ESAD: End-to-end Semi-supervised Anomaly Detection

Chaoqin Huang[1][2]
huangchaoqin@sjtu.edu.cn
Fei Ye[1]
yf3310@sjtu.edu.cn
Peisen Zhao[1][3]
pszhao@sjtu.edu.cn
Ya Zhang ✉[1][2]
ya_zhang@sjtu.edu.cn
Yanfeng Wang[1][2]
wangyanfeng@sjtu.edu.cn
Qi Tian[3]
tian.qi1@huawei.com

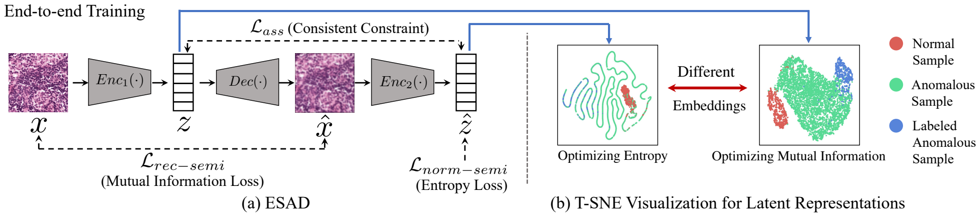
[1] Cooperative Medianet Innovation Center,
Shanghai Jiao Tong University
[2] Shanghai AI Laboratory
[3] Huawei Cloud & AI

Abstract

This paper explores semi-supervised anomaly detection, a more practical setting for anomaly detection where a small additional set of labeled samples are provided. We propose a new KL-divergence based objective function for semi-supervised anomaly detection, and show that two factors: the *mutual information* between the data and latent representations, and the *entropy* of latent representations, constitute an integral objective function for anomaly detection. To resolve the contradiction in simultaneously optimizing the two factors, we propose a novel encoder-decoder-encoder structure, with the first encoder focusing on optimizing the mutual information and the second encoder focusing on optimizing the entropy. The two encoders are enforced to share similar encoding with a consistent constraint on their latent representations. Extensive experiments have revealed that the proposed method significantly outperforms several state-of-the-arts on multiple benchmark datasets, including medical diagnosis and several classic anomaly detection benchmarks.

1 Introduction

Anomaly detection (AD), with broad application in medical diagnosis [45], credit card fraud detection [32], and autonomous driving [11], has received significant attention among the machine learning community. The main challenge in AD is that, it is prohibitive, even if not impossible, to collect a representative set of anomalous samples due to its remarkable scarcity in the population. To bypass the challenge, many approaches [31, 39, 54] have resorted to unsupervised learning so that only normal samples are needed for model training.



(a) ESAD (b) T-SNE Visualization for Latent Representations

Figure 1: The training processes of ESAD for semi-supervised anomaly detection. (a) ESAD leverages an encoder-decoder-encoder structure, where the two encoders are enforced to share similar encoding with a consistent constraint on their latent representations, with the first encoder targeting to optimize the mutual information and the second encoder focusing on the entropy. (b) shows the T-SNE [23] visualization results for the latent representations.

Semi-supervised anomaly detection, where a small set of labeled data are provided for training in addition to a large amount of unlabeled data, represents a more practical setting of anomaly detection. In the real-world scenario, it is feasible to obtain a small set of 'biased' anomalous data. Earlier semi-supervised AD methods follow the unsupervised learning paradigm and employ such a labeled anomalous set through a certain form of regularization [16, 26]. More recently, Deep SAD [37], the first deep semi-supervised AD method, builds upon the *Infomax principle* [4, 17, 22] that maximizes the mutual information between the data and the latent representations and enforces an additional regularization on the latent representations. Due to the contradiction between the mutual information-based objective and entropy-based regularization, named model collapse in [36, 37], Deep SAD adopts a two-stage process: (i) autoencoder pre-training for mutual information maximization; and (ii) encoder fine-tuning for entropy regularization. This sequential learning process cannot guarantee the two objectives are simultaneously optimized and cannot well resolve the contradiction between the mutual information and entropy during the optimization. The model tends to collapse when the entropy is minimized to zero at the second stage, and the model inevitably leads to low mutual information as all data are mapped into a constant [36, 37].

In this paper, we introduce ESAD, an end-to-end method for semi-supervised anomaly detection. We start with exploring an alternative optimization target for AD by maximizing the KL-divergence between the normal and the anomalous class. Considering the challenge in estimating the anomalous distribution which results in the infeasibility of direct optimization, the KL-divergence based objective function is relaxed and further decomposed into two factors: (i) *mutual information* between the data and the latent representations and (ii) *entropy* of latent representations. While the two factors in the final objective function seem to be the same as those of Deep SAD, the difference lies in that, here mutual information and entropy are considered an integral part of the single objective function and need to be optimized simultaneously in an end-to-end training fashion.

In addition, to resolve the contradiction between the mutual information and entropy during optimization, we extend the autoencoder structure widely adopted for deep anomaly detection into an encoder-decoder-encoder structure illustrated in Figure 1 (a), where two separate but closely resembled encoders are employed to emphasize different factors in the optimization so that the model can be trained end-to-end. Specifically, although the two encoders are enforced to share similar encoding via a consistent constraint on the outputs, the first encoder focuses on mutual information through targeting on good representations only for the normal data but not for the labeled anomalous data, while the second encoder focuses on entropy by enforcing the compacted representations for the normal data and scattered representations for the anomalous data. With the encoder-decoder-encoder structure, we achieve end-to-end training for semi-supervised anomaly detection. Figure 1 (b) shows that the two encoders actually result in quite different embeddings, confirming the difficulty in finding a common embedding that simultaneously optimizing both mutual information and entropy. However, embeddings from both encoders show a better separation between the normal and anomalous classes than that of Deep SAD.

