Diminishing Empirical Risk Minimization for Unsupervised Anomaly Detection

Shaoshen Wang*, Yanbin Liu†, Ling Chen*, and Chengqi Zhang*

*Australian Artificial Intelligence Institute, University of Technology Sydney, Sydney, Australia

†Centre for Medical Research, The University of Western Australia, Perth, Australia

{shaoshenwang, csyanbin}@gmail.com, {Ling.Chen, Chengqi.Zhang}@uts.edu.au

Abstract—Unsupervised anomaly detection (AD) is a challenging task in realistic applications. Recently, there is an increasing trend to detect anomalies with deep neural networks (DNN). However, most popular deep AD detectors cannot protect the network from learning contaminated information brought by anomalous data, resulting in unsatisfactory detection performance and overfitting issues. In this work, we identify one reason that hinders most existing DNN-based anomaly detection methods from performing is the wide adoption of the Empirical Risk Minimization (ERM). ERM assumes that the performance of an algorithm on an unknown distribution can be approximated by averaging losses on the known training set. This averaging scheme thus ignores the distinctions between normal and anomalous instances. To break through the limitations of ERM, we propose a novel Diminishing Empirical Risk Minimization (DERM) framework. Specifically, DERM adaptively adjusts the impact of individual losses through a well-devised aggregation strategy. Theoretically, our proposed DERM can directly modify the gradient contribution of each individual loss in the optimization process to suppress the influence of outliers, leading to a robust anomaly detector. Empirically, DERM outperformed the state-of-the-art on the unsupervised AD benchmark consisting of 18 datasets.

Index Terms—Unsupervised Anomaly Detection, Empirical Risk Minimization, Autoencoder

I. INTRODUCTION

Anomaly detection (AD) is an important research topic in data mining and machine learning [5], [7], [37]. It aims to identify data points that do not conform to expected behaviors. Since anomalies usually provide critical information, AD has been widely-used in various applications, such as health care [33], [35], network intrusion detection [21], [39], fraud detection [30] and other areas [8], [26]. In many realistic scenarios, there is no ground truth available to distinguish anomalous instances from the normal ones. The only assumption is that the proportion of normal instances is much higher than that of anomalies in a given dataset. According to this assumption, researchers usually resort to unsupervised approaches to cope with the situation, i.e. unsupervised anomaly detection (AD).

A multitude of methods have been proposed to tackle the unsupervised AD problem [32], [34], [46], [48], [49]. Recently, there is an increasing trend to use Deep Neural Networks (DNN) to solve the problem of unsupervised AD, as DNN-based methods show improved performance compared with traditional machine learning models, especially when the scale of data increases. For example, deep autoencoder has become the core of most deep unsupervised AD approaches [5], [7], as it can powerfully extract and preserve intrinsic information from data. Specifically, an autoencoder learns the latent representation from the original data by minimizing the reconstruction loss. The anomalies are then detected based on the assumption that normal instances are more likely to be compressed and reconstructed than the anomalous ones.

Despite the progress, there remains a major issue for deep AD detectors: most existing deep autoencoder AD methods cannot prevent the network from aggressively fitting the anomalies, which leads to the overfitting issue and the unsatisfactory performance in turn. One of the obvious reasons is that DNN are often over-parameterized and designed with non-linear activation function between layers, which makes DNN an universal approximator [12] fitting well with not only normal data but also the anomalies [45]. Recent works [24] have attempted to mitigate this issue by an elaborate design for model capacity.

In this work, we analyze that another reason which makes DNN non-ideal for the scenario of unsupervised AD is the concept of empirical risk minimization (ERM), which is widely adopted by DNN. ERM assumes that the performance of a learning algorithm on an unknown data distribution can be approximated by averaging the losses on the known training set, as follows:

$$\bar{R}(\theta) := \frac{1}{N} \sum_{i \in [N]} f(x_i; \theta), \tag{1}$$

where $f(\cdot; \theta)$ is a loss function parameterized by θ and x_i is the i-th training instance from a dataset of size N. Take the deep autoencoder anomaly detector as an example. An anomalous instance x_A tends to have a larger reconstruction loss $f(x_A; \theta)$. Under the ERM scheme, since the equivalent weights (i.e., $\frac{1}{N}$) treat all data equally during loss aggregation, an anomalous instance contributes even more to the overall loss $\bar{R}(\theta)$ than a normal instance. Consequently, the model parameters θ are optimized based on a total loss $\bar{R}(\theta)$ that is severely influenced by the losses incurred by anomalies, making the deep model

wrongly focus on and fit with the unwanted anomalies. In other words, ERM ignores the distinctions between the normal and the anomalous instances. While some methods [24], [36] have proposed to address this problem by either discarding or assigning low weights to potential anomalies during the training process, they are hard to well generalize on test data because these methods rely on inflexible ad-hoc or manual selection of potential anomalous instances.

To tackle the problem brought by ERM, we propose a novel Diminishing Empirical Risk Minimization (DERM) framework to adaptively adjust the impact of each individual loss. For $t \in \mathbb{R}^+$, DERM takes the following form:

$$\tilde{R}_{\text{DERM}}(t; \theta) := e^{\frac{1}{N} \sum\limits_{i \in [N]} \log(tf(x_i; \theta))}. \tag{2}$$

The effectiveness of our design in Eq. 2 comes from the intrinsic property of the logarithm function. Specifically, logarithm function is a slowly increasing function (i.e., its derivative is decreasing). When $f(x_i; \theta)$ has a larger value (usually for anomalies), the logarithm output $\log(tf(x_i; \theta))$ will suppress $f(x_i; \theta)$ more quickly. And the parameter t controls how intensively the suppression will take effect. In this way, Eq. 2 weakens the impact of potential anomalies (leading to larger loss values) in a dynamic and controllable manner. Moreover, by comparing the gradients of ERM (Eq. 1) and DERM (Eq. 2) w.r.t θ, we theoretically find that DERM can directly modify the gradient contribution of each individual loss term. In particular, if the loss $f(x_i; \theta)$ is larger than the average (i.e., $\tilde{R}_{\text{DERM}}(t; \theta)$), the gradient will be diminished. With this property, the potential anomalies will contribute less to the optimization process, leading to a more robust anomaly detection model.

