Deep Variational Semi-Supervised Novelty Detection

Tal Daniel
ECE Department, Technion
Haifa, Israel
taldanielm@campus.technion.ac.il

Thanard Kurutach
EECS Department, UC Berkeley
California, USA

Aviv Tamar
ECE Department, Technion
Haifa, Israel

Abstract

In anomaly detection (AD), one seeks to identify whether a test sample is abnormal, given a data set of normal samples. A recent and promising approach to AD relies on deep generative models, such as variational autoencoders (VAEs), for unsupervised learning of the normal data distribution. In semi-supervised AD (SSAD), the data also includes a small sample of labeled anomalies. In this work, we propose two variational methods for training VAEs for SSAD. The intuitive idea in both methods is to train the encoder to 'separate' between latent vectors for normal and outlier data. We show that this idea can be derived from principled probabilistic formulations of the problem, and propose simple and effective algorithms. Our methods can be applied to various data types, as we demonstrate on SSAD datasets ranging from natural images to astronomy and medicine, can be combined with any VAE model architecture, and are naturally compatible with ensembling. When comparing to state-of-the-art SSAD methods that are not specific to particular data types, we obtain marked improvement in outlier detection.

1 Introduction

Anomaly detection (AD) – the task of identifying abnormal samples with respect to some normal data – has applications in domains ranging from health-care, to security, and robotics [35]. In its common formulation, training data is provided only for normal samples, while at test time, anomalous samples need to be detected. In the probabilistic AD approach, a model of the normal data distribution is learned, and the likelihood of a test sample under this model is thresholded for classification as normal or not. Recently, deep generative models such as variational autoencoders (VAEs, [23]) and generative adversarial networks [17] have shown promise for learning data distributions in AD [1, 44, 38, 48, 50].

Here, we consider the setting of *semi-supervised* AD (SSAD), where in addition to the normal samples, a small sample of labeled anomalies is provided [18]. Most importantly, *this set is too small to represent the range of possible anomalies*. For example, consider normal data representing the digit '0' in the MNIST dataset, and being given a small sample from the digit '1'as anomalies. The goal in SSAD is to use this data to better classify as anomalies *all the digits* '1'-'9'. In this case, it is clear that classification methods (either supervised or semi-supervised) are not suitable. Instead, most approaches are based on 'fixing' an unsupervised AD method to correctly classify the labeled anomalies, while still maintaining AD capabilities for unseen outliers [18, 32, 37].

NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications

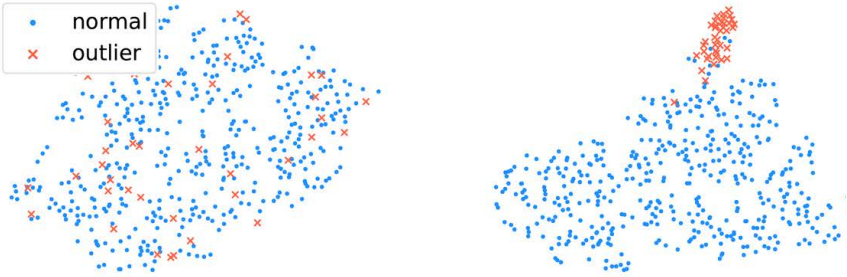

Figure 1: t-SNE of the latent space of a conventional VAE (left), and our proposed dual-prior VAE (right), trained on Thyroid dataset, and evaluated on an unseen test sample. Latent vectors for outliers (red crosses) in dual-prior VAE are pushed away from normal samples.

We present a variational approach for learning data distributions in the SSAD problem setting. We base our method on the VAE, and modify the training objective to account for the labeled outlier data. We propose two formulations for this problem. The first maximizes the log-likelihood of normal samples, while minimizing the log-likelihood of outliers, and we effectively optimize this objective by combining the standard evidence lower bound (ELBO) with the χ upper bound (CUBO, [13]). The second method is based on separating the VAE prior between normal and outlier samples. Effectively, both methods have a similar intuitive interpretation: they modify the VAE encoder to push outlier samples away from the prior distribution (Fig. 1).

The state-of-the-art Deep-SAD SSAD method [37] also separates the encoding of normal and outlier samples. In comparison, by building on variational inference theory, our approach does not place any restriction on the network architecture as Ruff et al. [37], and can be used to modify any VAE to account for outliers. Additionally, our derivation allows a principled use of ensembles, which we show to improve performance significantly.

We evaluate our methods in the comprehensive SSAD test-suite of Ruff et al. [37], which includes both image data and low-dimensional data sets from astronomy, medicine, and other domains. Our method is on par with or better than the state-of-the-art. For some image data sets, Nalisnick et al. [33] argued that VAEs are not informative novelty detectors. However, as demonstrated by Daniel and Tamar [9], incorporating a likelihood-based adversarial term in the objective of VAE results in excellent out-of-distribution detection in the same setting as Nalisnick et al. [33]. As we show here, even 1% of labeled anomalies can significantly improve detection in the setting of Ruff et al. [37].

Additionally, to the best of our knowledge, our approaches are the first principled probabilistic formulation for training VAEs with negative examples. We demonstrate this idea, and further establish the flexibility of our method, by modifying a conditional VAE used for generating sampling distributions for robotic motion planning to not generate way points that collide with obstacles [20], and also, on learning to identify novel sentiments in text, based on a VAE for sentences [4].

2 Related Work

Anomaly detection, a.k.a. outlier detection or novelty detection, is an active field of research. Shallow methods include one-class SVM (OC-SVM, [39]) and support vector data description (SVDD, [47]) and rely on hand-crafted features. Recently, there is interest in deep learning methods for AD [6]. Erfani et al. [14], Andrews et al. [2], Cao et al. [5], Chen et al. [7] follow a hybrid approach where deep unsupervised learning is used to learn features, which are then used within a shallow AD method. Specifically for image domains, Golan and El-Yaniv [16] learn features using a self-supervised paradigm – by applying geometric transformations to the image and learning to classify which transformation was applied. Recently, self-supervised and contrastive approaches for image-based AD [45, 40] have exhibited impressive results by using augmentations or features from pre-trained models. Lee et al. [27] learn a distance metric in feature space, for networks pre-trained on image classification. All of these methods display outstanding AD performance, but are limited to image domains, while our approach does not require particular data or pre-trained features. Extending

our approach to use unlabelled data, image augmentations, or pre-trained features is an interesting direction for future work. Several studies explored using deep generative models such as GANs and VAEs for AD [1, 44, 38, 48], and more recently, deep energy-based models [49, 43] have reported outstanding results. Our work extends the VAE approach to the SSAD setting, showing that even a 1% fraction of labelled anomalies can improve over the state-of-the-art AD scores of deep energy based models, demonstrating the importance of the SSAD setting.

Most studies on SSAD also require hand designed features. Muñoz-Marí et al. [32] proposed S^2OC-SVM, a modification of OC-SVM that introduces labeled and unlabeled samples, Görnitz et al. [18] proposed an approach based on SVDD, and Blanchard et al. [3] base their approach on statistical testing. Also related is the active learning approach to AD [10, 11, 19], which sequentially improves an anomaly detector by querying an expert for labels. Deep SSAD has been studied recently in specific contexts such as videos [25], network intrusion detection [29], or specific neural network architectures [15].

The most relevant prior work to our study is the recently proposed deep semi-supervised anomaly detection (Deep SAD) approach of Ruff et al. [37]. Deep SAD is a general method based on deep SVDD [36], which learns a neural-network mapping of the input that minimizes the volume of data around a predetermined point. While it has been shown to work on a variety of domains, Deep SAD must place restrictions on the network architecture, such as no bias terms (in all layers) and no bounded activation functions, to prevent degeneration of the minimization problem to a trivial solution. The methods we propose here do not place any restriction on the network architecture, and can be combined with any VAE model. Further, we show that the main strength of Deep SAD stems from its auto-encoder pre-training, making it effectively similar to a VAE-based approach. In this respect, our work provides a principled theoretical development of this approach. In addition, we show improved performance in almost all domains compared to Deep SAD's state-of-the-art results.

3 Background

We consider deep unsupervised learning under the variational inference setting [23]. Given some data x, one aims to fit the parameters θ of a latent variable model $p_\theta(x) = \mathbb{E}_{p(z)}\left[p_\theta(x|z)\right]$, where the prior $p(z)$ is known. For general models, the maximum-likelihood objective $\max_\theta \log p_\theta(x)$ is intractable due to the marginalization over z, and can be approximated using variational inference methods.

Evidence lower bound (ELBO) The evidence lower bound (ELBO) states that for some approximate posterior distribution $q(z|x)$:

$$\log p_\theta(x) \geq \mathbb{E}_{q(z|x)}\ \left[\log p_\theta(x|z)\right] - D_{\mathrm{KL}}(q(z|x)\|p(z)),$$

where the Kullback-Leibler divergence is $D_{\mathrm{KL}}(q(z|x)\|p(z)) = \mathbb{E}_{q(z)}\left[\log \frac{q(z|x)}{p(z)}\right]$. In the variational autoencoder (VAE, [23]), the approximate posterior is represented as $q_\phi(z|x) = \mathcal{N}(\mu_\phi(x), \Sigma_\phi(x))$ for some neural network with parameters ϕ, the prior is $p(z) = \mathcal{N}(\mu_0, \Sigma_0)$, and the ELBO can be maximized using the *reparameterization trick*. Since the resulting model resembles an autoencoder, the approximate posterior $q_\phi(z|x)$ is also known as the *encoder*, while $p_\theta(x|z)$ is termed the *decoder*.

χ upper bound (CUBO) Recently, Dieng et al. [13] derived variational upper bounds to the data log-likelihood. The χ^n-divergence is given by $D_{\chi^n}(p\|q) = \mathbb{E}_{q(z;\theta)}\left[\left(\frac{p(z|X)}{q(z;\theta)}\right)^n - 1\right]$. For $n > 1$, Dieng et al. [13] propose the following χ upper bound (CUBO):

$$\log p_\theta(x) \leq \frac{1}{n}\log \mathbf{E}_{q(z|x)}\left[\left(\frac{p_\theta(x|z)p(z)}{q(z|x)}\right)^n\right] \doteq CUBO_n(q(z|x)).$$

For a family of approximate posteriors $q_\phi(z|x)$, one can minimize the CUBO using Monte Carlo (MC) estimation. However, MC gives a lower bound to CUBO and its gradients are biased. As an alternative, Dieng et al. [13] proposed the following optimization objective: $\mathcal{L} = \exp\{n \cdot CUBO_n(q_\phi(z|x))\}$. By monotonicity of $\exp$, this objective reaches the same optima as $CUBO_n(q_\phi(z|x))$. This produces an unbiased estimate, and the number of samples only affects the variance of the gradients.