To validate the effectiveness of ESAD, we experiment with two medical image datasets [3, 48], three natural image benchmarks [20, 21, 50], and several classic AD benchmarks [35]. Extensive results and analysis have shown that ESAD outperforms state-of-the-art methods on almost all datasets. Ablation studies are conducted to show the effectiveness of the proposed objectives and the encoder-decoder-encoder architecture for ESAD.

Our main contribution is summarized as follows:

- We introduce a KL-divergence based objective for semi-supervised anomaly detection and show that it can be relaxed and decomposed into mutual information and entropy related objectives, which formulates the AD objective with information-theoretical terms.
- To achieve end-to-end training, we propose an encoder-decoder-encoder architecture to simultaneously optimize the two contradictory factors, mutual information and entropy.
- The proposed method outperforms state-of-the-arts on multiple AD benchmarks.

2 Related Works

Unsupervised Anomaly Detection. The vital challenge of unsupervised AD is that the training dataset contains only normal data. One-class classification based approaches tended to depict normal data directly with statistical approaches [10, 33, 36, 42, 51]. Self-supervised based approaches remedied the lack of supervision by introducing different self-supervisions, where the model was trained to optimize a self-supervised task, and then normal data can be separated with the assumption that anomalous data will perform differently. In this domain, reconstruction [1, 2, 8, 14, 27, 39, 40, 41, 49, 56] is the most popular self-supervision. Some approaches introduced other self-supervisions, *e.g.*, [13] applied dozens of image geometric transforms for transformation classification, and [52] proposed a restoration framework to further improve the feature learning.

Semi-supervised Anomaly Detection. Since classical semi-supervised approaches [7, 18, 28, 29, 34] are inappropriate and hardly detect new and unknown anomalies due to the cluster assumption [5], many semi-supervised approaches are still grounded on the unsupervised learning paradigm [16]. Along this line, Deep SSAD [16] has been studied recently in specific contexts such as videos [19], network intrusion detection [25], or specific neural network architectures [9]. Deep SAD [37], a general method based on deep SVDD [36], built upon the Infomax principle, where the training processes are consist of two stages. TLSAD [17] further consolidated the model's discriminative power with a transfer learning framework, which relied on an additional large-scale reference dataset for the model training.

Anomaly Detection on Medical Images is an important application but rarely considered in deep anomaly detection literature. [55] proposed P-Net for anomaly detection in retinal images by leveraging the specific relation between the image texture and the regular structure of retinal images, which is hard to generalize to other medical data. [45] relied on the classical autoencoder approach with a re-designed training pipeline to handle high-resolution, complex images. [53] proposed a confidence-aware anomaly detection model for detecting viral pneumonia with in-house data. In this paper, we conduct experiments on some well-organized and open-source medical image datasets [3, 48].

3 End-to-end Semi-supervised Anomaly Detection

Given the input space $\mathcal{X}$ consisting of normal data $\mathcal{X}_N$ and anomalous data $\mathcal{X}_A$, where $\mathcal{X} = \mathcal{X}_N \cup \mathcal{X}_A$. For semi-supervised anomaly detection (AD), we are given n unlabeled samples $\mathbf{x}_1^u, \cdots, \mathbf{x}_n^u \in \mathcal{X}$ and m labeled samples $(\mathbf{x}_1^l, y_1), \cdots, (\mathbf{x}_m^l, y_m) \in \mathcal{X} \times \mathcal{Y}$ with $\mathcal{Y} = \{-1, 1\}$ where $y = 1$ denotes normal samples and $y = -1$ denotes anomalous samples. We assume $m \ll n$. Suppose the output space is $\mathcal{Z}$, the goal of AD is to find $f_\theta : \mathcal{X} \to \mathcal{Z}$, parameterized by θ, that leads to the maximum distance between normal and anomalous data.

Targeting semi-supervised anomaly detection, we attempt to explore an objective function based on Kullback–Leibler (KL) divergence. Let X and Z be variables sampled from $\mathcal{X}$ and $\mathcal{Z}$, respectively. Denote the joint distribution of data and latent representations for normal and anomalous data as $p_N(X,Z)$ and $p_A(X,Z)$, respectively, and the objective function for semi-supervised AD is then formulated as: $\max_\theta \ \mathrm{KL}\left[p_N(X,Z) \| p_A(X,Z)\right]$. Here $p_N(X,Z)$ can be approximately estimated using the labeled normal samples and the large numbers of unlabeled data, with the widely adopted assumption for AD that almost all unlabeled data are normal [16, 31, 36, 37, 39, 54]. On the contrary, it is impossible to estimate $p_A(X,Z)$ due to the extremely limited labeled instances. We here introduce another distribution, $p_{\tilde{A}}(X,Z)$, and reformulated the objective function as follows:

$$\max_\theta \ \mathrm{KL}\left[p_N(X,Z) \| p_A(X,Z)\right] - \mathrm{KL}\left[p_{\tilde{A}}(X,Z) \| p_A(X,Z)\right], \tag{1}$$