In the experiments, we first verify the effectiveness of the proposed DERM on anomaly suppression and gradient diminishing using both the synthetic and real-world datasets. And then we conduct comprehensive ablation study and comparison with the state-of-the-art approaches. Our contributions can be summarized as follows:

- We propose a novel Diminishing Empirical Risk Minimization (DERM) framework for unsupervised anomaly detection. In DERM, the adverse effect of the potential anomalies are suppressed in a dynamic and controllable manner.
- We conduct theoretical analysis on DERM and reveal that DERM can directly modify the gradient contribution of each individual loss. Specifically, each gradient contribution is proportional to the ratio of the average loss to a single loss.
- We perform extensive experiments on both synthetic and real-world datasets to verify the efficacy of DERM. Experimental results demonstrate improved performance of DERM on the unsupervised AD benchmark consisting of 18 datasets.

II. RELATED WORKS

A. *Unsupervised Anomaly Detection (AD).*

Anomaly detection is a significant study field in machine learning and data mining [5], [7], [29], [37]. Unsupervised anomaly detection does not require any data with labelling. The only assumption is that the number of normal data points is larger than the number of anomalies. A number of methods have been proposed for unsupervised AD [5], [7], [29]. Traditional methods tend to choose Principal Component Analysis (PCA) [41], Isolation Forest [25] and Support Vector Machine (SVM) [9] to detect anomalies in an unsupervised manner.

Recently, a large amount of representation learning approaches equipped with deep neural network have aroused great interest in this space. The core idea is to learn an useful representation by minimizing the reconstruction loss. These reconstruction-based methods learn a low-dimensional vector in the latent space and then project it back to the original feature space. The reconstruction error between the input and the reconstructed output is treated as the anomaly score. Early work [42] in this branch is proposed on anomaly detection with the representation learning capability of autoencoder, which utilized the large reconstruction error to detect anomalies. Subsequent work combined diverse techniques or prior knowledge with autoencoder to enhance the detection efficacy. RDA [48] combined the robust PCA with an autoencoder to group the data into a mixture of normal and anomaly components. DAGMM [49] trained a Gaussian Mixture Model to learn the latent representation from autoencoder to determine anomalies jointly by reconstruction error and density estimation. In addition, one-class classification and its deep neural network variants are also widely used for anomaly detection [9], [32]. The decision boundary surrounding normal instances is also learned for anomaly detection.

However, these methods often rely on the DNN to extract information, and most existing deep AD approaches fail to prevent the DNN from aggressively learning the anomalies. Recent studies attempted to address this issue via different strategies. RCA [24] proposed to discard a proportion of suspicious anomaly data through a collaborative autoencoder. RSR [19] utilized a robust subspace recovery layer to extract a subspace from the given data and move the outliers further away from the subspace. Different from such algorithms, we resolve this issue from the learning principle perspective and propose the Diminishing Empirical Risk Minimization (DERM) framework. DERM adaptively adjusts the individual loss contribution of each instance to diminish the outliers.

B. *Autoencoders for Anomaly Detection*

Autoencoders have been widely-adopted in unsupervised anomaly detection [3], [11], [40], [43], [49]. The general idea is to train an autoencoder on the entire dataset (both normal and anomalous) and utilize the reconstruction error as the detection criterion. As there are fewer anomalous data for training, the reconstruction errors of anomalies are higher

than that of normal ones, which can be utilized to separate the anomalies from normal data. There are subsequent work that seek help from traditional machine learning algorithms to enhance the capability of vanilla autoencoder. For example, robust autoencoder [6] incorporates RPCA [4] in an autoencoder, where parameters of autoencoder and a sparse residual matrix are alternatively optimized. Normalized deep autoencoder [2] considers the situation of multiple modes for normal instances and also applies L_2 normalization for latent variables of the autoencoder. What's more, MemAE [11] proposes a memory-augmented autoencoder to improve the performance of unsupervised anomaly detection. Nevertheless, most existing approaches overlook the deficiency of the default ERM principle in AD applications. In this work, we put forward a novel DERM framework to circumvent the drawback of ERM.

C. Sample Re-weighting and Aggregation Schemes.

Approaches have been proposed to re-weight the influence of samples by modifying the ERM objective. In [15], [20], examples were re-weighted as per their loss values to intervene the optimization dynamics, which pays more attention to difficult examples. *Relaxed clipping* [44] performed the example re-weighting via loss clipping. There are other works trying to modify the loss aggregation scheme. One of the alternatives to traditional average loss in ERM is to consider a min-max objective, which tries to minimize the max loss. The min-max objective has been applied in application such as enhancing robustness under perturbations [38].

Sample re-weighting has also been applied on anomaly detection. Most of existing deep anomaly detection approaches fail to protect the neural network from interfering by anomalies during parameter learning. Some work address this issue via assigning different weights to corresponding data point. For instance, self-paced learning model [18] and Mentornet [13] assign higher weights to instances which are easier to be classified. Recently, TERM [22] proposed the tilted empirical risk minimization, which tuned the impact of individual losses by applying different gradient weights on them. However, existing methods are less effective when applied to unsupervised AD. By contrast, DERM is specially designed for the unsupervised AD task by adaptively re-weighting examples with a new learning principle. Moreover, DERM can directly modify the gradient contribution of each example from a theoretical perspective.

