in the appendix that Atypical Selection outperforms this style of negative selection across multiple benchmarks.

3.3 OVERALL OBJECTIVE

Let $\widetilde{\mathbb{Q}}_X$ be the distribution of all atypical samples and let $\mathbb{P}_X$ be the distribution of normal samples. We consider two different scenarios, first, when we don't have access to any anomalies during training, and the second case when we have access to a few anomalies.

Unsupervised case:

We train the AMA using the following min-max objective:

$$\min_{\mathbf{G}} \max_{\mathbf{D} \in Lip-1} \mathcal{L}_{normal} - \lambda_{neg} \mathcal{L}_{neg} \tag{2}$$

$\mathcal{L}_{normal}$ part of the loss is to improve the reconstructions of normal in-distribution samples . It consists of 3 terms, first term is inspired by Mirrored Wasserstein loss, making sure that reconstructions look like their ground truths, second term is to ensure the interpolated points look similar to normal points, and the third term is a regularization term on encodings. $-\lambda_{neg}\mathcal{L}_{neg}$ term penalizes the anomalies and ensures that anomalies are not reconstructed well. In this paper, since we assume that real anomalies are not available to us during training, we instead use generated *synthetic anomalies* in this term.

$$\begin{aligned}\mathcal{L}_{normal} =& \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_X}\Big[\mathbf{D}(\mathbf{x}, \mathbf{x}) - \mathbf{D}(\mathbf{x}, \hat{\mathbf{x}}) + \\ & \lambda_{inter}\left(\mathbf{D}(\mathbf{x}, \mathbf{x}) - \mathbf{D}(\mathbf{x}, \hat{\mathbf{x}}_{inter})\right) + \\ & \lambda_{reg}\|\mathbf{E}(\mathbf{x})\|\Big]\end{aligned} \tag{3}$$

$$\mathcal{L}_{neg} = \mathbb{E}_{\mathbf{x} \sim \widetilde{\mathbb{Q}}_X}\left[\mathbf{D}(\mathbf{x}, \mathbf{x}) - \mathbf{D}(\mathbf{x}, \hat{\mathbf{x}}_{neg})\right] \tag{4}$$

$\hat{\mathbf{x}}_{neg} = G(\mathbf{z}_{neg})$, where $\mathbf{z}_{neg}$ is the latent sampled by Atypical Selection. λ_{neg} is the Atypical Selection hyper-

parameter, λ_{inter} is the weight for the interpolation component, λ_{reg} is the latent space regularization weight and $\|E(x)\|$ acts as regularizer for the latent representations.

Semi-Supervised case: If we have a few true anomalies available during the training, we can use the same objective by using real anomalies instead of synthetic negatives in the $\mathcal{L}_{neg}$ term. Please refer to appendix for related experiments.

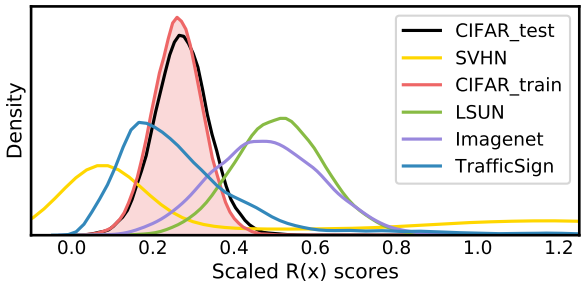

Figure 4: In this figure we show the density plots of R-scores predicted on various datasets by an AMA module trained on CIFAR-10 train data. We can note that CIFAR-10 train and test data distributions highly overlap. Depending on the dataset we are considering to be OOD distribution, R-scores of anomalous samples can trend lower or higher than the normal samples. If we use R-score directly to tag the anomalies, it will classify all the images with higher R-scores are anomalies, including many of the CIFAR-10 test samples. Meanwhile, all those to the left will be misidentified as normal samples. Instead of taking R-score on its face value, we propose to create A-score which weighs in the R-score of a given sample with respect to training data R-scores. R-score can be computed using Eq:5 and A-score can be computed using Eq:6

.