In the following, we denote by $ELBO_Q(X)$ and $CUBO_Q(X)$ the ELBO and CUBO for data X where the approximate posterior is $Q(z|x)$.

4 Deep Variational Semi-Supervised Anomaly Detection

In the anomaly detection problem, the goal is to detect whether a sample x was generated from some normal distribution $p_{normal}(x)$,[1] or not – making it an anomaly. In semi-supervised anomaly detection (SSAD), we are given N_{normal} samples from $p_{normal}(X)$, which we denote X_{normal}.[2] In addition, we are given $N_{outlier}$ examples of anomalous data, denoted by $X_{outlier}$, and we assume that $N_{outlier} \ll N_{normal}$. In particular, $X_{outlier}$ does not cover the range of possible anomalies, and thus classification methods (neither supervised nor semi-supervised) are not applicable.

Our approach for SSAD is to approximate $p_{normal}(x)$ using a deep latent variable model $p_\theta(x)$, and to decide whether a sample is anomalous or not based on thresholding its predicted likelihood. In the following, we propose two variational methods for learning $p_\theta(x)$. The first method, which we term *max-min likelihood* VAE (MML-VAE), maximizes the likelihood of normal samples while minimizing the likelihood of outliers. The second method, which we term *dual-prior* VAE (DP-VAE), assumes different priors for normal and outlier samples.

4.1 Max-Min Likelihood VAE

In this approach, we seek to find model parameters based on the following objective:

$$\max_{\theta} \quad \log p_\theta(X_{normal}) - \gamma \log p_\theta(X_{outlier}), \tag{1}$$

where $\gamma \geq 0$ is a weighting term. Note that, in the absence of outlier data or when $\gamma = 0$, Equation 1 is just maximum likelihood estimation. For $\gamma > 0$, however, we take into account the knowledge that outliers should not be assigned high probability.

We model the data distribution using a latent variable model $p_\theta(x) = \mathbb{E}_{p(z)}\left[p_\theta(x|z)\right]$, where the prior $p(z)$ is known. To optimize the objective in Equation 1 effectively, we propose the following variational lower bound:

$$\begin{aligned} &\log p_\theta(X_{normal}) - \gamma \log p_\theta(X_{outlier}) \\ &\geq ELBO_{Q_1}(X_{normal}) - \gamma CUBO_{Q_2}(X_{outlier}), \end{aligned} \tag{2}$$

where $Q_1(z|x), Q_2(z|x)$ are the variational auxiliary distributions. In principle, the objective in Equation 2 can be optimized using the methods of Kingma and Welling [23] and Dieng et al. [13], which would effectively require training two encoders: Q_1 and Q_2, and one decoder $p_\theta(x|z)$,[3] separately on the two datasets. However, it is well-known that training deep generative models requires abundant data. Thus, there is little hope for learning an informative Q_2 using the small dataset $X_{outlier}$. To account for this, our main idea is to use the same variational distribution $Q(z|x)$ for both loss terms, which effectively relaxes the lower bound as follows:

$$\begin{aligned} &\log p_\theta(X_{normal}) - \gamma \log p_\theta(X_{outlier}) \\ &\geq ELBO_{Q}(X_{normal}) - \gamma CUBO_{Q}(X_{outlier}). \end{aligned} \tag{3}$$

In other words, we use the **same encoder** for both normal and anomalous data. One can look at this design choice from a bias-variance perspective: sharing the encoder may loosen the lower bound (similar to adding bias), but in practice will work better when we have only a few negative samples (reducing variance). Finally, the loss function of the MML-VAE is: $\mathcal{L} = \gamma CUBO_Q(X_{outlier}) - ELBO_Q(X_{normal})$.

To gain intuition about the effect of the CUBO term in the loss function, it is instructive to assume that $p_\theta(x|z)$ is fixed. For $CUBO_2(q(z|x))$, the objective can be written as: $\mathbb{E}_q\left[\exp\left(-2\log\left(\frac{q(z|x)}{p(x|z)p(z)}\right)\right)\right]$. Minimizing this objective leads to pushing $q(z|x)$ away from $p(z)$. In this case, maximizing the lower bound only affects the variational distribution $q(z|x)$. Thus, intuitively, the CUBO term seeks to **separate** the latent distribution for normal samples, which will concentrate on high-likelihood regions of $p(z)$, from the latents of outliers, which will concentrate elsewhere. To further clarify, we plot the CUBO loss for a 1-dimensional Gaussian $q(z|x) \sim \mathcal{N}(\mu, \sigma)$,

[1] Throughout, subscript $\cdot_{normal}$ refers to the distribution of normal data, and not to the Gaussian distribution.

[2] For clarity, our exposition assumes that X_{normal} is clean, and does not contain any anomalous samples. We do verify that our methods can handle the important case of data polluted with anomalies in our experiments.

[3] Based on the ELBO and CUBO definitions, the decoder $p_\theta(x|z)$ is the same for both terms.

where the prior is $\mathcal{N}(0,1)$, and we fix $\mathcal{L}_R$. Observe that the CUBO loss drives $q(z|x)$ away from the prior to areas with $\mu > 0$ and small variance.

For the common choice of Gaussian prior and approximate posterior distributions [23], the CUBO component of the loss function is given by: $\mathcal{L}_{CUBO} = \exp\{\log|\Sigma_q| + \mu_q^T\Sigma_q^{-1}\mu_q + \log\mathbb{E}_q\left[\exp\{-2\cdot\mathcal{L}_R - z^Tz + z^T\Sigma_q^{-1}z - 2z^T\Sigma_q^{-1}\mu_q\}\right]\}$, where $\mathcal{L}_R = -\log p_\theta(x|z)$ is the reconstruction error. The expectation can be approximated with MC estimation. In our experiments, we found that computing $\exp(\log\mathbb{E}\left[\exp(\cdot)\right])$ is numerically more stable, as we can employ the log-sum-exp trick.

In practice, we found that updating only the encoder according to the CUBO loss results in better performance and stabler training when the CUBO is estimated with a small number of samples, corresponding with the bias-variance intuition above. The reason for this is that CUBO seeks to maximize the reconstruction error, which does not produce informative updates for the decoder. Since this decoder is shared with the ELBO term, the CUBO update can decrease performance for normal samples as well. Complete algorithm details and derivation of the CUBO objective are provided in Appendix E.

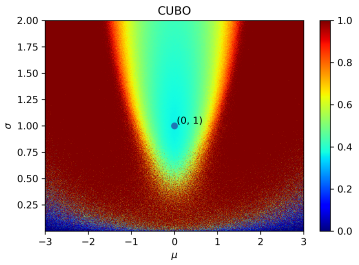

Figure 2: CUBO loss for 1D Gaussian.

4.2 Dual Prior VAE

The second method we propose is a simple yet effective modification of the VAE for SSAD, which we term dual-prior VAE (DP-VAE). Here, we assume that both normal and outlier data are generated from a single latent variable model $p_\theta(x) = \mathbb{E}_{p(z,y)}\left[p_\theta(x|z,y)\right]$, where y is an additional Bernoulli variable that specifies whether the sample is normal or not. Additionally, we assume the following conditional prior for z:

$$p(z|y) = \begin{cases} \mathcal{N}(\mu_{normal}, \Sigma_{normal}), & y = 1 \\ \mathcal{N}(\mu_{outlier}, \Sigma_{outlier}), & y = 0 \end{cases} .$$

Since our data is labeled, y is known for every x and does not require inference. Deriving the ELBO for normal and outlier samples in this case gives: $ELBO_{normal}(x) = -\mathcal{L}_R(x) - D_{\mathrm{KL}}[Q(z|x)\|p(z|y=1)]$, $ELBO_{outlier}(x) = -\mathcal{L}_R(x) - D_{\mathrm{KL}}[Q(z|x)\|p(z|y=0)]$. Here, similarly to the MML-VAE above, we assume *the same encoder* $Q(z|X)$ for normal and outlier samples, and the loss function is

$$\mathcal{L} = -\left[ELBO_{normal}(X_{normal}) + ELBO_{outlier}(X_{outlier})\right].$$

In our implementation, we set $\Sigma_{normal} = \Sigma_{outlier} = I$, $\mu_{normal} = 0$, and $\mu_{outlier} = \alpha\vec{1}$ where $\alpha \neq 0$. In this case, this loss function has the intuitive interpretation of training an encoder that separates between latent variables for normal and outlier data.

As $\mathcal{L}$ is simply a combination of VAE loss functions, optimization is straightforward. Similarly to the MML-VAE method, we freeze the decoder when minimizing $ELBO_{outlier}$, as we don't expect for an informative reconstruction of the outlier distribution from the limited data $X_{outlier}$.

We emphasize that while this model assumes a prior distribution for the outliers (e.g., following the example in the introduction, for MNIST digits of class '1'), it can still be used in the SSAD setting to generalize to unseen outliers (digits from '1' to '9'). The reason, as we elaborate on in the next section, is that we will only use the normal data prior (for digit '0') for anomaly detection. Thus, similarly to the MML-VAE model, the DP-VAE works only to **separate** the latents of observed outliers from the latents of the normal data, therefore improving the detection of normal samples.

4.3 VAE Architectures and Ensembles

Both MML-VAE and DP-VAE simply add loss terms to the objective function of the conventional VAE. Thus, they can be applied to any type of VAE without any restriction on the architecture. In addition, they can be used with ensembles of VAEs, a popular technique to robustify VAE training and increase its performance in classification and generations tasks [31, 46, 42, 7, 30]. We train K different VAEs (either MML-VAE or DP-VAE) with random initial weights, and set the score to be the average over the ELBOs in the ensemble. We found that ensembles significantly improve the SSAD performance of our methods, as demonstrated in our experiments. In section 5.3 we further report an ablative analysis to evaluate the effect of using an ensemble.