III. METHODOLOGY

A. Diminishing Empirical Risk Minimization (DERM)

a) **Problem Setting.**: Suppose we are given a dataset $X = \{x_1, x_2, \ldots, x_n\}$, where $x_i \in X \subseteq \mathbb{R}^d$ has d features. Our objective is to classify each x_i as either an anomalous instance or a normal one in an unsupervised manner.

b) **DERM Framework.**: DERM learns a set of k autoencoders $\{(E_j, D_j) | j = 1, 2 \ldots k\}$ with different initializations. Here, we set $k = 2$ for brevity (the discussion be easily generalized to the situation when $k > 2$). As shown in Figure 1,

in each iteration, a mini-batch of data is randomly sampled and fed into the autoencoders. Then, the reconstruction loss for each data instance x_i is calculated as:

$$f(x_i; \theta) = \sum_{j=1,2,\ldots,k} ||x_i - D_j(E_j(x_i))||_2^2. \qquad (3)$$

To obtain the final loss, as discussed before, previous work utilize the ERM (Eq. 1) to aggregate all the instance losses with equal weights. However, it ignores the distinctions between the normal and anomalous instances, thus overfitting the anomalies. To address this issue, we devise the DERM (Eq. 2) to automatically and dynamically adjust the loss weights for each instance. With the DERM, the normal instances are well learned and result in small losses, while the anomalous instances are less-trained and cause large losses due to their inconsistent behaviour and the lack of data. Then, a back-propagation step is conducted to update the parameters of the whole model by gradient-based optimization. During the testing phase, the reconstruction loss $f(x_i; \theta)$ is adopted as an anomaly score of the testing instance. The details of the DERM framework (k=2) is shown in Algorithm 1.

c) **Theoretical Analysis.**: To dig deep into the motivation and mechanism, we present the theoretical analysis for the proposed DERM framework. Using traditional ERM is equivalent to obtaining the mean loss of the training samples, which can be biased towards (or negatively affected by) outlier data. By the following Theorem, we show that DERM can suppress the anomalies by reducing the gradient contribution of the potential anomalous candidates.

Theorem 1: For a continuously differentiable (i.e., smooth) loss function $f(x; \theta)$, the gradient of DERM (Eq. 2) is:

$$\nabla_\theta \tilde{R}_{\text{DERM}}(t; \theta) = \sum_{i \in [N]} w_i(t; \theta) \nabla_\theta f(x_i; \theta)$$

$$\text{where } w_i(t; \theta) = \frac{1}{N} \frac{\tilde{R}_{\text{DERM}}(t; \theta)}{f(x_i; \theta)}.$$

Proof 1:

$$\nabla_\theta \tilde{R}_{\text{DERM}}(t; \theta)$$

$$= \nabla_\theta \left\{ e^{\frac{1}{N} \sum_{i \in [N]} \log t f(x_i; \theta)} \right\}$$

$$= \left(e^{\frac{1}{N} \sum_{i \in [N]} \log t f(x_i; \theta)} \right) \cdot \frac{1}{N} \sum_{i \in [N]} \nabla_\theta \log t f(x_i; \theta)$$

$$= \left(e^{\frac{1}{N} \sum_{i \in [N]} \log t f(x_i; \theta)} \right) \cdot \frac{1}{N} \sum_{i \in [N]} \frac{\nabla_\theta f(x_i; \theta)}{f(x_i; \theta)}$$

$$= \sum_{i \in [N]} \frac{1}{N} \cdot \frac{e^{\frac{1}{N} \sum_{i \in [N]} \log t f(x_i; \theta)}}{f(x_i; \theta)} \nabla_\theta f(x_i; \theta)$$

$$= \sum_{i \in [N]} \frac{1}{N} \cdot \frac{\tilde{R}_{\text{DERM}}(t; \theta)}{f(x_i; \theta)} \nabla_\theta f(x_i; \theta)$$
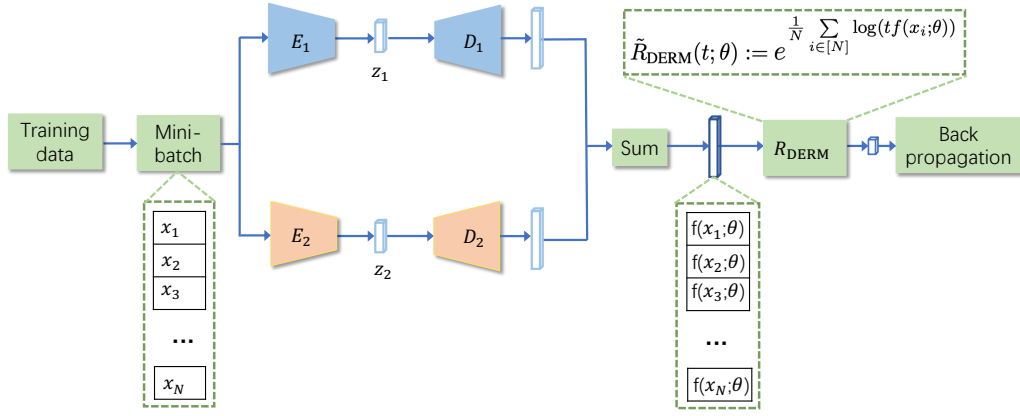
∎

Fig. 1. Illustration of the DERM framework in the training phase. A batch of training data is sampled as the input for k ($k = 2$ in this figure) collaborative autoencoders. The reconstruction loss for each instance is computed. Then, the total loss for an instance from the two autoencoders is summed to obtain a batch of loss. DERM aggregation scheme is applied on the loss and back propagation is performed to update parameters in neural networks of both autoencoders.

We can observe from Theorem 1 that the gradient of DERM is a weighted sum of the individual gradient terms w.r.t the instances. Specifically, for each instance x_i, the original gradient $\nabla_\theta f(x_i; \theta)$ is re-weighted proportional to the ratio between the DERM loss $\tilde{R}_{\text{DERM}}(t; \theta)$ and the instance loss $f(x_i; \theta)$. Considering the gradient ratio structure in Theorem 1, the DERM term in the numerator is fixed in a mini-batch, while the denominator $f(x_i; \theta)$ is a strong indicator for the anomalous/normal probability. In particular, a normal instance usually has a small loss $f(x_i; \theta)$, leading to a large re-weighting ratio on the original gradient. On the contrary, an anomalous instance tends to have a large loss, thus resulting in a small ratio on the original gradient. With this mechanism, the influence of the anomalous instances is weakened while the impact of the normal instances are reinforced, which alleviates the overfitting and universal approximation towards the anomalies.