3.4 ANOMALY SCORE

Prior work in GAN-based anomaly detection used discriminator output as anomaly score [Schlegl et al., 2017, Ngo et al., 2019]. Zenati et al. [2018] proposed an improvement by computing the distance between a sample and its reconstruction in the feature space of the discriminator, **R-score** (or R(x) score used interchangeably), can be written as:

$$R(x) = \|f(x, x) - f(x, \mathbf{G}(\mathbf{E}(x)))\|_1 \tag{5}$$

where $f(\cdot, \cdot)$ is the penultimate layer of the discriminator.

In Zenati et al. [2018], authors claim that the anomalous samples will have higher $R(x)$ values compared to that of normal samples. While this is true for the datasets considered in Zenati et al. [2018], we observed a counter-intuitive behaviour in some OOD detection scenarios. In CIFAR-10 vs SVHN OOD detection experiment, our model and many other AE-based anomaly detectors (including [Zenati et al., 2018]) assign lower R-scores to OOD samples as shown in Fig. 4. This behavior is similar to the observations in [Nalisnick et al., 2018, Choi et al., 2018] where sample likelihoods are used as anomaly scores. Even though the R-scores distribution of test CIFAR-10 samples overlaps with training distribution quite well, if we use the R-scores to compute AUROC, it results in a very low AUC value (0.442 from Table 1), meaning most of the anomalies are classified as normal samples. This suggests that this reconstruction-based score is not a robust anomaly scoring function in all OOD detection scenarios.

Hence we propose the following technique: (i) fit the R-scores of training data to a Gaussian distribution (ii) compute the anomaly score for a given test sample x_i as the likelihood of $R(x_i)$ under the Gaussian distribution. The proposed anomaly metric, **A-score** (or A(x) score used interchangeably) can be written as:

$$A(x_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(R(x_i)-\mu)^2/2\sigma^2} \tag{6}$$

where μ is the mean and σ^2 is the variance of the distribution of R-scores over the training data.

A-score metric measures how similar the behaviour of test-time sample to that of training data, while R-score looks at relative behaviour of samples only at the test time.

4 EXPERIMENTS AND RESULTS

4.1 EXPERIMENTAL SETTING

Datasets: Following the setting in [Ren et al., 2019, Choi et al., 2018, Serrà et al., 2019, Nalisnick et al., 2019], we use CIFAR-10 [Krizhevsky et al., 2009], SVHN [Sermanet et al., 2012] and FashionMNIST [Xiao et al., 2017] as normal datasets. We evaluate the performance of the models when the anomalies are coming from each of the OOD datasets, ImageNet (resized)[Deng et al., 2009], CIFAR-100 [Krizhevsky et al., 2009], MNIST [LeCun et al., 1998], Omniglot [Lake et al., 2019]. We also consider the case when anomalies arise within the same data manifold (i.e. same dataset). We evaluated this scenario on CIFAR-10 and MNIST datasets. For these experiments, we consider one class as normal and rest 9 classes as anomalies following the setup from [Ruff et al., 2018, Zenati et al., 2018].

Baselines: We compare our model against various generative model based anomaly detection approaches. Ren et al. [2019] uses likelihood based estimate from a Autoregressive model to discover anomalies. WAIC [Choi et al., 2018] proposes to use WAIC criteria on top of likelihood estimation methods to find anomalies. Serrà et al. [2019] leverages complexity estimate of images to detect OOD inputs. Typicality test [Nalisnick et al., 2019] proposes to calculate an empirical estimate of entropy of set of samples and use it to

Table 1: In this table we present AUROC scores on various OOD detection across various datasets. All methods in the table have no access to OOD data during training, but a small number of anomalies during validation to choose the best model. All the results are average AUROC values across test dataset, with one sample evaluated at a time except the result for Typicality test[Nalisnick et al., 2019] which corresponds to using batch size of 2 of the same type. In the bottom half of the table we show the ablation results of our model AMA with one component missing at a time from our pipeline.