where $p_{\tilde{A}}(X,Z)$ can be estimated by the limited labeled anomalous data. With this objective function, we attempt to simultaneously (i) maximize KL divergence between the normal class and the anomalous class and (ii) minimize the KL divergence between the labeled anomalous class and the real anomalous class. Considering that it is impossible to estimate p_A, we decompose the KL term $\mathrm{KL}\left[p_N(X,Z) \| p_A(X,Z)\right]$ as follows:

$$\begin{aligned} \mathrm{KL}\left[p_N(X,Z) || p_A(X,Z)\right] = \ & \mathbb{E}_{p_N(X,Z)}\left[\log \frac{p_N(X,Z)}{p_A(X,Z)}\right] \\ = \mathbb{E}_{p_N(X,Z)}&\left[\log\left(\frac{p_N(Z|X)}{p_N(Z)} \cdot p_N(Z) \cdot \frac{1}{p_A(Z|X)} \cdot \frac{p_N(X)}{p_A(X)}\right)\right] \\ = I(X_N, Z_N) - H(Z_N) &+ \mathbb{E}_{p_N(X)}\left[H(p_N(Z|X), p_A(Z|X))\right] + \mathrm{KL}\left[p_N(X) || p_A(X)\right], \end{aligned} \tag{2}$$

where $I(\cdot,\cdot)$ is the mutual information, $H(\cdot)$ is the entropy, and $H(\cdot,\cdot)$ is the cross entropy. With the non-negativity of the third and fourth terms (see the supplementary material for the proof), we get a lower bound to Eq. (2): $\mathrm{KL}\left[p_N(X,Z) || p_A(X,Z)\right] \geq I(X_N, Z_N) - H(Z_N)$. Similarly, $\mathrm{KL}\left[p_{\tilde{A}}(X,Z) \ \| p_A(X,Z)\right]$ is approximated with $I(X_{\tilde{A}}, Z_{\tilde{A}}) - H(Z_{\tilde{A}})$. The final objective function is thus formulated as:

$$\max_\theta \ \{[I(X_N, Z_N) - I(X_{\tilde{A}}, Z_{\tilde{A}})] - [H(Z_N) - H(Z_{\tilde{A}})]\}. \tag{3}$$

Note that this objective function is coincidentally similar to that of Deep SAD [37], by optimizing on both the mutual information and entropy. However, the objective function here differs from [37] in that: (i) we start with a KL based formulation and derive equal weights for the mutual information and entropy, while for Deep SAD, the entropy is introduced as regularization with a coefficient β; (2) the mutual information for our paper involves different directions of optimizations for normal and anomalous data, while Deep SAD treats the normal and anomalous data the same in maximizing the mutual information. In our formulation, the optimizations of mutual information and entropy are integral parts of the single anomaly detection objective function and hence need to be optimized simultaneously.

Architecture. We follow [37] and employ an autoencoder to optimize the mutual information $I(X,Z)$. To resolve the contradiction between mutual information and entropy and achieve end-to-end training, different from the straightforward solution by directly introducing two independent encoders [12], we propose to append an additional encoder to the autoencoder and introduce an encoder-decoder-encoder architecture, where the first encoder $Enc_1(\cdot)$ emphasizes mutual information optimization and the second encoder $Enc_2(\cdot)$ focuses on entropy optimization, and in the meanwhile, the two encoders are enforced to share similar encoding via a consistent constraint on their latent representations. The encoder-decoder-encoder architecture can be expressed as:

$$\mathbf{z} = Enc_1(\mathbf{x}),\ \hat{\mathbf{x}} = Dec(\mathbf{z}),\ \hat{\mathbf{z}} = Enc_2(\hat{\mathbf{x}}), \tag{4}$$

where $\hat{\mathbf{x}}$ is the output of the decoder, and $\mathbf{z}$ and $\hat{\mathbf{z}}$ are the latent representations from the first and second encoders, respectively. The wights for the two encoders are are not shared.

Losses. To optimize the two factors, *i.e.*, mutual information and entropy, in Eq. (3), we propose the corresponding losses as follows.

The optimization of mutual information is achieved with reconstruction or restoration [47]. With unlabeled samples $\mathbf{x}_1^u, \cdots, \mathbf{x}_n^u$ and labeled samples $\mathbf{x}_1^l, \cdots, \mathbf{x}_m^l$, we want the autoencoder to well reconstruct the normal data but erroneously reconstruct the labeled anomalous data, thus the reconstruction likelihood is maximized for the normal data and minimized for the labeled anomalous data. A straight-forward loss definition for the anomalous data is the negative squared norm loss. However, due to its unbounded nature, it is expected to result in an ill-posed optimization problem and caused optimization to diverge [37]. We therefore introduce a transformation function $\phi(\cdot)$ on the input, forcing the network to reconstruct the anomalous data $\mathbf{x}$ to its transformation $\phi(\mathbf{x})$, where $\phi(\mathbf{x}) \neq \mathbf{x}, \forall \mathbf{x} \in \mathcal{X}_A$. The transformation makes the network unable to correctly reconstruct the anomalous samples. The reconstruction loss is defined as follows:

$$\mathcal{L}_{rec-semi} = \frac{1}{n}\sum_{i=1}^{n}\|\hat{\mathbf{x}}_i^u - \mathbf{x}_i^u\|^2 + \frac{1}{m}\sum_{j=1}^{m}\|\hat{\mathbf{x}}_j^l - \Phi(\mathbf{x}_j^l)\|^2, \tag{5}$$

where $\Phi(\mathbf{x}_j^l) = \begin{cases} \mathbf{x}_j^l, & \text{if } y_j = 1, \\ \phi(\mathbf{x}_j^l), & \text{if } y_j = -1. \end{cases}$ For the data which is functioned as a vector, $\phi(\mathbf{x}_j^l)$ can be a version adding Gaussian noise or a random permutation between various dimensions; for the image data, it can be a noised and rotated version of the original images. Besides the proposed transformation function, we also try another strategy, which enforces the model to reconstruct the labeled anomalous data to the normal data [31]. But this task is too strict and difficult, especially for two types of samples that are quite different, which makes the model hard to converge.