Compared with the existing methods that only change the weights of the instance losses, our method directly re-weights the gradient terms to intervene the optimization process, which works on the parameter space in a more straightforward manner. Moreover, taking both the batch statistics (i.e., $\tilde{R}_{\text{DERM}}(t; \theta)$) and the individual loss $f(x_i; \theta)$ into consideration, DERM can automatically and dynamically modify the gradient weights. This avoids designating manual or ad-hoc weights according to the dataset, thus facilitating the generalization of our method to a diverse range of datasets without many extra adjustments.

B. Comparison with Existing Methods

Studies have been carried out to devise alternatives to the vanilla ERM in various scenarios, such as classification with imbalanced data [23], [28] and learning in the presence of corrupted data [14], [16]. Recently, tilted empirical risk minimization (TERM) [22] framework is proposed to address the sensitivity and poor generalization issues in ERM. TERM

Algorithm 1: DERM for unsupervised AD

Input: training set X_{train}, test set X_{test}, temperature t, learning rate η, max epoch M.

Output: anomaly scores for X_{test}.

1 **Initialize** encoders E_1, E_2 and decoders D_1, D_2 with $\theta = \{\theta_{E_1}, \theta_{E_2}, \theta_{D_1}, \theta_{D_2}\}$.

2 **Training phase:**

3 **for** $epoch \in \{1, 2, \dots M\}$ **do**

4   Sample a mini-batch $\{x_1, \dots, x_N\}$ from X_{train}.

5   Calculate $f(x_i; \theta) = \|x_i - D_1(E_1(x_i))\|_2^2 + \|x_i - D_2(E_2(x_i))\|_2^2$ for each x_i in the sampled mini-batch.

6   Aggregate all losses $\{f(x_i; \theta)\}_N$ with $\tilde{R}_{\text{DERM}}(t; \theta)$ according to Eq. 2.

7   Update $\theta \leftarrow \theta - \eta \nabla_\theta \tilde{R}_{\text{DERM}}(t; \theta)$.

8 **Testing phase:**

9 **for** *each x_i in X_{test}* **do**

10   Calculate anomaly score $s(i) = f(x_i; \theta)$ for x_i.

11 **return** anomaly scores s.

---

takes the following form:

$$\tilde{R}_{\text{TERM}}(t; \theta) := \frac{1}{t} \log \left(\frac{1}{N} \sum_{i \in [N]} e^{tf(x_i; \theta)} \right). \qquad (4)$$

TERM is shown to be effective in robust regression and classification tasks. However, it turns out to be less effective on unsupervised anomaly detection due to its intrinsic property. TERM suffers from the sensitivity to the absolute value of loss, resulting in unsatisfactory outcomes. We analyze the problem as follows, as a comparison to our DERM in Theorem 1. The gradient of TERM [22] is,

$$\nabla_\theta \tilde{R}_{\text{TERM}}(t; \theta) = \sum_{i \in [N]} w_i(t; \theta) \nabla_\theta f(x_i; \theta) \qquad (5)$$

$$\text{with } w_i(t;\theta) = \frac{e^{t(f(x_i;\theta) - \tilde{R}_{\text{TERM}}(t;\theta))}}{N} . \qquad (6)$$

When applying TERM to the unsupervised anomaly detection, we find a major drawback with respect to Eq. 5 (note that when applying TERM to anomaly detection, it requires $t < 0$ [22].). For a min-batch of data $X = \{x_1, x_2, ..., x_N\}$, according to DNN's universal approximation property, the reconstruction loss $\{f(x_i;\theta)\}$ has a high possibility to distribute close to 0. Consequently, the item $|f(x_i;\theta) - \tilde{R}_{\text{TERM}}|$ for a data point x_i is also highly likely to approach 0 since $\tilde{R}_{\text{TERM}}$ is a variant of average w.r.t X. This leads to an unsatisfactory trivial solution that the gradient weight $w_i(t;\theta)$ of each instance in Eq. 5 is close to $\frac{1}{N}$. To amend the issue, TERM heavily relies on the parameter t to scale $|f(x_i;\theta) - \tilde{R}_{\text{TERM}}|$ into a suitable range. Unfortunately, X often changes drastically for different instances even in the same dataset. In realistic applications, it takes efforts to determine the ideal value of t and sometimes the ideal t does not exists. If we want to apply the same algorithm to multiple datasets, the sensitivity of t requires frequently re-training the model and hinders the knowledge transferring across datasets, which restricts its application scenario.

The proposed DERM avoids the aforementioned issue in TERM. Specifically, in DERM, $w_i(t;\theta)$ is linear to $\frac{\tilde{R}_{\text{DERM}}(t;\theta)}{f(x_i;\theta)}$. By contrast, in TERM, $w_i(t;\theta)$ corresponds to e to the power of $(f(x_i;\theta) - \tilde{R}_{\text{TERM}})$. This formulation difference leads to following advantage of our proposed method: the weight $w_i(t;\theta)$ of an instance x_i is not sensitive to the distribution and numerical value of training data since it takes the ratio format instead of the subtraction between $\tilde{R}_{\text{DERM}}$ and $f(x_i;\theta)$. This avoids trivial solution discussed above and assigns discriminative weights to normal and anomalous instances, respectively. Moreover, the form of $w_i(t;\theta)$ has an additional signal magnification effect. Except for the anomalous instances, the normal instances can be dynamically weighted to reflect their various impacts. This way, DERM makes use of the pattern of normal data to emphasize more on high-quality normal data.