4.4 Anomaly Detection with MML-VAE and DP-VAE

Finally, we describe our SSAD method based on MML-VAE and DP-VAE. After training an MML-VAE, the $ELBO$ provides an approximation of the normal data log-likelihood. Thus, given a novel sample x, we can score its (log) likelihood of belonging to normal data as:

$$score(x) = ELBO(x).$$

After training a DP-VAE, $ELBO_{normal}$ provides a similar approximation, and in this case we set

$$score(x) = ELBO_{normal}(x).$$

Finally, for classifying the sample, we threshold its score.

4.5 Comparison with SVDD and Deep SAD

We discuss connections between our approaches and Deep SAD [37], the recent state-of-the-art that builds on deep SVDD. In SVDD [36], a deep network is trained to minimize the average distance in latent space between the features of the normal data samples and a predetermined point c, pushing the latent representations of normal data to concentrate in a hyper-sphere centered on c. The distance from c is then used to distinguish normal samples (small distance) from outliers (large distance). While in principle c could be arbitrary, in practice, the network is pre-trained using an auto-encoder loss, and c is chosen by averaging the features of normal data. As we show in Appendix G, choosing c using the auto-encoder pre-training is crucial for SVDD performance (and thus, also for Deep SAD). For SSAD, Ruff et al. [37] modify the SVDD objective to further push outliers away from c. While Ruff et al. [37] motivate this training using information theoretic terms, many steps in the explanation are not rigorously justified, such as the minimization of mutual information using the regularized auto-encoder, and using the inverse distance from c when mapping features for outlier data.

Interestingly, our methods, which are developed from a principled variational inference formulation, have a similar interpretation of learning features that minimize a (variational) auto-encoder loss, and 'pushing away' the features of outliers from the features of normal samples. However, our auto-encoder loss is well-established as $p_\theta(x|z)$ in the ELBO. Further, our loss for outliers is developed rigorously, as a sequence of well-defined approximations to the data likelihood, in both methods.

Furthermore, by building on well-established first principles, our method is not prone to the phenomena termed 'hypersphere collapse', where the network can reduce the latent distance from c to zero for all inputs. In order to avoid it, Ruff et al. [36, 37] place strong restrictions on the model architecture (no bias terms, unbounded activations), or require modifications to the data [8]. For the semi-supervised setting, Ruff et al. [37] notes that if there are sufficiently many labeled anomalies, hypersphere collapse is not a problem for Deep SAD. However, this is an empirical observation, and there is no result showing how many labeled anomalies are required to prevent hypersphere collapse. In practice, the restrictions of Deep SVDD were applied for obtaining the results in Ruff et al. [37].

We conclude with a note on ensembles. As we show in Section 5, ensembles reduce the variance of the stochastic VAE training, improve our method's performance, and lead to state-of-the-art results. For a fair comparison with Deep SAD, however, we have also evaluated ensembles of Deep SAD models, as reported in Appendix A.1. Interestingly, our investigation shows that ensembles have little effect on Deep SAD. This can be explained as follows. In Deep-SAD, confidence is measured according to distance to an arbitrary point c in the vector space. Thus, the scores from different networks in the ensemble are not necessarily calibrated (as they each have different c points). In our VAE approach, however, the score of each network is derived to approximate the likelihood of the sample – a calibrated quantity, naturally giving rise to the benefit of the ensemble.

5 Experiments

Our experiments aim to (1) establish our methods as SOTA for deep SSAD, following the extensive evaluation protocol of Ruff et al. [37]; and (2) demonstrate the generality of our approach for training VAEs with negative samples on various domains. The tasks and datasets are summarized in Table 1, and note that due to lack of space, some results are reported in the appendix. In particular, our experiments on detecting novel sentiments in text data using a recurrent neural network (RNN) based VAE are in Appendix H.

Domain	Task	Datasets	Architecture
Images	Anomaly Detection	MNIST, Fashion-MNIST, CIFAR10, CatsVsDogs	CNN
Classic AD Benchmark [41]	Anomaly Detection	arrhythmia, cardio, satellite, satimage-2, shuttle, thyroid	MLP
Robotics	Motion Planning	Path Planning [20]	MLP
NLP (Appendix H)	Sentiment Analysis	IMDB	RNN (LSTM)

Table 1: Summary of the experimented tasks and data

SSAD Evaluation: The SSAD evaluation protocol of Ruff et al. [37] includes strong baselines of shallow, deep, and hybrid algorithms for AD and SSAD. A brief overview of the methods is provided in Appendix B, and we refer the reader to Ruff et al. [37] for a more detailed description. Performance is evaluated by the area under the curve of the receiver operating characteristic curve (AUROC), a commonly used criterion for AD. There are two types of datasets: (1) high-dimensional datasets which were modified to be semi-supervised and (2) classic anomaly detection benchmark datasets. The first type includes MNIST, Fashion-MNIST and CIFAR-10, and the second includes datasets from various fields such as astronomy and medicine. For strengthening the baselines, Ruff et al. [37] grant the shallow and hybrid methods an unfair advantage of selecting their hyper-parameters to maximize AUROC on a subset (10%) of the test set. Here we also follow this approach.

Hyperparameters and Architectures: When comparing with the state-of-the-art Deep SAD method, we used the same network architecture but include bias terms in all layers. For MNIST, we found that this architecture did not work well, and instead used a standard convolutional neural network (see Appendix D for full details). Following the explanation in Section 4.5, we use an ensemble of $K = 5$ VAE models in all the experiments. We also report results without ensembling in Table 6 in the supplementary. Unlike Ruff et al. [37] which did not report on using a validation set to tune the hyper-parameters, we take a validation set (20%) out of the training data to tune hyper-parameters for the image datasets.

5.1 Image Datasets

MNIST, Fashion-MNIST and CIFAR-10 datasets all have ten classes. Similarly to Ruff et al. [37], we derive ten AD setups on each dataset. In each setup, one of the ten classes is set to be the normal class, and the remaining nine classes represent anomalies. During training, we set X_{normal} as the normal class samples, and $X_{outlier}$ as a small fraction of the data from **only one** of the anomaly classes. At test time, we evaluate on anomalies from **all** classes.

The experiment we perform, similarly to Ruff et al. [37], is evaluating the model's detection ability as a function of the ratio of anomalies presented to it during training. We set the ratio of labeled training data to be $\gamma_l = N_{outlier}/(N_{normal} + N_{outlier})$, and we evaluate different values of γ_l in each scenario. In total, there are 90 experiments per each value of γ_l. Note that for $\gamma_l = 0$, no labeled anomalies are presented, and we revert to standard *unsupervised* AD, which in our approach amounts to training a standard VAE ensemble. For pre-processing, pixels are scaled to $[0, 1]$.

Our results are presented in Table 2. The complete table with all of the competing methods can be found in Appendix A.1. Note that even a small fraction of labeled outliers ($\gamma_l = 0.01$) leads to significant improvement compared to the standard unsupervised VAE ($\gamma_l = 0$) and compared to the best-performing unsupervised methods on MNIST and CIFAR-10. Also, our methods consistently outperform DeepSAD and the other SSAD baselines in most domains, and the large standard deviation is a result of the high variability over the 90 different experiment settings. We validated that our improvement over Deep-SAD is statistically significant using a paired t-test ($p < 0.05$).

Following Ruff et al. [37], we also experiment with a setting where the training set is polluted with unknown outliers. The results, which are in Appendix A.1, further show the robustness and effectiveness of our methods.

CatsVsDogs Dataset In addition to the test domains of Ruff et al. [37], we also evaluate on the CatsVsDogs dataset, which is notoriously difficult for anomaly detection [16]. This dataset contains 25,000 images of cats and dogs in various positions, 12,500 in each class. Following Golan and El-Yaniv [16], We split this dataset into a training set containing 10,000 images, and a test set of 2,500 images in each class. We also rescale each image to size 64x64. We follow a similar experimental procedure as described above, and average results over the two classes.

Data	γ_l	OC-SVM Raw	OC-SVM Hybrid	Inclusive NRF	SSAD Raw	SSAD Hybrid	Deep SAD	Supervised Classifier	MML VAE	DP VAE
MNIST	.00	96.0±2.9	**96.3±2.5**	95.2±3.0	96.0±2.9	96.3±2.5	92.8±4.9		94.2±3.0	94.2±3.0
	.01				96.6±2.4	96.8±2.3	96.4±2.7	92.8±5.5	**97.3±2.1**	97.0±2.3
	.05				93.3±3.6	97.4±2.0	96.7±2.4	94.5±4.6	**97.8±1.6**	97.5±2.0
	.10				90.7±4.4	97.6±1.7	96.9±2.3	95.0±4.7	**97.8±1.6**	97.6±2.1
	.20				87.2±5.6	97.8±1.5	96.9±2.4	95.6±4.4	**97.9±1.6**	**97.9±1.8**
F-MNIST	.00	**92.8±4.7**	91.2±4.7		**92.8±4.7**	91.2±4.7	89.2±6.2		90.8±4.6	90.8±4.6
	.01				**92.1±5.0**	89.4±6.0	90.0±6.4	74.4±13.6	91.2±6.6	90.9±6.7
	.05				88.3±6.2	90.5±5.9	90.5±6.5	76.8±13.2	91.6±6.3	**92.2±4.6**
	.10				85.5±7.1	91.0±5.6	91.3±6.0	79.0±12.3	**91.7±6.4**	**91.7±6.0**
	.20				82.0±8.0	89.7±6.6	91.0±5.5	81.4±12.0	91.9±6.0	**92.1±5.7**
CIFAR-10	.00	62.0±10.6	63.8±9.0	**70.0±4.9**	62.0±10.6	63.8±9.0	60.9±9.4		52.7±10.7	52.7±10.7
	.01				73.0±8.0	70.5±8.3	72.6±7.4	55.6±5.0	73.7±7.3	**74.5±8.4**
	.05				71.5±8.1	73.3±8.4	77.9±7.2	63.5±8.0	**79.3±7.2**	79.1±8.0
	.10				70.1±8.1	74.0±8.1	79.8±7.1	67.7±9.6	80.8±7.7	**81.1±8.1**
	.20				67.4±8.8	74.5±8.0	81.9±7.0	80.5±5.9	82.6±7.2	**82.8±7.3**

Table 2: Results for image datasets. We report the average and standard deviation of AUROC over 90 experiments, for various ratios of labeled anomalies in the data γ_l.