For the entropy $H(Z)$, assuming Z follows an isotropic Gaussian [6], $Z \sim N(\mu, \sigma^2 I)$ with $\sigma > 0$, the entropy of Z is proportional to its log-variance, *i.e.*, $H(Z) \propto \log \sigma^2$ (see the supplementary material for the proof). In this case, for $\mathbf{z} \sim p(Z)$, an L_2 norm can be used for the optimization of the entropy, since it minimizes the empirical variance and thus minimizes the entropy of a latent Gaussian.

$$\mathcal{L}_{norm-semi} = \frac{1}{n}\sum_{i=1}^{n}\|\hat{\mathbf{z}}_i^u\|_2 + \frac{1}{m}\sum_{j=1}^{m}(\|\hat{\mathbf{z}}_j^l\|_2)^{y_j}, \tag{6}$$

where $y_j = -1$ for the labeled anomalous data while $y_j = 1$ for the labeled normal data. This loss enforces the compacted representation for the normal data and scattered representation for the labeled anomalous data. Note that the inverse squared norm loss used for labeled anomalous data here is bounded from below and smooth, which are crucial properties for losses used in deep learning [15]. Compared with the SVDD loss in [37] where a pre-training process is necessary for initializing an additional hypersphere center, $\mathcal{L}_{norm-semi}$ does not need the pre-training, which indicates that the end-to-end training can be achieved.

To define the consistency between the two encoders, similar to the assistant loss [1], we resort to a consistent constraint between their corresponding latent representations:

$$\mathcal{L}_{ass} = \frac{1}{n+m} \sum_{i=1}^{n+m} ||\hat{\mathbf{z}}_i - \mathbf{z}_i||^2. \tag{7}$$

Finally, we define our training loss as follow:

$$\mathcal{L}_{semi} = \mathcal{L}_{rec-semi} + \lambda_1 \mathcal{L}_{norm-semi} + \lambda_2 \mathcal{L}_{ass}, \tag{8}$$

where λ_1 and λ_2 are two hyperparameters. We will further discuss the impacts of these two hyperparameters in the experiment section. To this end, we achieve end-to-end training for semi-supervised anomaly detection.

Anomaly Score Measurement. We discuss how we calculate the anomaly score in the test phase. Since both the mutual information and the entropy are related to the performance of anomaly detection, we use both $\mathcal{L}_{rec-semi}$ and $\mathcal{L}_{norm-semi}$ to measure the anomaly score for the given samples, which are related to the mutual information and the entropy, respectively. We calculate the reconstruction error of each input sample $\mathbf{x}$ and the value of L_2 norm for its representation $\hat{\mathbf{z}}$ for anomaly detection. The anomaly score is formulated as:

$$\mathcal{S}_{test} = ||\hat{\mathbf{x}} - \mathbf{x}||^2 + \lambda_1 ||\hat{\mathbf{z}}||_2, \tag{9}$$

where λ_1 is the same as the setting in the training process. We will further discuss the impact of λ_1 in Section 4.5. To the best of our knowledge, it is the first time considering both the terms of the mutual information and the entropy for the anomaly score measurement. On the contrary, most one-class classification based methods, *e.g.*, OC-SVM [42], only consider the term of the entropy. Similarly, Deep SVDD [36], Deep SAD [37] and TLSAD [12] also consider only the term of the entropy, since they only use the SVDD loss as the final anomaly score. Most reconstruction based methods or restoration based methods, including the vanilla AE [24] and ARNet [52], only consider the term of mutual information, since they only use the reconstruction or restoration loss as the anomaly score. Results show that considering both of the two terms significantly improves the performance of anomaly detection.

4 Experiments

In this section, we conduct substantial experiments to validate our method. The ESAD is first evaluated on multiple AD benchmark datasets, comparing with several state-of-the-arts. Then we present the respective effects of different designs through ablation study. Finally, we visualize the latent representations of ESAD through T-SNE.

4.1 Experimental Setups

Datasets. We conduct semi-supervised anomaly detection experiments on three popular natural image datasets MNIST [21], Fashion-MNIST [50] and CIFAR-10 [20], together with six non-image classic AD datasets [35], all following the settings in [37]. To validate our method on real-world AD scenarios, *i.e.*, with higher resolution and with more complex anomalies, we additionally conduct experiments on two medical image datasets Camelyon16 [3] and the NIH dataset [48]. For all datasets, the training and testing partitions remain as default. More details are shown in the supplementary material.

Table 1: Results of anomaly detection on natural image datasets, where we increase the ratio of labeled anomalies γ_l in the training set. We report the avg. AUC in % with st. dev. computed over 90 experiments at various γ_l. Results of SSAD Hybrid, SS-DGM and Deep SAD are borrowed from [37]. Results of TLSAD are borrowed from [12].