To better understand how DERM overcomes the limitations of TERM, we conduct experiments on both synthetic and real-world dataset. For synthetic dataset, we generate reconstruction loss for 200 normal instances from $\mathcal{N}(0.03, 0.002)$ and 10 anomalies from $\mathcal{N}(0.06, 0.005)$ that constitute a set $\{f(x_i;\theta)\}_{i \in [210]}$ to simulate the losses for both normal and anomalous data. For real-world dataset, we adopt the *shuttle* dataset from OODS library [31]. The gradient weights of all instances are shown in Figure 2, from which we can draw several remarks as follows: (1) the gradient weights of TERM approach $\frac{1}{N}$, since the exponential powers in Eq. 5 are close to 0, (2) normal and anomalous instances are better distinguished and separated by the proposed DERM method, (3) except for the normal/anomalous instance separation, DERM also induces dynamic weights on the normal instances. These remarks are consistent with the above analysis and demonstrates our DERM method as a better anomaly detector.
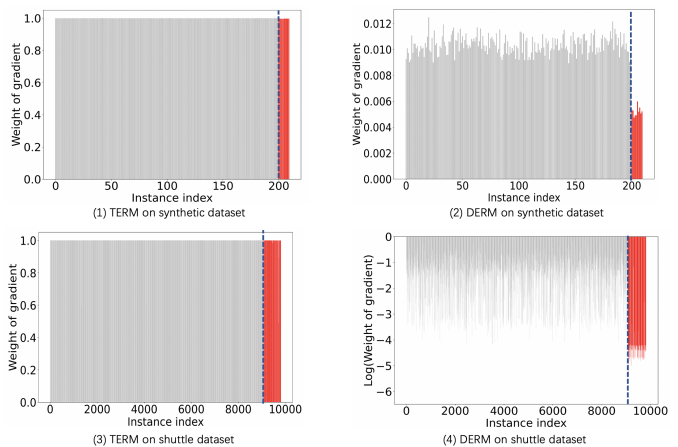

Fig. 2. Gradient comparison between DERM and TERM [22] on real and synthetic datasets. The height of grey bar represents the gradient weights of the normal instances. The height of red bar represents that of anomalous instances. The gradient weight is obtained by $\omega_i / \frac{1}{N}$ for clear visualization. Blue dash line separates the normal and anomalous data. In (4), a base-10 logarithm scale is applied on the y-axis for a better visualization.

C. Collaborative Autoencoders (cAE)

Employing only one autoencoder, the model has the risk of quickly converging to an unsatisfactory solution due to the low reconstruction loss to compute the gradient [45]. The premature convergence of loss function fails to explore the loss surface sufficiently [24]. Meanwhile, ensemble of model outputs has shown a high efficacy in previous studies [1], [10], [25], [47]. To alleviate above issue, inspired by ensemble, we propose the Collaborative Autoencoders (cAE) with different weight initializations for each autoencoder.

Typically, ensemble is a collaboration of a set of models that are trained individually. To enable the end-to-end model training in DERM, we utilize a diagram of optimizing multiple autoencoders in parallel to mimic the traditional ensemble procedure. That is, all autoencoders are optimized by gradient based methods simultaneously and then collaboratively contribute to the subsequent testing phase. After training, each individual autoencoder would have distinct parameters as they are initialized with random values. Subsequently, an one forward pass is performed over the data to obtain multiple reconstruction losses for each test data point. The final loss is a summation of all reconstruction losses, which reduces the chance of model becoming overfitting to certain data points. This collaboration of multiple AEs design endows more robustness and reduce the testing variance, which further enhances the entire performance for detecting anomalies. Collaborative or ensemble structure has also been adopted in a recent work [24], which performs two forward passes of each autoencoder, resulting in an increased computation cost. It also adopts dropouts to simulate ensemble process, which can bring unstability and undermine the performance. In contrast, our proposed cAE structure requires only one forward pass of each autoencoder. There is no dropout in proposed structure, thus avoiding potential negative effect on model capability.

	RCA	VAE	SO_GAAL	RSR	DAGMM	D-SVDD	OCSVM	IF	DERM(Ours)
vowels	0.917±.016	0.503±.045	0.555±.219	0.930±.019	0.340±.103	0.190±.062	0.767±.044	0.765±.031	**0.972**±.011
pima	**0.711**±.016	0.648±.015	0.629±.054	0.660±.050	0.531±.025	0.363±.051	0.633±.016	0.673±.012	0.704±.017
optdigits	0.890±.041	0.909±.016	0.495±.185	0.885±.180	0.290±.042	0.550±.059	0.555±.039	0.726±.049	**0.893**±.036
sensor	0.950±.030	0.913±.003	0.698±.238	0.980±.013	0.924±.085	0.614±.100	0.941±.002	0.949±.009	**0.996**±.001
letter	0.802±.036	0.521±.042	0.414±.094	0.665±.045	0.433±.034	0.465±.073	0.531±.060	0.638±.021	**0.829**±.022
cardio	0.905±.012	**0.944**±.006	0.449±.105	0.937±.009	0.862±.031	0.505±.047	0.932±.008	0.930±.008	0.886±.021
arrhythmia	0.806±.044	0.811±.034	0.558±.077	0.893±.006	0.603±.095	0.635±.065	0.811±.061	0.807±.007	**0.898**±.002
breastw	0.978±.003	0.950±.006	0.985±.011	0.956±.007	0.976±.000	0.406±.047	0.955±.015	**0.987**±.002	0.951±.003
musk	**1.000**±.000	0.944±.002	0.840±.218	0.964±.002	0.903±.130	0.829±.072	**1.000**±.000	0.998±.004	**1.000**±.000
mnist	**0.858**±.012	0.778±.009	0.767±.058	0.754±.065	0.652±.077	0.538±.069	0.820±.012	0.796±.014	0.803±.018
satimage-2	0.977±.008	0.966±.008	0.772±.158	**1.000**±.000	0.853±.113	0.739±.137	0.999±.002	0.993±.001	0.997±.001
satellite	**0.712**±.011	0.538±.016	0.634±.049	0.649±.048	0.667±.189	0.631±.023	0.653±.014	0.702±.021	0.661±.041
mammo	0.844±.014	**0.864**±.014	0.232±.005	0.769±.028	0.834±.000	0.272±.048	0.830±.027	0.862±.010	0.801±.041
thyroid	0.956±.008	0.839±.011	**0.984**±.032	0.940±.023	0.582±.095	0.704±.076	0.893±.026	0.979±.003	0.951±.020
annthyroid	0.688±.016	0.589±.021	0.640±.033	0.646±.024	0.506±.020	0.591±.030	0.597±.023	**0.827**±.011	0.692±.027
ionosphere	0.846±.015	0.763±.015	0.838±.043	0.946±.019	0.467±.082	0.735±.074	0.838±.056	0.853±.006	**0.977**±.016
pendigits	0.856±.011	0.931±.006	0.272±.062	0.884±.057	0.872±.068	0.613±.052	**0.957**±.007	0.950±.015	0.866±.023
shuttle	0.935±.013	0.987±.001	0.715±.310	0.989±.003	0.890±.109	0.531±.260	0.984±.003	**0.997**±.001	0.981±.001
Average	0.868	0.800	0.638	0.859	0.676	0.539	0.808	0.855	**0.881**