We chose a VAE architecture similar to the autoencoder architecture in Golan and El-Yaniv [16], and for the *Deep SAD* baseline, we modified the architecture to not use bias terms and bounded activations. We report our results in Table 3, along with the numerical scores for baselines taken from Golan and El-Yaniv [16]. Note that without labeled anomalies, our method is not informative, predicting roughly at chance level, and this aligns with the baseline results reported by Golan and El-Yaniv [16]. However, *even just* 1% *labeled outliers is enough to significantly improve predictions and produce informative results*, demonstrating the potential of the SSAD. In this domain, the geometric transformations of Golan and El-Yaniv [16] allow for significantly better performance even without labeled outliers. While this approach is domain specific, incorporating similar ideas into probabilistic AD methods is an interesting direction for future research.

Data	γ_l	OC-SVM Raw	OC-SVM Hybrid	DAGMM	DSEBM	ADGAN	DADGT	Deep SAD	MML VAE	DP VAE
CatsVsDogs	.00	51.7	52.5	47.7	51.6	49.4	**88.8**	49.9	50.7	50.7
	.01							54.1	59.4	**64.0**
	.05							60.1	68.4	**70.3**
	.10							64.4	71.5	**75.3**
	.20							67.2	73.3	**78.2**

Table 3: Results for ratio of labeled anomalies γ_l in the training set, on the CatsVsDogs dataset.

5.2 Classic Anomaly Detection Datasets

AD is a well-studied field of research, with many publicly available AD benchmarks and well-established baselines [41]. These datasets are lower-dimensional than the image datasets above, and by evaluating our method on them, we aim to demonstrate the flexibility of our method for general types of data. We follow Ruff et al. [37], and consider random train-test split of 60:40 with stratified sampling to ensure correct representation of normal and anomaly classes. We evaluate over ten random seeds with 1% of anomalies, i.e., $\gamma_l = 0.01$. There are no specific anomaly classes, thus we treat all anomalies as one class. For pre-processing, we preform standardization of the features to have zero mean and unit variance. As most of these datasets have a very small amount of labeled anomalous data, it is inefficient to take a validation set from the training set. To tune hyper-parameters, we measure AUROC performance on the training data, and we follow Ruff et al. [37] and set a 150 training epochs limit. The complete hyper-parameters table is found in Appendix D.2.

In Table 4 we present the results of the best-performing algorithms along with our methods. The complete table with all the competing methods can be found in Appendix A.2. Our methods outperform the baselines on most datasets, demonstrating the flexibility of our approach.

5.3 Ablative Analysis

We perform an ablative analysis of DP-VAE method on the cardio, satellite and CIFAR-10 datasets (with 1% outliers). We evaluate our method with the following properties: (1) Frozen and unfrozen

Dataset	OC-SVM Raw	SSAD Raw	Supervised Classifier	Deep SAD	VAE	MML VAE	DP VAE
arrhythmia	84.5±3.9	**86.7±4.0**	39.2±9.5	75.9±8.7	85.6 ± 2.4	85.7 ± 2.3	**86.7 ± 1.7**
cardio	98.5±0.3	98.8±0.3	83.2±9.6	95.0±1.6	95.5 ± 0.8	**99.2 ± 0.4**	99.1 ± 0.4
satellite	95.1±0.2	**96.2±0.3**	87.2±2.1	91.5±1.1	77.7 ± 1.0	92.0 ± 1.5	89.2 ± 1.6
satimage-2	99.4±0.8	**99.9±0.1**	**99.9±0.1**	**99.9±0.1**	99.6 ± 0.6	99.7 ± 0.3	**99.9 ± 0.1**
shuttle	99.4±0.9	99.6±0.5	95.1±8.0	98.4±0.9	98.1 ± 0.6	99.8 ± 0.1	**99.9 ± 0.04**
thyroid	98.3±0.9	97.9±1.9	97.8±2.6	98.6±0.9	86.9 ± 3.7	**99.9 ± 0.04**	**99.9 ± 0.04**

Table 4: Results on classic AD benchmark datasets in the setting with a ratio of labeled anomalies of $\gamma_l = 0.01$ in the training set. We report the avg. AUROC with st.dev. computed over 10 seeds.

decoder, (2) separate and same encoder for normal data and outliers and (3) the effect of using ensembles of VAEs instead of one. Table 5 summarizes the analysis; the complete table can be found in Appendix C. It can be seen that using the same encoder is necessary as we expected, since training a neural network requires sufficient amount of data. Moreover, when dealing with a small pool of outliers, the effect on the decoder is minimal. Hence, freezing the decoder contributes little to the improvement.

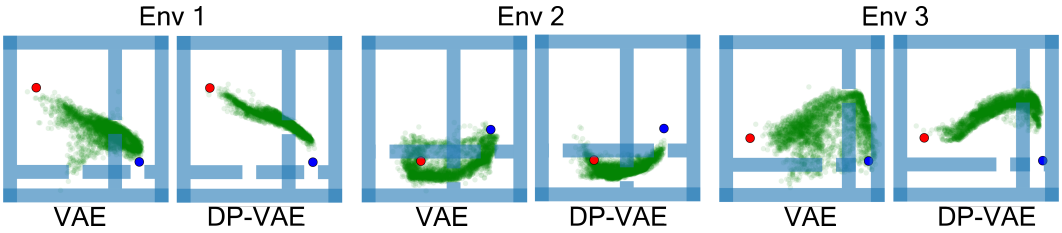

Figure 3: Motion planning with VAE. Following the work of Ichter et al. [20], a conditional VAE is trained to generate robot configurations to be used within a sampling based motion planning algorithm. The VAE generates 6D robot configurations (robot position and velocity; here projected onto the 2D plane) given an image of obstacles, an initial state (blue) and a goal state (red). Using our DP-VAE method, we added to the VAE training negative samples on obstacle boundaries. We compare samples from the standard VAE (left) and DP-VAE (right) on three unseen obstacle maps. Note that our method results in much more informative samples for planning obstacle avoidance.

Encoder	Decoder	Ensemble	Cardio AUROC	Satellite AUROC	CIFAR-10 AUROC
Separate	Unfreeze	5	96.6±0.6	65.3±1.1	49.8±10.7
Separate	Freeze	5	96.6±0.6	66.9±1.2	50.0±10.9
Same	Unfreeze	5	98.8±0.06	88.9±1.7	73.9±8.9
Same	Freeze	5	**99.1±0.4**	**91.7±1.5**	**74.5±8.4**
Same	Freeze	1	97.8±1.1	87.6±1.7	72.3±9.0

Table 5: Ablative analysis (DP-VAE). AUROC is reported over average of 10 seeds for the *satellite* and *cardio*. For *CIFAR-10*, results are reported for 1% of anomalies, averaged over 90 experiments.

5.4 Sample-Based Motion Planning Application

While our focus is on SSAD, our methods can be used to enhance any VAE generative model when negative samples are available. We demonstrate this idea in a motion planning domain [26]. Sampling-based planning algorithms search for a path for a robot between obstacles on a graph of points sampled from the feasible configurations of the system. In order to accelerate the planning process, Ichter et al. [20] suggest to learn non-uniform sampling strategies that concentrate in regions where an optimal solution might lie, given a training set of planning problems to learn from. Ichter et al. [20] proposed a conditional VAE that samples configurations conditioned on an image of the obstacles in the domain, and is trained using samples of feasible paths on a set of training domains.

Here, we propose to enhance the quality of the samples by including a few outliers (5% of the normal data) during training, which we choose to be points on the obstacle boundaries – as such points clearly do not belong to the motion plan. We used the publicly available code base of Ichter et al.

[20], and modified only the CVAE training using our DP-VAE method. Exemplary generated samples for different obstacle configurations (unseen during training) are shown in Figure 3. It can be seen that adding the outliers led to a VAE that is much more focused on the feasible parts of the space, and thus generates significantly less useless points in collision.

6 Conclusion

We proposed two VAE modifications that account for negative data examples, and used them for semi-supervised anomaly detection. We showed that these methods can be derived from natural probabilistic formulations of the problem, and that the resulting algorithms are general and effective – competitive or better than the state-of-the-art on diverse datasets. We further demonstrated that even a small fraction of outlier data can significantly improve anomaly detection on various datasets, and that our methods can be combined with VAE applications such as in motion planning and NLP.

We see great potential in the probabilistic approach to AD using deep generative models: it has a principled probabilistic interpretation, it is agnostic to the particular type of data, and it can be implemented using expressive generative models. For specific data such as images, however, discriminative approaches that exploit domain specific methods such as geometric transformations are currently the best performers. Developing similar self-supervised methods for generative approaches is an exciting direction for future research.

7 Acknowledgements

This work is partly funded by the Israel Science Foundation (ISF-759/19) and the Open Philanthropy Project Fund, an advised fund of Silicon Valley Community Foundation.

References

[1] Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE*, 2(1), 2015.

[2] Jerone Andrews, Edward Morton, and Lewis Griffin. Detecting anomalous data using auto-encoders. *International Journal of Machine Learning and Computing*, 6:21, 01 2016.

[3] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-supervised novelty detection. *J. Mach. Learn. Res.*, 11:2973–3009, December 2010.

[4] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/K16-1002. URL `https://www.aclweb.org/anthology/K16-1002`.

[5] Van Loi Cao, Miguel Nicolau, and James Mcdermott. A hybrid autoencoder and density estimation model for anomaly detection. In *PPSN*, volume 9921, pages 717–726, 09 2016.

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

[7] Jinghui Chen, Saket Sathe, Charu Aggarwal, and Deepak Turaga. Outlier detection with autoencoder ensembles. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 90–98. SIAM, 2017.

[8] Penny Chong, Lukas Ruff, Marius Kloft, and Alexander Binder. Simple and effective prevention of mode collapse in deep one-class classification, 2020.

[9] Tal Daniel and Aviv Tamar. Soft-introvae: Analyzing and improving the introspective variational autoencoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4391–4400, 2021.