Data	γ_l	SSAD Hybrid [16]	SS-DGM [18]	Deep SAD [37]	TLSAD [12]	ESAD (ours)
MNIST	.00	96.3 ± 2.5	-	92.8 ± 4.9	-	**98.5** ± **1.3**
	.01	96.8 ± 2.3	89.9 ± 9.2	96.4 ± 2.7	94.1	**99.2** ± **0.7**
	.05	97.4 ± 2.0	92.2 ± 5.6	96.7 ± 2.4	96.9	**99.4** ± **0.3**
	.10	97.6 ± 1.7	91.6 ± 5.5	96.9 ± 2.3	97.7	**99.5** ± **0.4**
	.20	97.8 ± 1.5	91.2 ± 5.6	96.9 ± 2.4	98.3	**99.6** ± **0.3**
F-MNIST	.00	91.2 ± 4.7		89.2 ± 6.2	-	**94.0** ± **4.5**
	.01	89.4 ± 6.0	65.1 ± 16.3	90.0 ± 6.4	88.4	**95.3** ± **4.2**
	.05	90.5 ± 5.9	71.4 ± 12.7	90.5 ± 6.5	91.4	**95.6** ± **4.1**
	.10	91.0 ± 5.6	72.9 ± 12.2	91.3 ± 6.0	92.0	**95.8** ± **4.0**
	.20	89.7 ± 6.6	74.7 ± 13.5	91.0 ± 5.5	93.2	**95.9** ± **4.0**
CIFAR-10	.00	63.8 ± 9.0		60.9 ± 9.4	-	**78.8** ± **6.5**
	.01	70.5 ± 8.3	49.7 ± 1.7	72.6 ± 7.4	74.4	**83.7** ± **6.4**
	.05	73.3 ± 8.4	50.8 ± 4.7	77.9 ± 7.2	80.0	**86.9** ± **6.8**
	.10	74.0 ± 8.1	52.0 ± 5.5	79.8 ± 7.1	84.8	**87.8** ± **6.7**
	.20	74.5 ± 8.0	53.2 ± 6.7	81.9 ± 7.0	86.3	**88.5** ± **6.9**

Evaluation Protocol. We quantify the model performance using the area under the Receiver Operating Characteristic (ROC) curve metric (AUC). It is commonly adopted as performance measurement in anomaly detection (AD) tasks.

Model Configuration. For ESAD, the architecture of the autoencoder and the data preprocessing for the image dataset is aligned with [52]. Different from Deep SAD, which uses different networks for each dataset, ESAD uses the same autoencoder network since it is robust enough. For non-image classic AD datasets, we use the autoencoder network aligned with [37]. The hyperparameter λ_1 and λ_2 are set to 1 as default. We give the full details in the supplementary material.

4.2 Experiments on Natural Images

Competing Methods. We consider several semi-supervised anomaly detection state-of-the-arts, including SSAD [16], SS-DGM [18], Deep SAD [37] and TLSAD [12] as baselines. Following [37], as [16] is sensitive to hyperparameters, SSAD Hybrid here uses a subset (10%) of the test set for hyperparameter selection to establish a strong baseline. More details for these baseline methods are shown in the supplementary material.

Experiment Settings. For a dataset with C classes, we conduct a batch of C experiments respectively with each of the C classes set as the normal class once. We then evaluate performance on an independent test set, which contains samples from all classes, including normal and anomalous data.

Comparison with State-of-the-art Methods. The effectiveness of adding labeled anomalies during training is investigated. By adding labeled anomalous samples $\mathbf{x}_1, \ldots, \mathbf{x}_m$ to the training set, we increase the ratio of labeled training data $\gamma_l = m/(n+m)$. The number of anomaly classes included in the labeled training data is set as 1, *i.e.*, there are still eight unseen classes at testing time. We iterate this training set generation process and report the average results over the ten kinds of normal classes × nine labeled anomalous classes, *i.e.*, over 90 experiments per labeled ratio γ_l. The corresponding results are shown in Table 1. On all involved datasets, results present that the average AUCs of ESAD outperform all other methods, including TLSAD which utilizes a large-scale additional dataset (ImageNet [38]) as the reference data for the model training. Results when $\gamma_l > 0$ are much better than the

Table 2: Performance of anomaly detection methods on medical image datasets. We report the avg. AUC in % with st. dev. computed over 3 runs.

Method	Cam.16	NIH (a sub.)	NIH (PA)	NIH (AP)
DAOL [43]	-	80.5 ± 2.1	-	-
DGEO [13]	45.9 ± 2.1	85.3 ± 1.0	63.6 ± 0.6	54.4 ± 0.6
PIAD [44]	89.5 ± 0.6	87.3 ± 0.9	68.7 ± 0.5	58.6 ± 0.3
DIF [30]	90.6 ± 0.3	85.3 ± 0.4	47.2 ± 0.4	56.1 ± 0.2
Deep SAD [37]	92.1 ± 0.4	90.9 ± 0.2	51.9 ± 0.8	59.8 ± 0.1
DPA [45]	93.4 ± 0.3	92.6 ± 0.2	**70.8 ± 0.1**	58.5 ± 0.0
ESAD (ours)	**96.8 ± 0.4**	**94.6 ± 0.4**	68.9 ± 0.2	**60.1 ± 0.2**

Table 3: Results on classic anomaly detection benchmark datasets with a ratio of labeled anomalies of $\gamma_l = 0.01$. We report the avg. AUC in % with st. dev. computed over 10 seeds.