IV. EXPERIMENTAL RESULTS

To evaluate the proposed DERM framework, we perform experiments on real-world outlier detection datasets from diverse domains with both continuous and categorical features. We follow the setting in [24] to carry out experiments on 18 datasets from the OODS library [31], on which previous algorithms also perform experiments. Specifically, we split data such that 80% is used for training, and the remaining 20% for testing.

For a single autoencoder, both the encoder and decoder are implemented by multi-layer feedforward neural networks with two hidden layers. The model is optimized with Adam optimizer with a default learning rate of 0.001. The setting for training is consistent with all DNN-based benchmarks. Our method is not sensitive to the parameter t (as will be discussed in Section 4.2 and shown in Fig. 4), we thus empirically set it to 0.01 for all datasets. The commonly-used Area under ROC curve (AUC) score is adopted as the evaluation metric for all methods.

A. Benchmarks and settings

We compare DERM with the following benchmarks:

- RCA [24], which is a recent state-of-the-art robust framework using collaborative autoencoders to jointly identify normal observations from the data while learning its feature representation.
- Variational AutoEncoder (VAE) [17], which is a probabilistic model aiming to learn a Bayesian latent variable model by maximizing the log-likelihood of the training data.
- SO-GAAL [27], which is a novel Single-Objective Generative Adversarial Active Learning method. It directly generates informative potential outliers based on the minimax game between a generator and a discriminator.
- RSR [19], which is a neural network model with a novel robust subspace recovery layer. This layer extracts the underlying subspace from a latent representation of the given data and removes outliers that lie away from this subspace.
- DAGMM [49], which trains a Gaussian Mixture Model to learn the latent representation from the autoencoder to determine anomalies jointly by the reconstruction error and the density estimation.
- Deep-SVDD [32], which learns a neural network transformation from input space to output space. The transformation attempts to map most of the data representations into a hyper-sphere with radius of minimum volume.
- OCSVM [9], which is based on the one-class SVM and fits a tight hyper-sphere in the non-linearly transformed feature space to include most of the data based on the positive examples.
- Isolation Forest (IF) [25], which builds an ensemble of trees for a given dataset. Anomalies are then identified as instances that have short average path lengths on the trees.

Among them, OCSVM and IF are traditional AD detectors, others are recently proposed DNN-based approaches. For all the compared methods, we adopt the optimal settings from their official implementations with minimal modifications so that they can adapt to the datasets from OODS library. To ensure the fair comparison, the same neural network architecture is applied for all DNN-based algorithms. Experiments are repeated for 20 times with random initializations and the average±std are reported.

B. Comparison with the State-of-the-Art

We show the comparison results on 18 datasets in Table I. DERM achieves the best average AUC and has the most number of best-performing datasets (i.e., 7 datasets). This demonstrates the effectiveness of DERM framework for unsupervised anomaly detection. The DNN-based methods such as DAGMM, Deep-SVDD and SO-GAAL fail to perform well when dealing with contaminated data. As analyzed, the reason

|	Autoencoder	TERM+cAE	cAE	DERM+cAE
vowels	0.919±.034	0.950±.022	0.949±.025	**0.972**±.011
pima	**0.704**±.017	0.659±.030	0.672±.029	0.625±.048
optdigits	0.854±.065	**0.927**±.023	0.889±.028	0.893±.036
sensor	0.960±.015	0.975±.026	0.970±.020	**0.996**±.001
letter	**0.829**±.022	0.822±.016	0.796±.040	0.819±.054
cardio	**0.886**±.021	0.867±.025	0.868±.031	0.878±.036
arrhythmia	0.886±.003	0.890±.010	0.885±.002	**0.898**±.002
breastw	0.933±.016	0.942±.007	0.939±.006	**0.951**±.003
musk	0.977±.085	**1.000**±.001	0.989±.022	**1.000**±.000
mnist	0.780±.016	**0.850**±.017	0.781±.027	0.803±.018
satimage-2	0.923±.027	0.942±.027	0.960±.022	**0.997**±.001
satellite	0.657±.021	**0.664**±.015	0.654±.015	0.661±.041
mammo	**0.833**±.033	0.828±.030	0.832±.028	0.801±.041
thyroid	0.907±.038	0.921±.031	0.936±.025	**0.951**±.020
annthyroid	0.656±.023	0.671±.019	0.667±.017	**0.692**±.027
ionosphere	0.969±.008	**0.984**±.002	0.968±.007	0.977±.016
pendigits	0.799±.047	0.803±.040	0.826±.043	**0.866**±.023
shuttle	0.838±.144	0.812±.080	0.813±.059	**0.981**±.001
Average	0.845	0.860	0.855	**0.881**

Fig. 3. Change of average weight of gradients for normal (blue) and anomalous (red) instances in $vowels$ and $pendigits$ dataset w.r.t. training iterations in DERM framework.

is that most DNN-based methods aggressively fit the anomalies and learn inaccurate features from them, which are supposed to be learned from normal data. The ERM scheme widely-used in these methods exacerbates this issue. RCA is also inferior to the proposed method because RCA arbitrarily discards the suspicious instances, which impairs the detection capacity.