[10] S. Das, W. Wong, T. Dietterich, A. Fern, and A. Emmott. Incorporating expert feedback into active anomaly discovery. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 853–858, 2016.

[11] Shubhomoy Das, Weng-Keen Wong, Alan Fern, Thomas G. Dietterich, and Md Amran Siddiqui. Incorporating feedback into tree-based anomaly detection. *CoRR*, abs/1708.09441, 2017.

[12] Lucas Deecke, Robert Vandermeulen, Lukas Ruff, Stephan Mandt, and Marius Kloft. Anomaly detection with generative adversarial networks, 2018.

[13] Adji Bousso Dieng, Dustin Tran, Rajesh Ranganath, John Paisley, and David Blei. Variational inference via χ upper bound minimization. In *Advances in Neural Information Processing Systems*, pages 2732–2741, 2017.

[14] Sarah M. Erfani, Sutharshan Rajasegarar, Shanika Karunasekera, and Christopher Leckie. High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning. *Pattern Recogn.*, 58(C):121–134, October 2016.

[15] Tolga Ergen, Ali Hassan Mirza, and Suleyman Serdar Kozat. Unsupervised and semi-supervised anomaly detection with lstm neural networks. *arXiv preprint arXiv:1710.09207*, 2017.

[16] Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. In *Advances in Neural Information Processing Systems*, pages 9758–9769, 2018.

[17] 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.

[18] Nico Görnitz, Marius Kloft, Konrad Rieck, and Ulf Brefeld. Toward supervised anomaly detection. *Journal of Artificial Intelligence Research*, 46:235–262, 2013.

[19] Martin Grill and Tomáš Pevný. Learning combination of anomaly detectors for security domain. *Comput. Netw.*, 107(P1):55–63, October 2016. ISSN 1389-1286. doi: 10.1016/j.comnet.2016.05.021.

[20] B. Ichter, J. Harrison, and M. Pavone. Learning sampling distributions for robot motion planning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7087–7094, May 2018.

[21] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder Variational Autoencoders. *arXiv e-prints*, art. arXiv:1602.02282, Feb 2016.

[22] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

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

[24] Diederik P. Kingma, Danilo Jimenez Rezende, Shakir Mohamed, and Max Welling. Semi-supervised learning with deep generative models. *CoRR*, abs/1406.5298, 2014.

[25] B Kiran, Dilip Thomas, and Ranjith Parakkal. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *Journal of Imaging*, 4(2):36, 2018.

[26] Jean-Claude Latombe. *Robot motion planning*, volume 124. Springer Science & Business Media, 2012.

[27] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, pages 7167–7177, 2018.

[28] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE, 2008.

[29] Erxue Min, Jun Long, Qiang Liu, Jianjing Cui, Zhiping Cai, and Junbo Ma. Su-ids: A semi-supervised and unsupervised framework for network intrusion detection. In *International Conference on Cloud Computing and Security*, pages 322–334. Springer, 2018.

[30] Yisroel Mirsky, Tomer Doitshman, Yuval Elovici, and Asaf Shabtai. Kitsune: An ensemble of autoencoders for online network intrusion detection. *CoRR*, abs/1802.09089, 2018.

[31] Decebal Constantin Mocanu and Elena Mocanu. One-shot learning using mixture of variational autoencoders: a generalization learning approach. *arXiv preprint arXiv:1804.07645*, 2018.

[32] Jordi Muñoz-Marí, Francesca Bovolo, Luis Gómez-Chova, Lorenzo Bruzzone, and Gustavo Camp-Valls. Semisupervised one-class support vector machines for classification of remote sensing data. *IEEE transactions on geoscience and remote sensing*, 48(8):3188–3197, 2010.

[33] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations*, 2019.

[34] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *In EMNLP*, 2014.

[35] Marco AF Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.

[36] Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Lucas Deecke, Shoaib A. Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 4393–4402, 2018.

[37] Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. In *ICLR*, 2020.

[38] 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.

[39] Bernhard Schölkopf, John C. Platt, John C. Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Comput.*, 13(7):1443–1471, July 2001.

[40] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. *arXiv preprint arXiv:2103.12051*, 2021.

[41] Rayana Shebuti. Odds library, 2016. URL `http://odds.cs.stonybrook.edu`.

[42] Wenxian Shi, Hao Zhou, Ning Miao, and Lei Li. Dispersed exponential family mixture vaes for interpretable text generation, 2020.

[43] Yunfu Song and Zhijian Ou. Learning neural random fields with inclusive auxiliary generators. *ArXiv*, abs/1806.00271, 2018.

[44] Suwon Suh, Daniel H Chae, Hyon-Goo Kang, and Seungjin Choi. Echo-state conditional variational autoencoder for anomaly detection. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 1015–1022. IEEE, 2016.

[45] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *arXiv preprint arXiv:2007.08176*, 2020.

[46] Ping Liang Tan and Robert Peharz. Hierarchical decompositional mixtures of variational autoencoders, 09–15 Jun 2019.

[47] David M. J. Tax and Robert P. W. Duin. Support vector data description. *Mach. Learn.*, 54(1):45–66, January 2004. ISSN 0885-6125.

[48] William Wang, Angelina Wang, Aviv Tamar, Xi Chen, and Pieter Abbeel. Safer classification by synthesis. *CoRR*, abs/1711.08534, 2017.

[49] Shuangfei Zhai, Yu Cheng, Weining Lu, and Zhongfei Zhang. Deep structured energy based models for anomaly detection. *CoRR*, abs/1605.07717, 2016.

[50] Ev Zisselman and Aviv Tamar. Deep residual flow for out of distribution detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

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

A Complete Results

A.1 MNIST, Fashion-MNIST, CIFAR-10

Table 7 includes the complete results. The results for an ensemble of Deep SAD models are in parenthesis (CIFAR-10). The ensemble method is implemented the same way as ours: we train $K = 5$ separate models (i.e., each model has its own c), and the score (which is the distance from c) is the average of scores from all models in the ensemble.

Anomaly Ratio	**Deep SAD**		**MML VAE**		**DP VAE**	
	$K = 1$	$K = 5$	$K = 1$	$K = 5$	$K = 1$	$K = 5$
0.0	60.9±9.4	60.7±9.4	56.4±6.5	52.7±10.7	56.4±6.5	52.7±10.7
0.01	72.6±7.4	71.7±8.4	72.6 ±8.7	73.7±7.3	72.5±8.2	74.5±8.4
0.05	77.9±7.2	77.7±7.8	77.1 ± 7.6	79.3±7.2	77.3±8.3	79.1±8.0
0.1	79.8±7.1	79.6±7.4	78.7 ± 7.7	80.8±7.7	79.1± 8.4	81.1±8.1
0.2	81.9±7.0	81.2±7.5	80.5± 7.3	82.6±7.2	80.4±8.4	82.8±7.3

Table 6: CIFAR-10 AUROC comparison between ensemble size K of Deep SAD, MML-VAE and DP-VAE models. We report the avg. AUROC with st.dev. computed over 90 experiments at various ratios γ_l.

Table 9 includes the complete results for the setting with a ratio of γ_p polluted samples. We follow Ruff et al. [37] and further investigate the robustness of our method in a scenario where the training set is polluted with unknown outliers. We fix the ratio of labeled training samples at $\gamma_l = 0.05$ and similarly to the previous experiment we only draw known outliers form one anomaly class. The results show that our methods outperform Deep SAD most of the time or show competitive performance.

A.2 Classic Anomaly Detection

Table 8 includes the complete results. We also experimented with a hybrid model, trained by combining the loss functions for MML-VAE and DP-VAE, and can be seen as a DP-VAE with the additional loss of minimizing the likelihood of outlier samples in $ELBO_{normal}(x)$. This model obtained similar performance.

B Competing Methods

We compare our methods to reported performance of both deep and shallow learning approaches, as detailed by Ruff et al. [37] and Golan and El-Yaniv [16]. For completeness, we give a brief overview of the methods.

OC-SVM/SVDD The one-class support vector machine (OC-SVM) is a kernel based method for novelty detection [39]. It is typically employed with an RBF kernel, and learns a collection of closed sets in the input space, containing most of the training samples. SVDD [47] is equivalent to to OC-SVM for the RBF kernel. OC-SVM are both granted an unfair advantage by selecting its hyper-parameters to maximize the AUROC on a subset (10%) of the test set to establish a strong baseline.

Isolation Forest (IF) Proposed by Liu et al. [28], IF is a tree-based method that explicitly isolates anomalies instead of constructing a profile of normal instances and then identify instances that do not conform to the normal profile as anomalies. As recommended in the original paper, the number of trees is set to $t = 100$, and the sub-sampling size to $\psi = 256$.

Kernel Density Estimator (KDE) The bandwidth h of the Gaussian Kernel is selected via 5-fold cross-validation using the log-likelihood following Ruff et al. [36].