Data	Deep SVDD [36]	SSAD Hybrid [16]	SS-DGM [18]	Deep SAD [37]	ESAD (ours)
arrhythmia	74.6 ± 9.0	78.3 ± 5.1	50.3 ± 9.8	75.9 ± 8.7	**85.2 ± 2.9**
cardio	84.8 ± 3.6	86.3 ± 5.8	66.2 ± 14.3	95.0 ± 1.6	**98.8 ± 0.5**
satellite	79.8 ± 4.1	86.9 ± 2.8	57.4 ± 6.4	91.5 ± 1.1	**92.5 ± 0.7**
satimage-2	98.3 ± 1.4	96.8 ± 2.1	99.2 ± 0.6	**99.9 ± 0.1**	**99.9 ± 0.1**
shuttle	86.3 ± 7.5	97.7 ± 1.0	97.9 ± 0.3	98.4 ± 0.9	**99.1 ± 1.1**
thyroid	72.0 ± 9.7	95.3 ± 3.1	72.7 ± 12.0	98.6 ± 0.9	**99.6 ± 0.2**

results when $\gamma_l = 0$, showing the effectiveness of the semi-supervised training scheme.

4.3 Experiments on Medical Images

Medical images, such as H&E stained images, X-ray, etc., have extremely high resolution compared to natural images. In addition, the patient's lesions may only occupy a small part of the entire image, which brings great challenges to AD. To validate the AD performance of ESAD on real-world AD scenarios, we examined two challenging medical problems with different image characteristics and abnormality appearance, *i.e.*, Camelyon16 [3] and chest X-rays in NIH [48]. We consider several state-of-the-arts, including DAOL [43], DGEO [13], PIAD [44], DIF [30] Deep SAD [37], and DPA [45]. Note that for [13, 30, 44], anomalous samples in the training set are used for the validation. We re-train Deep SAD [37] with the same encoder and decoder network as ESAD to obtain a better baseline.

Anomaly Detection on Chest X-Rays. NIH images without any disease marker were considered normal. Pulmonary and cardiac abnormalities in this dataset are all considered anomalous. Following [43, 45], we split the dataset into two sub-datasets having only posteroanterior (PA) or anteroposterior (AP) projections. The labeled anomalous samples contain only the most frequent disease ('Infiltration') out of fourteen possibilities and there are still thirteen unseen possibilities of anomalies in the test set. We also experiment on a subset containing clearer normal/anomalous cases [43]. Default preprocessing of chest X-rays involved a 768 × 768 central crop and resize to 64 × 64. As shown in Table 2, the anomaly detection performance of ESAD outperforms all state-of-the-arts on the clearer subset [43] and AP subset. DPA [45] performs better than ESAD on the subset of PA. Note that DPA uses a higher resolution version of the images (256 × 256) for validation, so it has a greater advantage compared with other methods.

Metastases Detection in Digital Pathology. For the Camelyon16 Challenge [3], we sample the Vahadane-normalized [46] 64 × 64 tiles from the fully normal slides with magnification of 10×, and treat these as normal. Tiles with metastases are treated as anomalous. It contains 7612 normal and 200 anomalous training images, and 4000 (normal) + 817 (anomalous)

Table 4: Ablation study on different designs of architecture and loss functions for ESAD on two different datasets. We report the avg. AUC in % with st. dev. computed over 10 seeds.

Architecture	Loss Functions			Dataset	
	$\mathcal{L}_{ass}$	$\mathcal{L}_{rec-semi}$			
Encoder-decoder-encoder		Gaussian	Permutation	satellite	cardio
✗	✗	✗	✗	87.9 ± 1.7	96.5 ± 1.1
✓	✗	✗	✗	90.0 ± 1.2	97.2 ± 0.9
✓	✓	✗	✗	90.4 ± 1.1	97.9 ± 1.0
✓	✓	✓	✗	92.0 ± 1.1	98.2 ± 0.6
✓	✓	✗	✓	92.5 ± 1.0	98.6 ± 0.6
✓	✓	✓	✓	92.5 ± 0.7	98.8 ± 0.5

Table 5: Ablation study on shallow and deep networks, for both the encoder and the decoder. 'Shallow' is a LeNet-type network utilized in [37]. 'Deep' is the network utilized in ESAD. We report the avg. AUC in % with st. dev. computed over 90 experiments at various γ_l on F-MNIST. Results with * are lower than expected because of the model collapse problem for Deep SAD under the small labeled anomalies ratio.

Network	Method	$\gamma_l = 0.0$	$\gamma_l = 0.01$	$\gamma_l = 0.05$	$\gamma_l = 0.1$	$\gamma_l = 0.2$
Shallow	Deep SAD [37]	89.2 ± 6.2	90.0 ± 6.4	90.5 ± 6.5	91.3 ± 6.0	91.0 ± 5.5
	ESAD (ours)	93.6 ± 4.5	94.9 ± 4.2	95.3 ± 4.2	95.4 ± 4.1	95.5 ± 4.1
Deep	Deep SAD [37]	72.5 ± 7.0*	87.0 ± 8.7*	90.3 ± 6.4	91.8 ± 7.7	92.0 ± 7.0
	ESAD (ours)	94.0 ± 4.5	95.3 ± 4.2	95.6 ± 4.1	95.8 ± 4.0	95.9 ± 4.0

images for the test. As shown in Table 2, the anomaly detection performance of ESAD outperforms all state-of-the-art methods.

4.4 Experiments on Classic Anomaly Detection Benchmark Datasets

We examine the performance of the various methods on well-established classic AD benchmark datasets [35] with $\gamma_l = 0.01$. Networks of both the encoder and the decoder are aligned with [37]. The corresponding results are shown in Table 3. Comparing with other state-of-the-arts, ESAD shows the highest AUCs and stability. It shows that unlike other deep approaches [8, 9, 13, 19, 25], ESAD is not domain or data-type specific.