C. Ablation Study and Parameter Analysis

In order to validate the effectiveness of the DERM principle, we conduct ablation study on DERM in comparison with AE and TERM, as shown in Table II. All datasets are divided into training and testing data with a ratio of 0.8:0.2. The baseline is an autoencoder with mean square error loss and mean square error anomaly score, which is usually the default choice for autoencoder. The result clearly shows the advanced performance of DERM over TERM and vanilla Autoencoder, validating the effectiveness of the DERM principle and the cAE design.

To better understand the effectiveness of the training dynamics for the proposed DERM principle, we compute and plot the change of average weight of gradients for all normal and anomalous instances in $vowels$ and $pendigits$ datasets w.r.t. training iterations, as shown in Fig. 3. It clearly demonstrates that the weights of anomalies are almost consistently suppressed along the training process, which ensures the stability of optimization.

Fig. 4 shows the sensitivity analysis for mean AUCs among 18 datasets used in Fig. I by varying t from 0.001 to 10 for DERM and -1 to -0.01 for TERM respectively. It can be observed that DERM is generally insensitive to the variation of parameters t compared to TERM. It further validates the low testing variance and high robustness of DERM.

V. CONCLUSION

We propose a Diminishing Empirical Risk Minimization (DERM) framework for unsupervised anomaly detection to
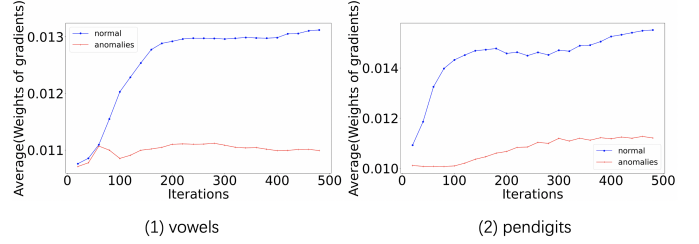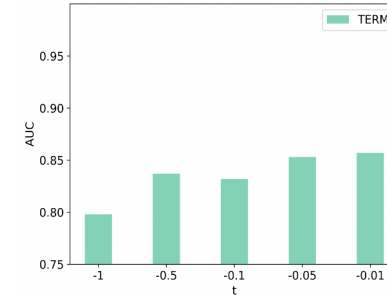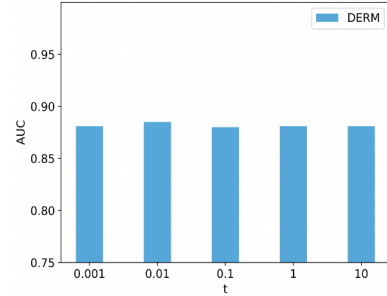

Fig. 4. Mean AUCs among 18 datasets with various t via DERM and TERM respectively. Note that when applying TERM to anomaly detection, it requires $t < 0$. Hence, different ranges of t are chosen for DERM and TERM for the sensitivity analysis. DERM shows robustness with even a larger variation of t.

mitigate the limitation of existing DNN-based methods incurred by the traditional ERM learning principle. DERM is well-devised to adaptively control the weights of gradient for corresponding instances via an innovative loss aggregation scheme. Theoretical analysis demonstrates the effectiveness of DERM in suppressing outliers that contaminate the training data. Experimental results show that our DERM framework achieves wide applicability, high flexibility and improved performance on a variety of real-world benchmarks consisting of 18 datasets from diverse domains. In this work, experiments on DERM mainly depend on the assumption that anomalies have larger reconstruction loss. Future studies could investigate the effectiveness of aggregation frameworks on anomaly detectors that adopt different format of loss.

REFERENCES

[1] Aggarwal, C.C., Sathe, S.: Outlier ensembles: An introduction. Springer (2017)

[2] Aytekin, C., Ni, X., Cricri, F., Aksu, E.: Clustering and unsupervised anomaly detection with l 2 normalized deep auto-encoder representations. In: 2018 International Joint Conference on Neural Networks (IJCNN). pp. 1–6. IEEE (2018)

[3] Borghesi, A., Bartolini, A., Lombardi, M., Milano, M., Benini, L.: Anomaly detection using autoencoders in high performance computing systems. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9428–9433 (2019)

[4] Candès, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? Journal of the ACM (JACM) 58(3), 1–37 (2011)

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

[6] Chalapathy, R., Menon, A.K., Chawla, S.: Robust, deep and inductive anomaly detection. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 36–51. Springer (2017)

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

[8] Chen, L., Cao, J., Chen, H., Liang, W., Tao, H., Zhu, G.: Attentive multi-task learning for group itinerary recommendation. Knowledge and Information Systems 63(7), 1687–1716 (2021)

[9] Chen, Y., Zhou, X.S., Huang, T.S.: One-class svm for learning in image retrieval. In: Proceedings 2001 International Conference on Image Processing. IEEE (2001)

[10] Emmott, A., Das, S., Dietterich, T., Fern, A., Wong, W.K.: A meta-analysis of the anomaly detection problem. arXiv preprint arXiv:1503.01158 (2015)

[11] Gong, D., Liu, L., Le, V., Saha, B., Mansour, M.R., Venkatesh, S., Hengel, A.v.d.: Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: CVPR (2019)