Data	γ_l	OC-SVM Raw	OC-SVM Hybrid	IF Raw	IF Hybrid	KDE Raw	KDE Hybrid	CAE	Deep SVDD
MNIST	.00	96.0±2.9	96.3±2.5	85.4±8.7	90.5±5.3	95.0±3.3	87.8±5.6	92.9±5.7	92.8±4.9
	.01								
	.05								
	.10								
	.20								
F-MNIST	.00	92.8±4.7	91.2±4.7	91.6±5.5	82.5±8.1	92.0±4.9	69.7±14.4	90.2±5.8	89.2±6.2
	.01								
	.05								
	.10								
	.20								
CIFAR-10	.00	62.0±10.6	63.8±9.0	60.0±10.0	59.9±6.7	59.9±11.7	56.1±10.2	56.2±13.2	60.9±9.4
	.01								
	.05								
	.10								
	.20								

Data	γ_l	Inclusive NRF	SSAD Raw	SSAD Hybrid	SS-DGM	Deep SAD	Supervised Classifier	MML VAE	DP VAE
MNIST	.00	95.26±3.0	96.0±2.9	96.3±2.5		92.8±4.9		94.2±3.0	94.2±3.0
	.01		96.6±2.4	96.8±2.3	89.9±9.2	96.4±2.7	92.8±5.5	97.3±2.1	97.0±2.3
	.05		93.3±3.6	97.4±2.0	92.2±5.6	96.7±2.4	94.5±4.6	97.8±1.6	97.5±2.0
	.10		90.7±4.4	97.6±1.7	91.6±5.5	96.9±2.3	95.0±4.7	97.8±1.6	97.6±2.1
	.20		87.2±5.6	97.8±1.5	91.2±5.6	96.9±2.4	95.6±4.4	97.9±1.6	97.9±1.8
F-MNIST	.00		92.8±4.7	91.2±4.7		89.2±6.2		90.8±4.6	90.8±4.6
	.01		92.1±5.0	89.4±6.0	65.1±16.3	90.0±6.4	74.4±13.6	91.2±6.6	90.9±6.7
	.05		88.3±6.2	90.5±5.9	71.4±12.7	90.5±6.5	76.8±13.2	91.6±6.3	92.2±4.6
	.10		85.5±7.1	91.0±5.6	72.9±12.2	91.3±6.0	79.0±12.3	91.7±6.4	91.7±6.0
	.20		82.0±8.0	89.7±6.6	74.7±13.5	91.0±5.5	81.4±12.0	91.9±6.0	92.1±5.7
CIFAR-10	.00	70.0±4.9	62.0±10.6	63.8±9.0		60.9±9.4 [60.7±9.4]		52.7±10.7	52.7±10.7
	.01		73.0±8.0	70.5±8.3	49.7±1.7	72.6±7.4 [71.7±8.4]	55.6±5.0	73.7±7.3	74.5±8.4
	.05		71.5±8.1	73.3±8.4	50.8±4.7	77.9±7.2 [77.7±7.8]	63.5±8.0	79.3±7.2	79.1±8.0
	.10		70.1±8.1	74.0±8.1	52.0±5.5	79.8±7.1 [79.6±7.4]	67.7±9.6	80.8±7.7	81.1±8.1
	.20		67.4±8.8	74.5±8.0	53.2±6.7	81.9±7.0 [81.2±7.5]	80.5±5.9	82.6±7.2	82.8±7.3

Table 7: Complete results of the experiment where we increase the ratio of labeled anomalies γ_l in the training set. We report the avg. AUROC with st.dev. computed over 90 experiments at various ratios γ_l. In parenthesis, the results of an ensemble of Deep SAD models ($K = 5$).

Data	OC-SVM Raw	OC-SVM Hybrid	CAE	Deep SVDD	SSAD Raw	SSAD Hybrid	SS-DGM	Deep SAD	Supervised Classifier	VAE	MML VAE	DP VAE	MML-DP VAE
arrhythmia	84.5±3.9	76.7±6.2	74.0±7.5	74.6±9.0	86.7±4.0	78.3±5.1	50.3±9.8	75.9±8.7	39.2±9.5	85.6 ± 2.4	85.7 ± 2.3	86.7 ± 1.7	87.3 ± 1.7
cardio	98.5±0.3	82.8±9.3	94.3±2.0	84.8±3.6	98.8±0.3	86.3±5.8	66.2±14.3	95.0±1.6	83.2±9.6	95.5 ± 0.8	99.3 ± 0.3	99.1 ± 0.4	99.2 ± 0.4
satellite	95.1±0.2	68.6±4.8	80.0±1.7	79.8±4.1	96.2±0.3	86.9±2.8	57.4±6.4	91.5±1.1	87.2±2.1	77.7 ± 1.0	92.0 ± 1.5	89.2 ± 1.6	91.7 ± 1.5
satimage-2	99.4±0.8	96.7±2.1	99.9±0.0	98.3±1.4	99.9±0.1	96.8±2.1	99.2±0.6	99.9±0.1	99.9±0.1	99.6 ± 0.6	99.7 ± 0.3	99.9 ± 0.1	99.8 ± 0.1
shuttle	99.4±0.9	94.1±9.5	98.2±1.2	86.3±7.5	99.6±0.5	97.7±1.0	97.9±0.3	98.4±0.9	95.1±8.0	98.1 ± 0.6	99.8 ± 0.1	99.9 ± 0.04	99.9± 0.09
thyroid	98.3±0.9	91.2±4.0	75.2±10.2	72.0±9.7	97.9±1.9	95.3±3.1	72.7±12.0	98.6±0.9	97.8±2.6	86.9 ± 3.7	99.9 ± 0.04	99.9 ± 0.04	99.9 ± 0.03

Table 8: Complete results on classic AD benchmark datasets in the setting with a ratio of labeled anomalies of $\gamma_l = 0.01$ in the training set. We report the avg. AUROC with st.dev. computed over 10 seeds.

Data	γ_p	OC-SVM Raw	OC-SVM Hybrid	IF Raw	IF Hybrid	KDE Raw	KDE Hybrid	CAE	Deep SVDD
MNIST	.00	96.0±2.9	96.3±2.5	85.4±8.7	90.5±5.3	95.0±3.3	87.8±5.6	92.9±5.7	92.8±4.9
	.01	94.3±3.9	95.6±2.5	85.2±8.8	90.6±5.0	91.2±4.9	87.9±5.3	91.3±6.1	92.1±5.1
	.05	91.4±5.2	93.8±3.9	83.9±9.2	89.7±6.0	85.5±7.1	87.3±7.0	87.2±7.1	89.4±5.8
	.10	88.8±6.0	91.4±5.1	82.3±9.5	88.2±6.5	82.1±8.5	85.9±6.6	83.7±8.4	86.5±6.8
	.20	84.1±7.6	85.9±7.6	78.7±10.5	85.3±7.9	77.4±10.9	82.6±8.6	78.6±10.3	81.5±8.4
F-MNIST	.00	92.8±4.7	91.2±4.7	91.6±5.5	82.5±8.1	92.0±4.9	69.7±14.4	90.2±5.8	89.2±6.2
	.01	91.7±5.0	91.5±4.6	91.5±5.5	84.9±7.2	89.4±6.3	73.9±12.4	87.1±7.3	86.3±6.3
	.05	90.7±5.5	90.7±4.9	**90.9±5.9**	85.5±7.2	85.2±9.1	75.4±12.9	81.6±9.6	80.6±7.1
	.10	89.5±6.1	89.3±6.2	**90.2±6.3**	85.5±7.7	81.8±11.2	77.8±12.0	77.4±11.1	76.2±7.3
	.20	86.3±7.7	88.1±6.9	88.4±7.6	86.3±7.4	77.4±13.6	82.1±9.8	72.5±12.6	69.3±6.3
CIFAR-10	.00	62.0±10.6	63.8±9.0	60.0±10.0	59.9±6.7	59.9±11.7	56.1±10.2	56.2±13.2	60.9±9.4
	.01	61.9±10.6	63.8±9.3	59.9±10.1	59.9±6.7	59.2±12.3	56.3±10.4	56.2±13.1	60.5±9.4
	.05	61.4±10.7	62.6±9.2	59.6±10.1	59.6±6.4	58.1±12.9	55.6±10.5	55.7±13.3	59.6±9.8
	.10	60.8±10.7	62.9±8.2	58.8±10.1	59.1±6.6	57.3±13.5	54.9±11.1	55.4±13.3	58.6±10.0
	.20	60.3±10.3	61.9±8.1	57.9±10.1	58.3±6.2	56.2±13.9	54.2±11.1	54.6±13.3	57.0±10.6

Data	γ_p	SSAD Raw	SSAD Hybrid	SS-DGM	Deep SAD	Supervised Classifier	MML VAE	DP VAE
MNIST	.00	**97.9±1.8**	97.4±2.0	92.2±5.6	96.7±2.4	94.5±4.6	97.8±1.6	97.5±2.0
	.01	**96.6±2.4**	95.2±2.3	92.0±6.0	95.5±3.3	91.5±5.9	94.7±4.2	95.7±4.0
	.05	93.4±3.4	89.5±3.9	91.0±6.9	93.5±4.1	86.7±7.4	91.8±5.3	**93.9±4.8**
	.10	90.7±4.4	86.0±4.6	89.7±7.5	91.2±4.9	83.6±8.2	89.3±6.0	**92.3±5.4**
	.20	87.4±8.6	86.6±6.6	79.7±9.4	85.4±7.7	**88.8±6.3**		
F-MNIST	.00	**94.0±4.4**	90.5±5.9	71.4±12.7	90.5±6.5	76.8±13.2	91.6±6.3	92.2±4.6
	.01	**92.2±4.9**	87.8±6.1	71.2±14.3	87.2±7.1	67.3±8.1	88.1±9.8	85.7±10.6
	.05	88.3±6.2	82.7±7.8	71.9±14.3	81.5±8.5	59.8±4.6	89.8±7.5	80.3±13.4
	.10	85.6±7.0	79.8±9.0	72.5±15.5	78.2±9.1	56.7±4.1	88.9±11.2	75.9±15.7
	.20	81.9±8.1	74.3±10.6	70.8±16.0	74.8±9.4	53.9±2.9	**89.2±8.3**	71.5±17.4
CIFAR-10	.00	73.8±7.6	73.3±8.4	50.8±4.7	77.9±7.2	63.5±8.0	**79.3±7.2**	79.1±8.0
	.01	73.0±8.0	72.8±8.1	51.1±4.7	76.5±7.2	62.9±7.3	76.9±8.2	**78.5±8.7**
	.05	71.5±8.2	71.0±8.4	50.1±2.9	74.0±6.9	62.2±8.2	75.3±8.3	**76.8±8.9**
	.10	69.8±8.4	69.3±8.5	50.5±3.6	71.8±7.0	60.6±8.3	73.5±8.9	**75.2±9.2**
	.20	67.8±8.6	67.9±8.1	50.1±1.7	68.5±7.1	58.5±6.7	71.0±9.6	**72.2±10.3**

Table 9: Complete results of the setting where we pollute the unlabeled part of the training set with (unknown) anomalies. We report the avg. AUROC with st. dev. computed over 90 experiments at various ratios γ_p.

Semi-Supervised Anomaly Detection (SSAD) A kernel method suggested by Görnitz et al. [18] which is a generalization to SVDD to both labeled and unlabled examples. Also granted the same unfair adavantage as OC-SVM/SVDD.

Convolutional Autoencoder (CAE) Autoencoders with convolution and deconvolution layers in the encoder and decoder, respectively. We use the same architectures described in D.1 for our VAE.