4.5 Ablation Study

The model architecture and different losses for ESAD are discussed in Table 4 through ablation studies. Experiments are conducted on two datasets, *i.e.*, cardio and satellite. Firstly, for the model architecture, results show that without the encoder-decoder-encoder architecture, ESAD with vanilla autoencoder shows relatively low and unstable AUCs (the entropy loss is conducted on the first encoder in this case). Secondly, ablation studies on two proposed losses, *i.e.*, $\mathcal{L}_{ass}$ and $\mathcal{L}_{rec-semi}$, show impressive results. Comparing with vanilla reconstruction loss, $\mathcal{L}_{rec-semi}$ utilizes two transformations for changing the supervisions of labeled anomalous data. Without these transformations, it degrades to the vanilla reconstruction loss where the original data are used as the reconstruction supervisions, leading to relatively lower AUCs. Note that the entropy loss should always be utilized in all experiments since it is highly relative to the anomaly score measurement, but its importance can be shown towards the following discussions for the hyperparameters.

Then we analyze the influence of the network choices. For the natural image datasets, Deep SAD [37] uses different LeNet-type networks for each dataset. ESAD does not fol-

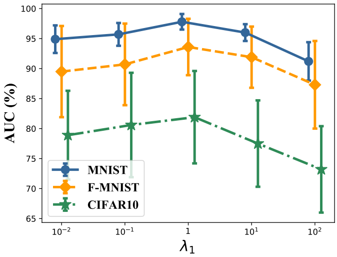
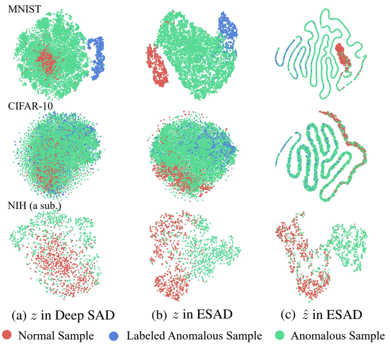


Figure 2: ESAD sensitivity analysis w.r.t. λ_1. We report avg. AUC in % with st. dev. over 90 experiments. Best viewed in color.



(a) z in Deep SAD (b) z in ESAD (c) $\hat{z}$ in ESAD

Figure 3: T-SNE visualization of latent representations on MNIST (top), CIFAR-10 (middle) and NIH (bottom).

low [37] and uses the same network for all datasets. To show the influence of network choices, as shown in Table 5, we experiment with different networks and show that: i) ESAD is robust to different networks, with a performance gap of 0.3% - 0.4% between using the shallow or deep network on F-MNIST, while Deep SAD encounters model collapse with certain networks. ii) ESAD outperforms Deep SAD for both shallow and deep networks. Results on more datasets are shown in the supplemental material.

We further analyze the sensitivity of the hyperparameters of ESAD. According to Eq. (8), λ_1 has a certain impact on the performance of semi-supervised AD. The larger λ_1 means more attention is paid to the entropy, while the smaller λ_1 pays more attention to the mutual information. Figure 2 shows the ESAD performance with different λ_1. The results show that the best AUC can be obtained when λ_1 is set as 1 in all datasets. When λ_1 is relatively too small or too large, relatively poor AD performance will be achieved. Fortunately, the relationship between AUCs and the λ_1 presents the same pattern in all datasets, which means that when changing datasets, we may not need to spend too many resources on the adjustment of λ_1. For λ_2, we found through experiments that modifying λ_2 has a relatively small impact. We thus always set λ_2 as 1. More details are shown in the supplementary material.

4.6 Visualization Analysis

We show that the latent representations extracted by ESAD can better be used to distinguish samples of different categories through T-SNE [23] analysis. We conduct experiments on MNIST, CIFAR-10 and the medical image dataset NIH. Figure 3 (a) shows the results using latent representations extracted by Deep SAD. Figure 3 (b) and (c) visualize different latent representations, *i.e.*, z and $\hat{z}$, extracted by ESAD, which are more discriminative than the baseline. In Figure 3 (c), $\hat{z}$ shows a more specific structure. It shows that the two latent representations have learned different information.

5 Conclusion

In this paper, we show that factors of *mutual information* and *entropy* constitute an integral objective function for anomaly detection. We achieve end-to-end training by proposing a novel model architecture. The proposed information theoretic framework can also be applied to more semi-supervised tasks, opening avenues for future research.

Acknowledgements. This work is supported by the National Key Research and Development Program of China (No. 2019YFB1804304), SHEITC (No. 2018-RGZN-02046), 111 plan (No. BP0719010), and STCSM (No. 18DZ2270700), and State Key Laboratory of UHD Video and Audio Production and Presentation.

References

[1] Samet Akçay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *ACCV*, 2018.

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

[3] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 2017.

[4] Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 1995.

[5] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning. *IEEE Transactions on Neural Networks*, 2009.

[6] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[7] Zihang Dai, Zhilin Yang, Fan Yang, William W Cohen, and Russ R Salakhutdinov. Good semi-supervised learning that requires a bad gan. In *NeurIPS*, 2017.

[8] Lucas Deecke, Robert Vandermeulen, Lukas Ruff, Stephan Mandt, and Marius Kloft. Image anomaly detection with generative adversarial networks. In *Joint european conference on machine learning and knowledge discovery in databases*, 2018.

[9] Tolga Ergen, Ali Mirza, and Suleyman Kozat. Unsupervised and semi-supervised anomaly detection with lstm neural networks. *TNNLS*, 2017.