[12] Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural networks 2(5), 359–366 (1989)

[13] Jiang, L., Zhou, Z., Leung, T., Li, L., Fei-Fei, L.M.: Learning data-driven curriculum for very deep neural networks on corrupted labels. arxiv 2017. arXiv preprint arXiv:1712.05055

[14] Jiang, L., Zhou, Z., Leung, T., Li, L.J., Fei-Fei, L.: Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In: International Conference on Machine Learning. pp. 2304–2313. PMLR (2018)

[15] Katharopoulos, A., Fleuret, F.: Biased importance sampling for deep neural network training. arXiv preprint arXiv:1706.00043 (2017)

[16] Khetan, A., Lipton, Z.C., Anandkumar, A.: Learning from noisy singly-labeled data. arXiv preprint arXiv:1712.04577 (2017)

[17] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)

[18] Kumar, M.P., Packer, B., Koller, D.: Self-paced learning for latent variable models. In: NIPS. vol. 1, p. 2 (2010)

[19] Lai, C.H., Zou, D., Lerman, G.: Robust subspace recovery layer for unsupervised anomaly detection. arXiv preprint arXiv:1904.00152 (2019)

[20] Leqi, L., Prasad, A., Ravikumar, P.: On human-aligned risk minimization (2019)

[21] Leung, K., Leckie, C.: Unsupervised anomaly detection in network intrusion detection using clusters. In: Proceedings of the Twenty-eighth Australasian conference on Computer Science-Volume 38. pp. 333–342 (2005)

[22] Li, T., Beirami, A., Sanjabi, M., Smith, V.: Tilted empirical risk minimization. arXiv preprint arXiv:2007.01162 (2020)

[23] Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)

[24] Liu, B., Wang, D., Lin, K., Tan, P.N., Zhou, J.: Rca: A deep collaborative autoencoder approach for anomaly detection. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21) (2021)

[25] Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation forest. In: 2008 eighth ieee international conference on data mining. pp. 413–422. IEEE (2008)

[26] Liu, S., Xue, S., Wu, J., Zhou, C., Yang, J., Li, Z., Cao, J.: Online active learning for drifting data streams. IEEE Transactions on Neural Networks and Learning Systems (2021)

[27] Liu, Y., Li, Z., Zhou, C., Jiang, Y., Sun, J., Wang, M., He, X.: Generative adversarial active learning for unsupervised outlier detection. IEEE Transactions on Knowledge and Data Engineering 32(8), 1517–1528 (2019)

[28] Malisiewicz, T., Gupta, A., Efros, A.A.: Ensemble of exemplar-svms for object detection and beyond. In: 2011 International conference on computer vision. pp. 89–96. IEEE (2011)

[29] Pang, G., Shen, C., Cao, L., Hengel, A.V.D.: Deep learning for anomaly detection: A review. ACM Computing Surveys (CSUR) 54(2), 1–38 (2021)

[30] Pourhabibi, T., Ong, K.L., Kam, B.H., Boo, Y.L.: Fraud detection: A systematic literature review of graph-based anomaly detection approaches. Decision Support Systems 133 (2020)

[31] Rayana, S.: Odds library (2016), http://odds.cs.stonybrook.edu

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

[33] Šabić, E., Keeley, D., Henderson, B., Nannemann, S.: Healthcare and anomaly detection: using machine learning to predict anomalies in heart rate data. AI & SOCIETY pp. 1–10 (2020)

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

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

[36] Shen, Y., Sanghavi, S.: Learning with bad training data via iterative trimmed loss minimization. In: International Conference on Machine Learning. pp. 5739–5748. PMLR (2019)

[37] Singh, K., Upadhyaya, S.: Outlier detection: applications and techniques. International Journal of Computer Science Issues (IJCSI) 9(1), 307 (2012)

[38] Sinha, A., Namkoong, H., Volpi, R., Duchi, J.: Certifying some distributional robustness with principled adversarial training. arXiv preprint arXiv:1710.10571 (2017)

[39] Song, J., Takakura, H., Okabe, Y., Nakao, K.: Toward a more practical unsupervised anomaly detection system. Information Sciences 231, 4–14 (2013)

[40] Wang, C., Wang, B., Liu, H., Qu, H.: Anomaly detection for industrial control system based on autoencoder neural network. Wireless Communications and Mobile Computing (2020)

[41] Wold, S., Esbensen, K., Geladi, P.: Principal component analysis. Chemometrics and intelligent laboratory systems 2(1-3), 37–52 (1987)

[42] Xia, Y., Cao, X., Wen, F., Hua, G., Sun, J.: Learning discriminative reconstructions for unsupervised outlier removal. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1511–1519 (2015)

[43] Xu, H., Chen, W., Zhao, N., Li, Z., Bu, J., Li, Z., Liu, Y., Zhao, Y., Pei, D., Feng, Y., et al.: Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications. In: Proceedings of the 2018 World Wide Web Conference. pp. 187–196 (2018)

[44] Yu, Y., Yang, M., Xu, L., White, M., Schuurmans, D.: Relaxed clipping: A global training method for robust regression and classification. In: NIPS. pp. 2532–2540 (2010)

[45] Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning (still) requires rethinking generalization. Communications of the ACM 64(3), 107–115 (2021)

[46] Zhang, C., Song, D., Chen, Y., Feng, X., Lumezanu, C., Cheng, W., Ni, J., Zong, B., Chen, H., Chawla, N.V.: A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 1409–1416 (2019)

[47] Zhao, Y., Nasrullah, Z., Li, Z.: Pyod: A python toolbox for scalable outlier detection. arXiv preprint arXiv:1901.01588 (2019)

[48] Zhou, C., Paffenroth, R.C.: Anomaly detection with robust deep autoencoders. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 665–674 (2017)

[49] Zong, B., Song, Q., Min, M.R., Cheng, W., Lumezanu, C., Cho, D., Chen, H.: Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In: International conference on learning representations (2018)