Trained on:	FashionMNIST		CIFAR-10			SVHN		
OOD data:	MNIST	Omniglot	SVHN	Imagenet	CIFAR-100	CIFAR-10	Imagenet	CIFAR-100
WAIC on WGAN ensemble [Choi et al., 2018]	0.871	0.832	0.623	0.626	-	-	-	-
Likelihood-ratio on PixelCNN++ [Ren et al., 2019]	**0.994**	-	0.931	-	-	-	-	-
Typicality test on Glow model [Nalisnick et al., 2019]	0.140	-	0.420	0.640	-	0.980	**1.000**	-
DeepSVDD [Ruff et al., 2018]	0.864	0.999	0.533	0.387	0.478	0.795	0.823	0.819
S using PixelCNN++ and FLIF [Serrà et al., 2019]	0.967	**1.000**	0.929	0.589	0.535	-	-	-
AMA w/o Mirrored Wass. Loss (Ours)	0.653	0.899	0.800	0.526	0.510	0.503	0.693	0.592
AMA w/o Simplex Interpolation (Ours)	0.960	0.998	0.820	0.847	0.537	0.991	0.993	0.987
AMA w/o Atypical selection (Ours)	0.894	0.997	0.861	0.812	0.535	0.990	0.991	0.987
AMA w/o new anomaly scoring (Ours)	0.991	0.997	0.442	0.890	0.501	**0.993**	**1.000**	**0.988**
AMA (Ours)	0.987	0.998	**0.958**	**0.911**	**0.551**	**0.993**	**1.000**	**0.988**

recognize anomalies. DeepSVDD [Ruff et al., 2018] optimizes the latent representations of images their distances in latent space as complexity measure.

In addition to these, another set of methods [Akçay et al., 2019, Zenati et al., 2018, Schlegl et al., 2017, Ngo et al., 2019, Ruff et al., 2018] addresses the setting in which anomalies come from the same data manifold (i.e. same dataset). We compared our model to these methods in this setting as well and we believe these methods can be extended to the case of OOD samples coming from different data manifold. For these experiments we follow the setup from [Zenati et al., 2018, Ruff et al., 2018, Schlegl et al., 2017], where one class is considered normal and the rest of the classes from the same dataset as anomalies. All the results shown in Table 2 are for this setting. In DeepSVDD, Global Contrast Normalization is used on the data prior to the training. We removed this additional normalization step to make the method comparable to other baselines.

Note that discriminative models such as [Hendrycks et al., 2018, Vyas et al., 2018, Hsu et al., 2020] achieve higher performance in OOD detection benchmarks, but assume access to the class labels during training. For brevity, we consider only unsupervised baselines.

BatchNorm Issue: We noticed that one of the earlier work [Akçay et al., 2019][1] evaluated their model in the training mode instead of setting to the evaluation mode. Due to this issue, the BatchNorm is calculated for the test batch, rather than using the train-time statistics.

Hence, while reporting results for Akçay et al. [2019], we re-evaluate their models by freezing the BatchNorm statistics during the test time. We follow the same protocol for all the models.

Network Architectures and Training: The generator and the discriminator in our model have residual architectures and are borrowed from SN-GAN [Miyato et al., 2018]. Our Encoder is a 4 layered convolution network with BatchNorm and LeakyRelu nonlinearity. Please refer to the appendix for the complete model architecture and training details.

Following the setting in [Zenati et al., 2018, Ren et al., 2019] we assume that we have access to a small number of anomalies during validation time (≈ 50 in number). To generate the test set, we randomly sample anomalies from the an OOD dataset, 20% the size of normal samples, compared to sampling equal number of normal and anomalies scenario presented in [Ren et al., 2019, Choi et al., 2018]. We believe our scenario is far more realistic and more stringent.We keep the test data and normalizations same during training for our model as well as the baselines to make them comparable.