[10] Eleazar Eskin. Anomaly detection over noisy data using learned probability distributions. In *ICML*, 2000.

[11] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, et al. Robust physical-world attacks on deep learning visual classification. In *CVPR*, 2018.

[12] Zhe Feng, Jie Tang, Yishun Dou, and Gangshan Wu. Learning discriminative features for semi-supervised anomaly detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[13] Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. In *NeurIPS*, 2018.

[14] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *ICCV*, 2019.

[15] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[16] Nico Görnitz, Marius Kloft, Konrad Rieck, and Ulf Brefeld. Toward supervised anomaly detection. *Journal of Artificial Intelligence Research*, 2013.

[17] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019.

[18] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *NeurIPS*, 2014.

[19] B Ravi Kiran, Dilip Mathew Thomas, and Ranjith Parakkal. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *Journal of Imaging*, 2018.

[20] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[21] Yann LeCun. The mnist database of handwritten digits. http://yann.lecun.com/exdb/mnist/, 1998.

[22] Ralph Linsker. Self-organization in a perceptual network. *Computer*, 1988.

[23] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 2008.

[24] Jonathan Masci, Ueli Meier, Dan Cireşan, and Jürgen Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In *International Conference on Artificial Neural Networks (ICANN)*, 2011.

[25] 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*, 2018.

[26] Jordi Mũnoz-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*, 2010.

[27] Miguel Nicolau, James McDermott, and Van Loi Cao. A hybrid autoencoder and density estimation model for anomaly detection. In *International Conference on Parallel Problem Solving from Nature*. Springer, 2016.

[28] Augustus Odena. Semi-supervised learning with generative adversarial networks. In *Workshop on Data-Efficient Machine Learning (ICML)*, 2016.

[29] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *NeurIPS*, 2018.

[30] Khalil Ouardini, Huijuan Yang, Balagopal Unnikrishnan, Manon Romain, Camille Garcin, Houssam Zenati, et al. Towards practical unsupervised anomaly detection on retinal images. In *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*. Springer, 2019.

[31] Pramuditha Perera, Ramesh Nallapati, and Bing Xiang. Ocgan: One-class novelty detection using gans with constrained latent representations. In *CVPR*, 2019.

[32] Clifton Phua, Vincent Lee, Kate Smith, and Ross Gayler. A comprehensive survey of data mining-based fraud detection research. In *International Conference on Intelligent Computation Technology and Automation*, 2010.

[33] Mostafa Rahmani and George K. Atia. Coherence pursuit: Fast, simple, and robust principal component analysis. *IEEE Transactions on Signal Processing*, 2017.

[34] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *NeurIPS*, 2015.

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

[36] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *ICML*, 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] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

[39] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. Adversarially learned one-class classifier for novelty detection. In *CVPR*, 2018.

[40] Mayu Sakurada and Takehisa Yairi. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Mlsda Workshop on Machine Learning for Sensory Data Analysis*, 2014.

[41] 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*, 2017.

[42] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 2001.

[43] Yu-Xing Tang, You-Bao Tang, Mei Han, Jing Xiao, and Ronald M Summers. Deep adversarial one-class learning for normal and abnormal chest radiograph classification. In *Medical Imaging 2019: Computer-Aided Diagnosis*, 2019.

[44] Nina Tuluptceva, Bart Bakker, Irina Fedulova, and Anton Konushin. Perceptual image anomaly detection. In *Asian Conference on Pattern Recognition*, 2019.

[45] Nina Tuluptceva, Bart Bakker, Irina Fedulova, Heinrich Schulz, and Dmitry V Dylov. Anomaly detection with deep perceptual autoencoders. *arXiv preprint arXiv:2006.13265*, 2020.

[46] Abhishek Vahadane, Tingying Peng, Amit Sethi, Shadi Albarqouni, Lichao Wang, Maximilian Baust, Katja Steiger, Anna Melissa Schlitter, Irene Esposito, and Nassir Navab. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE transactions on medical imaging*, 2016.

[47] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, 2008.

[48] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, 2017.

[49] Yan Xia, Xudong Cao, Fang Wen, Gang Hua, and Jian Sun. Learning discriminative reconstructions for unsupervised outlier removal. In *ICCV*, 2015.

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

[51] Huan Xu, C Caramanis, and S Sanghavi. Robust pca via outlier pursuit. *IEEE Transactions on Information Theory*, 2012.

[52] Fei Ye, Chaoqin Huang, Jinkun Cao, Maosen Li, Ya Zhang, and Cewu Lu. Attribute restoration framework for anomaly detection. *IEEE Transactions on Multimedia*, 2020.

[53] Jianpeng Zhang, Yutong Xie, Zhibin Liao, Guansong Pang, Johan Verjans, Wenxin Li, Zongji Sun, Jian He, and Chunhua Shen Yi Li. Viral pneumonia screening on chest x-ray images using confidence-aware anomaly detection. *IEEE transactions on medical imaging*, 2021.

[54] Zhiwei Zhang, Shifeng Chen, and Lei Sun. P-kdgan: Progressive knowledge distillation with gans for one-class novelty detection. In *IJCAI*, 2020.

[55] Kang Zhou, Yuting Xiao, Jianlong Yang, Jun Cheng, Wen Liu, Weixin Luo, et al. Encoding structure-texture relation with p-net for anomaly detection in retinal images. In *ECCV*, 2020.

[56] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, et al. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *ICLR*, 2018.