Hybrid Methods In all of the the hybrid methods mentioned in the results, the inputs are representations from a converged autoencoder, instead of raw inputs.

Unsupervised Deep SVDD Two end-to-end variants of OC-SVM methods called Soft-Boundary Deep SVDD and One-Class Deep SVDD proposed by Ruff et al. [36]. They use an objective similar to that of the classic SVDD to optimize the weights of a deep architecture.

Deep Semi-supervised Anomaly Detection (Deep SAD) Recently proposed by Ruff et al. [37], Deep SAD is a general method based on deep SVDD, which learns a neural-network mapping of the input that minimizes the volume of data around a predetermined point.

Semi-Supervised Deep Generative Models (SS-DGM) [24] proposed a deep variational generative approach to semi-supervised learning. In this approach, a classifier is trained on the latent space embeddings of a VAE, which are lower in dimension than the original input. They also propose a probabilistic model that describes the data using the available labels. These two models are fused together to a stacked semi-supervised model.

Deep Structured Energy-Based Models (DSEBM) Proposed by Zhai et al. [49], DSEBM is a deep neural technique, whose output is the energy function (negative log probability) associated with an input sample. The chosen architecture, as described by Golan and El-Yaniv [16] is the same as that of the encoder part in the convolutional autoencoder used by OC-SVM Hybrid.

Deep Autoencoding Gaussian Mixture Model (DAGMM) Proposed by Zong et al. [51], DAGMM is an end-to-end deep neural network that leverages Gaussian Mixture Modeling to perform density estimation and unsupervised anomaly detection in a low-dimensional space learned by deep autoencoder. It simultaneously optimizes the parameters of the autoencoder and the mixture model in an end-to-end fashion, thus leveraging a separate estimation network to facilitate the parameter learning of the mixture model. The architecture, as described by Golan and El-Yaniv [16] is the same as of the autoencoder we used is similar to that of the convolutional autoencoder used in OC-SVM Hybrid.

Anomaly Detection with a Generative Adversarial Network (ADGAN) A GAN-based model, proposed by Deecke et al. [12]. Anomaly detection is done with GANs by searching the generator's latent space for good sample representations. In the experiments performed by Golan and El-Yaniv [16], the generative model of the ADGAN had the same architecture used by the authors of the original paper.

Deep Anomaly Detection using Geometric Transformations (DADGT) A stae-of-the-art deep anomaly detection in images method proposed by Golan and El-Yaniv [16]. In this method, features are learned using a self-supervised paradigm – by applying geometric transformations to the image and learning to classify which transformation was applied.

Inclusive Neural Random Fields (Inclusive-NRF) A state-of-the-art energy-based model proposed by Song and Ou [43]. The inclusive-NRF learns neural random fields for continuous data by developing inclusive-divergence minimized auxiliary generators and stochastic gradient sampling. As this model directly provides a density estimate, it is an efficient tool for AD, as the estimated density can be used as the decision criterion.

C Complete Ablative Analysis

The complete ablative analysis is reported in table 10.

Encoder	Decoder	Ensemble	Cardio AUROC	Satellite AUROC	CIFAR-10 AUROC
Separate	Unfreeze	5	96.6±0.6	65.3±1.1	49.8±10.7
Separate	Freeze	5	96.6±0.6	66.9±1.2	50.0±10.9
Same	Unfreeze	5	98.8±0.06	88.9±1.7	73.9±8.9
Same	Freeze	5	**99.1±0.4**	**91.7±1.5**	**74.5±8.4**
Separate	Unfreeze	1	82.5±3.3	63.1±5.3	50.4±10.2
Separate	Freeze	1	88.8±0.9	66.2±3.4	49.2±10.4
Same	Unfreeze	1	96.8±1.5	90.0±2.9	72.8±8.6
Same	Freeze	1	97.8±1.1	87.6±1.7	72.3±9.0

Table 10: Complete ablative analysis of the Dual Prior method. AUROC is reported over average of 10 seeds for the *satellite* and *cardio*. For *CIFAR-10*, results are reported for the experiment with 1% of anomalies, averaged over 90 experiments.

D Implementation Details

We provide essential implementation information and describe how we tuned our models.

D.1 Network Architectures

Our work is based on deep variational autoencoders which require an encoder network and a decoder network. For the architectures, we follow Ruff et al. [37] architectures for the autoencoders, with the only differences being the use of bias weights in our architectures and instead of autoencoders we use *variational* autoencoders (i.e. the encoder outputs mean and standard deviation of the latent variables).

For the image datasets, LeNet-type convolutional neural networks (CNNs) are used. Each convolutional layer is followed by batch normalization and leaky ReLU activations ($\alpha = 0.1$) and 2×2-max pooling. On Fashion-MNIST, we use two convolutional layers, one the first with $16 \times (5 \times 5)$ filters and second with $32 \times (5 \times 5)$ filters. Following are two dense layers of 64 and 32 units respectively. On CIFAR-10, three convolutional layers of sizes $32 \times (5 \times 5)$,$64 \times (5 \times 5)$ and $128 \times (5 \times 5)$ filters are followed by a final dense layer of 128 units (i.e. the latent space is of dimension 128). For MNIST, we use a different architecture where the first layer is comprised of $64 \times (4 \times 4)$-filters and the second $128 \times (4 \times 4)$-filter with ReLU activations and no batch normalization. We use two dense layers, the first with 1024 units and the final has 32 units. For CatsVsDogs we follow [16] architecture for the autoencoder. We employ three convolutional layers, each followed by batch normalization and ReLU activation. The number of filters in each layer are $128 \times (3 \times 3)$, $256 \times (3 \times 3)$ and $512 \times (3 \times 3)$ respectively. The final dense layer has 256 units, which is the latent space dimension.

For the classic AD benchmark datasets, we use regular multi-layer perceptrons (MLPs). On arrhythmia, a 3-layer MLP with 128-64-32 units. On cardio, satellite, satimage-2 and shuttle, we use a 3-layer MLP with 32-16-8 units. On thyroid a 3-layer MLP with 32-16-4 units.

For the motion planning, we follow Ichter et al. [20] and use two-layer MLP with 512 hidden units and latent space dimension of 256. We use dropout ($p = 0.5$) as regularization and ReLU activation for the hidden layers.

D.2 Hyper-parameters and Tuning

Tuning models in a semi-supervised setting is not a trivial task, as usually there is abundance of data from one class and a small pool of data from the other classes. Thus, it is not clear whether one should allocate a validation set out of the training set (and by doing that, reducing the number of available samples for training) or just evaluate the performance on the training set and hope for the best. Ruff et al. [37] didn't use a validation set, but predetermined the total number of epochs to run and finally evaluated the performance on the test set. We, on the other hand, decided to take a validation set out of the training set for the image datasets, as we have enough data. The validation set is composed of unseen samples from the normal class and samples from the *current* outlier class unlike the test set, which is composed of samples from all ten classes (9 outlier classes). On the classic AD benchmark datasets, as there is very few outlier data, we evaluate the performance during training on the training set itself without taking a validation set. Finally, we evaluate the performance on the test set. For all datasets, we used a batch size of 128 and an ensemble of 5 VAEs. For the image datasets, we run 200 epochs and for the classical AD benchmarks we run

150 epochs. For the MML method, we found $\gamma = 1$ to perform well on all datasets. Furthermore, similarly to the additional hyper-parameter β_{KL} in the ELBO term, we add β_{CUBO} to the CUBO term as derived in Equation 4. The motivation for adding this balancing hyper-parameter is that since $\mathcal{L} = \exp\{n \cdot CUBO_n(q_\phi(z|x))\}$ reaches the same optima as $CUBO_n(q_\phi(z|x))$ as explained earlier, so does $\mathcal{L} = \exp\{\beta \cdot n \cdot CUBO_n(q_\phi(z|x))\}$. Finding a good value for β_{CUBO} can contribute to the performance, as reflected in the results. Moreover, for the optimization of the CUBO component, we used gradient clipping and learning rate scheduling (every 50 epochs, learning rate is multiplied by 0.1). Tables 11 and 12 summarize all the hyper-parameters chosen for our model per dataset.

Data	**ND Update Interval**	**Learning Rate**	β_{KL}	α
MNIST	2	0.001	0.005	10
Fashion-MNIST	2	0.001	0.005	10
CIFAR-10	2	0.001	0.005	10
CatsVsDogs	1	0.001	0.005	10
arrhythmia	1	0.0005	0.5	2
cardio	1	0.001	0.05	5
satellite	1	0.001	0.05	5
satimage-2	1	0.0005	0.05	10
shuttle	1	0.001	0.05	5
thyroid	1	0.0001	0.05	10

Table 11: Hyper-parameters for the Dual Prior VAE

Data	**ND Update Interval**	**Learning Rate**	β_{KL}	β_{CUBO}
MNIST	2	0.0005 (0.0002)	0.005	0.005 (0.02)
Fashion-MNIST	2	0.001	0.005	0.005 (0.25)
CIFAR-10	2	0.001	0.005	0.005
CatsVsDogs	1	0.0005	0.005	0.005
arrhythmia	1	0.0005	0.5	0.5
cardio	1	0.001	0.05	0.05
satellite	1	0.001	0.05	0.05
satimage-2	1	0.001	0.05	0.05
shuttle	1	0.001	0.05	0.05
thyroid	1	0.0001	0.05	0.05

Table 12: Hyper-parameters for the Max-Min Likelihood VAE. In parenthesis, changes for the setting with polluted samples.

For training, we follow Kaae Sønderby et al. [21] recommendations for training VAEs and use a 20-epoch annealing for the KL-divergence component, that is, the KL-divergence coefficient, β_{KL}, is linearly increased from zero to its final value. Moreover, we allow the VAE to first learn a good representation of the normal data in a warm-up period of 50 epochs, and then we begin applying the novelty detection updates. For optimization, we use Adam [22] with a learning rate schedule to stabilize training when outlier samples are fed after the warm-up period.