4.2 ANOMALY DETECTION PERFORMANCE

We consider two common scenarios used in literature to benchmark the performance of our model. In the first scenario, we consider images from a given dataset as the normal samples and images from a different dataset (typically with a different underlying distribution) as anomalies (Eg. CIFAR-10 vs SVHN). In the second scenario, we consider images from one of the categories in the dataset as normal images while all other as anomalies (Eg. digit 0 vs rest in MNIST). Note that, in some papers, these two scenarios are referred as as out-of-distribution (OOD) and in-distribution anomalies respectively. Even though they are treated as different problems in previous work, they share the common goal of flagging samples that are different from input distribution. Hence we do not make this distinction and use the term "anomalies" to refer to the either scenario. We show that in both the scenarios, our model outperforms or matches the performance of current generative based anomaly detectors.

Images from different dataset as anomalies In Table 1, we show the performance of our model and the baselines against 3 different cases. Our first set of experiments uses gray-scale images from Fashion MNIST as normal images

[1] https://github.com/samet-akcay/skip-ganomaly

Table 2: Here we show performance of anomaly detection task when anomalies come from an unseen class of the same dataset. Each column denotes the normal class and the rest 9 classes from that respective dataset are considered as anomalies. The performance is measured using AUROC scores, higher the better.

MNIST	0	1	2	3	4	5	6	7	8	9	Average
FGAN[Ngo et al., 2019]	0.754	0.307	0.628	0.566	0.390	0.490	0.538	0.313	0.645	0.408	0.504
ALAD[Zenati et al., 2018]	0.962	0.915	0.794	0.821	0.702	0.79	0.843	0.865	0.771	0.821	0.828
Ano-GAN[Schlegl et al., 2017]	0.902	0.869	0.623	0.785	0.827	0.362	0.758	0.789	0.672	0.720	0.731
Skip-Ganomaly[Akçay et al., 2019]	0.297	0.877	0.393	0.486	0.618	0.540	0.455	0.633	0.426	0.584	0.531
DeepSVDD[Ruff et al., 2018]	0.971	0.995	0.809	0.884	**0.920**	0.869	0.978	0.940	**0.900**	0.946	0.921
AMA (Ours)	**0.986**	**0.998**	**0.882**	**0.891**	0.894	**0.938**	**0.981**	**0.983**	0.876	**0.948**	**0.938**
CIFAR-10	airplane	automobile	bird	cat	deer	dog	frog	horse	ship	truck	Average
FGAN[Ngo et al., 2019]	0.572	0.582	0.505	0.544	0.534	0.535	0.528	0.537	0.664	0.338	0.567
ALAD[Zenati et al., 2018]	0.679	0.397	0.685	**0.652**	0.696	0.550	0.704	0.463	**0.787**	0.391	0.601
Ano-GAN[Schlegl et al., 2017]	0.602	0.439	0.637	0.594	**0.755**	0.604	**0.730**	0.498	0.675	0.445	0.598
Skip-Ganomaly[Akçay et al., 2019]	0.655	0.406	0.663	0.598	0.739	0.617	0.638	0.519	0.746	0.387	0.597
Deep SVDD[Ruff et al., 2018]	0.682	0.477	0.679	0.573	0.752	0.628	0.710	0.511	0.733	0.567	0.631
AMA (Ours)	**0.752**	**0.634**	**0.696**	0.603	0.733	**0.650**	0.658	**0.582**	0.754	**0.632**	**0.669**

while the images from MNIST and Omniglot as OOD images. This is a relatively simple experiment and nearly all the baselines and our model achieve almost perfect AUROC. Even though our model does not have the best AUROC, it is well within the margin of error of the best performing the model.