E CUBO Loss Derivation

The χ-divergence is defined as follows:

$$D_{\chi^2}(p\|q) = \mathbb{E}_{q(z;\theta)}\big[\big(\frac{p(z|x)}{q(z;\theta)}\big)^2 - 1\big]$$

The general upper bound as derived in [13]:

$$\mathcal{L}_{\chi^2}(\theta) = CUBO_2 = \frac{1}{2}\log \mathbb{E}_{q(z;\theta)}\big[\big(\frac{p(x,z)}{q(z;\theta)}\big)^2\big]$$

The optimized CUBO:

$$\mathcal{L} = \exp\{2 \cdot CUBO_2(\theta)\} = \mathbb{E}_{q(z;\theta)}\big[\big(\frac{p(x,z)}{q(z;\theta)}\big)^2\big]$$

In the our VAE framework, we denote $\mathcal{L}_R = -\log p(X|z)$, the reconstruction error between the output of the decoder and the original input. We assume that $q(z|X) \sim \mathcal{N}(\mu_q(X), \Sigma_q(X))$ and that

in general, $p_{outlier}(z) \sim \mathcal{N}(\mu_o, I)$, but we note that in our case, we set $\mu_o = 0$, for both CUBO and hybrid methods. As the CUBO loss is derived only for the anomalous data, we omit the labels in the following. β_{CUBO} is a balancing hyper-parameter we add, as mentioned in D.2.

$$\mathcal{L}_{CUBO} = \mathbb{E}_{q(z;\theta)}\big[\big(\frac{p(x,z)}{q(z;\theta)}\big)^2\big] = \mathbb{E}_q\big[\exp\{2\log p(X|z) + 2\log\big(\frac{p(z)}{q(z|X)}\big)\}\big] =$$

$$\mathbb{E}_q\big[\exp\{-2\cdot\mathcal{L}_R + 2\beta_{CUBO}\cdot(\log p(z) - \log q(z|X))\}\big] =$$

$$\mathbb{E}_q\big[\exp\{-2\cdot\mathcal{L}_R + 2\beta_{CUBO}\cdot\big(-\frac{1}{2}(z-\mu_o)^T(z-\mu_o) - [-\frac{1}{2}\log|\Sigma_q| - \frac{1}{2}(z-\mu_q)^T\Sigma_q^{-1}(z-\mu_q)]\big)\}\big] =$$

$$\mathbb{E}_q\big[\exp\{-2\cdot\mathcal{L}_R + \beta_{CUBO}\cdot\big(-[z^Tz - 2z^T\mu_o + \mu_o^T\mu_o] + \log|\Sigma_q| + z^T\Sigma_q^{-1}z - 2z^T\Sigma_q^{-1}\mu_q + \mu_q^T\Sigma_q^{-1}\mu_q\big)\}\big] =$$

$$\mathbb{E}_q\big[\exp\{-2\cdot\mathcal{L}_R + \beta_{CUBO}\cdot\big(\log|\Sigma_q| + \mu_q^T\Sigma_q^{-1}\mu_q - \mu_o^T\mu_o - z^Tz + 2z^T\mu_o + z^T\Sigma_q^{-1}z - 2z^T\Sigma_q^{-1}\mu_q\big)\}\big] =$$

$$\exp\{\beta_{CUBO}\cdot\big(\log|\Sigma_q| + \mu_q^T\Sigma_q^{-1}\mu_q - \mu_o^T\mu_o\big)\}\cdot \mathbb{E}_q\big[\exp\{-2\cdot\mathcal{L}_R + \beta_{CUBO}\cdot\big(-z^Tz + 2z^T\mu_o + z^T\Sigma_q^{-1}z - 2z^T\Sigma_q^{-1}\mu_q\big)\}\big] =$$

$$\exp\{\beta_{CUBO}\cdot\big(\log|\Sigma_q| + \mu_q^T\Sigma_q^{-1}\mu_q - \mu_o^T\mu_o\big)\}\cdot \exp\{\log\mathbb{E}_q\big[\exp\{-2\cdot\mathcal{L}_R + \beta_{CUBO}\cdot\big(-z^Tz + 2z^T\mu_o + z^T\Sigma_q^{-1}z - 2z^T\Sigma_q^{-1}\mu_q\big)\}\big]\} =$$

$$\exp\{\beta_{CUBO}\cdot\big(\log|\Sigma_q| + \mu_q^T\Sigma_q^{-1}\mu_q - \mu_o^T\mu_o\big) + \log\mathbb{E}_q\big[\exp\{-2\cdot\mathcal{L}_R + \beta_{CUBO}\cdot\big(-z^Tz + 2z^T\mu_o + z^T\Sigma_q^{-1}z - 2z^T\Sigma_q^{-1}\mu_q\big)\}\big]\} \quad (4)$$

The expectation is estimated with Monte Carlo and for numeric stability we employ the commonly used *log-sum-exp trick*.

F ELBO With Gaussian Prior

We provide the derivation for the ELBO objective function where the prior, $p(z)$ is a Gaussian with non-zero mean, that is, $z \sim \mathcal{N}(\mu_o, I)$. Under the i.i.d. assumption in the VAE framework, we assume that $z_i \sim \mathcal{N}(\mu_{o_i}, 1)$ and $z_i|x \sim \mathcal{N}(\mu_{q_i}, \sigma_{ii}^2)$. Thus, it holds that:

$$\mathbb{E}_q[z_i^2] = \sigma_{ii}^2 + \mu_{q_i}^2$$

We now derive the KL-divergence component of the ELBO:

$$D_{\mathrm{KL}}[q(z|X)\|p(z)] = \mathbb{E}_q\big[\frac{\log q(z|X)}{\log p(z)}\big] = \mathbb{E}_q[\log q(z|X)] - \mathbb{E}_q[\log p(z)]$$

$$\mathbb{E}_q[\log q(z|X)] = -\frac{1}{2}\log|\Sigma_q| - \frac{1}{2}\mathbb{E}_q[(z-\mu_q)^T\Sigma_q^{-1}(z-\mu_q)] =$$

$$-\frac{1}{2}\log|\Sigma_q| + \frac{1}{2}\mu_q^T\Sigma_q^{-1}\mu_q - \frac{1}{2}\mathbb{E}_q[z^T\Sigma_q^{-1}z] =$$

$$-\frac{1}{2}\sum_{i=1}^{n}[\log\sigma_{ii}^2 - \frac{\mu_{q_i}^2}{\sigma_{ii}^2} + 1 + \frac{\mu_{q_i}^2}{\sigma_{ii}^2}] = -\frac{1}{2}\sum_{i=1}^{n}[\log\sigma_{ii}^2 + 1]$$

$$\mathbb{E}_q[\log p(z)] = -\frac{1}{2}\mathbb{E}_q[(z-\mu_o)^T(z-\mu_o)] =$$

$$\frac{1}{2}\sum_{i=1}^{n}[\sigma_{ii}^2 + \mu_{q_i}^2 - 2\mu_{q_i}\mu_{o_i} + \mu_{o_i}^2]$$

Finally:

$$D_{\mathrm{KL}}[q(z|X)\|p(z)] = -\frac{1}{2}\sum_{i=1}^{n}[1 + \log\sigma_{ii}^2 - \sigma_{ii}^2 - \mu_{q_i}^2 + 2\mu_{q_i}\mu_{o_i} - \mu_{o_i}^2]$$

G Deep SVDD/SAD with Ill-trained Autoencoder

To demonstrate that Deep SAD and Deep SVDD are highly dependent on the representation ability of the autoencoder, we train the autoencoder for 10 epochs instead of 150, and then train normally using the objective of Deep SAD/Deep SVDD. Experiment was conducted on the CIFAR-10 dataset. The results presented in Table 13 clearly show Deep SVDD/SAD's objective function has no effect on anomaly detection.

Anomaly Ratio	**CIFAR-10 AUROC**
0.0	59.8±9.8
0.01	59.4±9.6
0.05	58.8±9.4
0.1	58.4±10.0
0.2	56.9±11.2

Table 13: Deep SAD results for 10 epochs of autoencoder pre-training on *CIFAR-10*. Results are reported for the experiment are averaged over 90 experiments. For no anomalies, Deep SAD reduces to Deep SVDD.

H Sentiment Analysis Experiment Details

In this section, we further demonstrate the generality of our method by considering a Natural Language Processing (NLP) application. One of the most important tasks in NLP is the task of *sentiment analysis*. The objective is to identify opinions expressed in a piece of text, e.g., to determine whether the writer's attitude towards a particular topic is positive, negative, or neutral. Bowman et al. [4] proposed learning latent representations of entire sentences using a recurrent neural network (RNN) based VAE. Typically, sentiment analysis is learned in a supervised setting. Here, we consider identifying sentiment anomalies. E.g., having seen only data for positive opinions, identify when a negative opinion is encountered. Using our SSAD method, we shall show that even a small number of labelled outliers can significantly improve detection.

We experiment on the IMDB review dataset, which contains 30K positive and negative movie reviews. In our experiment, we use only the first sentence of the review (which leads to a noisy dataset as the first sentence does not necessarily express the reviewer opinion, e.g., includes a short summary of the movie). Unlike images, text data does not have a fixed size or structure. Moreover, samples from different domains/sentiments can have the same structure (e.g. "this is the best movie"–"this is the worst movie"), with the difference being limited only to certain words. Since the VAE represents whole sentences in the latent space, it seems that detecting sentiment novelties would be challenging.

We first take the first sentence of each review that has at least 5 words and no more than 20 words. We then separate to train (0.7) and test (0.3) sets and build a vocabulary of word representations using GloVe [34]. We follow the same architecture and training procedure as in the original paper [4], including annelaing for the KL term and word dropout (0.5), and train with unlabeled data for 480 epochs. Finally, we supply labeled anomalies for 5 epochs and measure the AUROC. As for the MML-VAE parameters, we used $\gamma = 1$.

We demonstrate results using MML-VAE as the SSAD method in Table 14.As our results show, using only 1% of outliers (i.e., reviews from the opposing sentiment), leads to a significant improvement.

Outlier Ratio	**Positive AUROC**	**Negative AUROC**	**Mean AUROC**
0.0	52.8±0.01	60.9±0.004	56.9±4.0
0.01	56.6±5.0	64.8±1.5	60.7±5.4

Table 14: Complete MML-VAE AUROC results for the task of sentiment analysis on the IMDB dataset. Each column corresponds to the unlabeled train data. The results are averaged over 5 runs.