Next two cases are a bit more challenging as the images are colored and more diverse. In first case, we use normal samples from CIFAR-10, and anomalies from SVHN, Imagenet, and CIFAR-100. In the second case, we use normal samples from SVHN, while the anomalies coming from CIFAR-10, Imagenet and CIFAR-100. Our model outperforms all the baselines in both these experiments. This shows that our model, AMA is optimizing the latent space of normal samples well which leads to an impressive generalization behavior. Even though AUROC scores are greater than 0.9 in most of the cases, our model falls short in case of CIFAR-10 vs CIFAR-100 (similar behavior is observed for the other baselines as well). This is a really hard scenario and even humans will have tough time deciding whether a given image is from CIFAR-10 or CIFAR-100.

Images from different categories as anomalies In Table-2, we show Anomaly Detection experiments when anomalies arise from the same data manifold (i.e. same dataset). Each column shows the results of a normal class with the rest of 9 classes as anomalies. Our method (AMA) outperforms other methods in terms of average scores with 1.7% AUROC gain over the next best method on MNIST and 3.7% gain on CIFAR-10 dataset. In terms of an individual case comparison, we achieve best results in 8 out of 10 cases on MNIST, while 6 out of 10 cases on CIFAR-10.

4.3 ABLATION STUDIES

We introduced 3 main ideas in this paper: Mirrored Wasserstein loss, Latent space regularization using Simplex Interpolation and Atypical Selection, and an alternative Anomaly scoring technique. In the second half of the Table 1, we show the ablation results, removing one component at a time. As expected, removing Mirrored Wasserstein loss reduces the AUROC scores the most. AUROC scores are reduced by an order of ~ 0.1 points whenever a part of Latent space regularization is removed. We see that in most of the cases, removing Atypical Selection reduces the scores a bit more than removing Simplex Interpolation. The new anomaly scoring metric contributes the most when the normal sample distribution is more diverse than the OOD distribution, eg: the case of CIFAR-10 as normal and SVHN as OOD. When we used R-score to identify anomalies in this scenario, most of the SVHN samples are tagged normal while most of the CIFAR-10 images tagged as anomalies, thus resulting in lower AUROC.

When does Atypical Selection help? In Table 1, we see how the performance of the model gets impacted when we removed Atypical Selection component. In most of the cases, the drop in the performance is fairly low. When we examine the R-score distributions between the pairs of data sets, we notice a trend that the Atypical selection helps the most when both the distributions are highly overlap.

5 CONCLUSION

In this paper, we introduced a new method for the unsupervised anomaly detection problem, Adversarial Mirrored Autoencoder (AMA), equipped with Mirrored Wasserstein loss and a latent space regularizer. Our method outperforms existing generative model based anomaly detectors on several benchmark tasks. We also show how each of the components in our proposed approach contribute to the model's performance in diverse data settings. While our model is quite powerful in OOD detection, it still does not fare well in harder experiments such as CIFAR-10 (as in-distribution) vs CIFAR-100 (as OOD). This is rather similar to the setting of anomalies arising from the same data manifold. While we

showed some early results in Table 2, we can further extend this work to improve for such scenarios. In the present work, in OOD case, we only consider anomalies from a single dataset, hence while doing Atypical Selection, we choose the pseudo-negatives which are closer to or further away from the origin in latent space as compared to in-distribution points. However this may not work when anomalies are coming from wide range of distributions, which we hope to address in future work.

ACKNOWLEDGEMENTS

This project was supported in part by NSF CAREER AWARD 1942230, an IBM faculty award, a grant from Capital One, and a Simons Fellowship on Deep Learning Foundations. This work was supported through the IBM Global University Program Awards initiative. Authors thank Ritesh Soni, Steven Loscalzo, Bayan Bruss, Samuel Sharpe, Jason Wittenbach and Kamal Gupta for helpful discussions.

References

Samet Akçay, Amir Atapour-Abarghouei, and Toby P Breckon. Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection. In *IJCNN*, pages 1–8. IEEE, 2019.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

David Berthelot, Colin Raffel, Aurko Roy, and Ian Goodfellow. Understanding and improving interpolation in autoencoders via an adversarial regularizer. *arXiv preprint arXiv:1807.07543*, 2018.

Avrim Blum, John Hopcroft, and Ravindran Kannan. Foundations of data science. *Vorabversion eines Lehrbuchs*, 5, 2016.

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

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

Hyunsun Choi, Eric Jang, and Alexander A Alemi. Waic, but why? generative ensembles for robust anomaly detection. *arXiv preprint arXiv:1810.01392*, 2018.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Dipankar Dasgupta and Nivedita Sumi Majumdar. Anomaly detection in multidimensional data using negative selection algorithm. In *Congress on Evolutionary Computation.*, volume 2, pages 1039–1044. IEEE, 2002.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009.

Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

Stephanie Forrest, Alan S Perelson, Lawrence Allen, and Rajesh Cherukuri. Self-nonself discrimination in a computer. In *Proceedings of 1994 IEEE computer society symposium on research in security and privacy*, pages 202–212. Ieee, 1994.

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*, pages 2672–2680, 2014.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

Victoria Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial intelligence review*, 22(2):85–126, 2004.

Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *CVPR*, 2020.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE CVPR*, pages 1125–1134, 2017.

Chen Jinyin and Yang Dongyong. A study of detector generation algorithms based on artificial immune in intrusion detection system. In *2011 3rd International Conference on Computer Research and Development*, volume 1, pages 4–8. IEEE, 2011.

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

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in neural information processing systems*, pages 10215–10224, 2018.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. The omniglot challenge: a 3-year progress report. *Current Opinion in Behavioral Sciences*, 29:97–104, 2019.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, pages 7167–7177, 2018.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15:2018, 2018.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

SC Martin Arjovsky and Leon Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Prateek Munjal, Akanksha Paul, and Narayanan C Krishnan. Implicit discriminator in variational autoencoder. In *IJCNN*, pages 1–8. IEEE, 2020.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*, 2018.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using typicality. *arXiv preprint arXiv:1906.02994*, 2019.

Cuong Ngo, Amadeus Aristo Winarto, Connie Kou Khor Li, Sojeong Park, Farhan Akram, and Hwee Kuan Lee. Fence gan: Towards better anomaly detection. *arXiv e-prints*, pages arXiv–1904, 2019.

A Nguyen, J Yosinski, and J Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. arxiv, cs, 2014.

Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016.

Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *NeurIPS*, pages 14707–14718, 2019.

Mihaela Rosca, Balaji Lakshminarayanan, David Warde-Farley, and Shakir Mohamed. Variational approaches for auto-encoding generative adversarial networks. *arXiv preprint arXiv:1706.04987*, 2017.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *ICML*, pages 4393–4402, 2018.

Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *arXiv preprint arXiv:2009.11732*, 2020.

Tim Sainburg, Marvin Thielk, Brad Theilman, Benjamin Migliori, and Timothy Gentner. Generative adversarial interpolative autoencoding: adversarial training on latent space interpolations encourage convex latent distributions. *arXiv preprint arXiv:1807.06650*, 2018.

Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*, 2017.

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

Pierre Sermanet, Soumith Chintala, and Yann LeCun. Convolutional neural networks applied to house numbers digit classification. In *ICPR*, pages 3288–3291. IEEE, 2012.

Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. *arXiv preprint arXiv:1909.11480*, 2019.

John Sipple. Interpretable, multidimensional, multimodal anomaly detection with negative sampling for detection of device failure. *arXiv preprint arXiv:2007.10088*, 2020.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

David MJ Tax and Robert PW Duin. Support vector data description. *Machine learning*, 54(1):45–66, 2004.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Apoorv Vyas, Nataraj Jammalamadaka, Xia Zhu, Dipankar Das, Bharat Kaul, and Theodore L Willke. Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In *ECCV*, 2018.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv:1708.07747*, 2017.

Houssam Zenati, Manon Romain, Chuan-Sheng Foo, Bruno Lecouat, and Vijay Chandrasekhar. Adversarially learned anomaly detection. In *2018 ICDM*. IEEE, 2018.