Elsa: Energy-based Learning for Semi-supervised Anomaly Detection

Sungwon Han*[1,2]
lion4152@gmail.com

Hyeonho Song*[1,2]
hyun78.song@gmail.com

Seungeon Lee[1,2]
marinearchon159@gmail.com

Sungwon Park[1,2]
deu30303@gmail.com

Meeyoung Cha[2,1]
meeyoungcha@ibs.re.kr

[1] School of Computing
Korean Advanced Institute of Science and Technology
Daejeon, South Korea

[2] Data Science Group
Institute of Basic Science
Daejeon, South Korea

Abstract

Contrastive learning has brought important advances in improving anomaly detection. Yet these techniques rely on clean training data, which cannot be guaranteed in real-world scenarios. This paper presents a theoretical interpretation of when and how contrastive learning alone fails to detect anomalies under data contamination. To address the shortcomings, we propose Elsa, a novel semi-supervised anomaly detection approach, that unifies the concept of energy-based models with unsupervised contrastive learning. Elsa instills robustness against various practical scenarios by a carefully designed fine-tuning step that uses the energy function to divide the normal data into prototype classes or subclasses that reflect heterogeneity of the data distribution. By using a small set of anomaly labels, Elsa improves anomaly detection performance in both clean and contaminated data scenarios by 0.9 and 6.6 AUROC, respectively.

1 Introduction

Anomaly detection [4], also known as novelty detection [31], identifies out-of-distribution (OOD) instances from the predominant normal data. Conventional detection approaches model the probability distribution $p(\mathbf{x})$ of the normal data as *normality score* implicitly or explicitly and identify deviant input with a small normality score. Various models are used for estimating $p(\mathbf{x})$, including generative adversarial networks [35, 43], autoencoders [1, 7], one-class classifiers [33, 34], and discriminative models with surrogate tasks [41].

Among them is CSI [40], the latest novelty detection method based on contrastive learning. CSI treats augmented input as positive samples and the distributionally-shifted input as negative samples, which leads to a substantial performance gain. Yet, it shares a common limitation with extant methods in that the model assumes clean training data and fails to

*These authors contributed equally to this work.

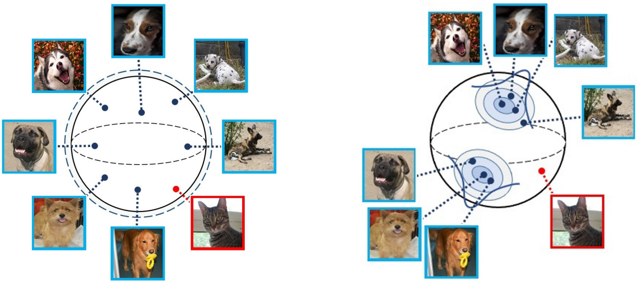


Figure 1: Embedding of CSI is a uniformly distributed hypersphere where anomalies (marked red) are hard to detect (left). Energy-based fine-tuning of Elsa embeds similar images nearby, leaving room for anomalies to be separated (right).

learn $p(\mathbf{x})$ when the data contains unknown anomalies as in various real-world scenarios. This limitation occurs because CSI uses a hypersphere embedding space that is uniformly distributed [5, 42]. The uniformity makes it challenging to distinguish OOD samples. Fig. 1 demonstrates this limitation.

We present Elsa (Energy based learning for semi-supervised anomaly detection), an anomaly detection method that unifies contrastive learning and energy-based functions. Elsa benefits from the high representation power of unsupervised contrastive learning via its pre-training step, which can accommodate existing algorithms [6, 13, 15, 29]. It applies a carefully designed energy function over the pre-trained embedding to learn the probability distribution $p(\mathbf{x})$ of normal data, with the help of a small set of labels that indicate whether given samples are normal or OOD. The energy-based fine-tuning step embeds similar samples nearby and can distinguish OOD samples from the mostly normal data via the energy score: low-energy corresponds to compatible data distribution (e.g., dog images in Fig. 1), and high-energy represents incompatibility (e.g., a cat image in Fig. 1).

Elsa's energy function does not require any explicit density estimator. Instead, it transforms an unsupervised contrastive problem into a non-parametric classification task by introducing the concept of *prototypes*, where a prototype vector functions as a subclass that reflects the heterogeneity of the normal data distribution. The energy score is computed as logits from the discriminative classifier, which denote the cosine similarity between a data instance and prototypes. Elsa is structurally different from other energy-based models that require knowledge of the ground-truth class information [24]. Furthermore, Elsa's training is stable compared to other works that utilized energy functions as generative models [10, 12].

Our results show nontrivial improvement over the best-performing models tested on CIFAR-10 and other benchmark datasets. Experiments confirm this gain is attributed to directly learning $p(\mathbf{x})$ via the energy function applied on unsupervised embedding learning. We empirically demonstrate the model's robustness by considering three practical scenarios. Codes for Elsa are released via a GitHub repository.[*]

2 Related Works

Reconstruction-based learning This approach assumes that generative model cannot fluently recover OOD samples, using the reconstruction error as anomaly score. Recent studies

[*] https://github.com/archon159/elsa

proposed methods to utilize such reconstruction errors. For example, multiple autoencoders can define reconstruction error with a synthetically generated blurred image [7]. Another study utilized gradients from back-propagation [20]. However, studies have also found that anomalies do not always yield a high reconstruction error when classes are similar [44].

Some studies employed GAN (generative adversarial network) to complement the reconstruction loss, for instance, via utilizing a generator and a discriminator [44] or using the reconstruction and discrimination loss [30, 35]. A series of studies proposed exploiting the generator's capability; for example, [43] suggested training a network by distinguishing synthetic anomalies generated by an ensemble of the generator's old state with the current states of generator and discriminator. Some studies have regularized the latent feature space in GAN-based models [30, 39]. However, GAN-based models are known to produce a suboptimal solution and hence are inapplicable for complex datasets [24, 28].

Self-supervised learning Self-supervised learning approaches are known to produce a robust representation of normal data, which leads to low confidence in classification probability for anomaly detection. Transformations like rotation, shift, and patch re-arranging can be used to augment pseudo-labels [16, 40, 41]. As discussed earlier, the state-of-the-art method in this domain, CSI [40], utilizes augmented pseudo-labels.

One-class classifiers This approach tries to learn the decision boundary between the distribution of training data and OOD samples. [33] suggested a loss function that forces training data samples to reside in a prefixed clustering centroid. [34] extended the problem into a semi-supervised learning objective and set the loss function to pull all unlabeled or positive samples closer to the centroid while pushing negative samples away from the centroid. [3] improved the detection performance by separately assigning a centroid to each augmentation. However, these methods are limited in their representation ability if the data distribution is complex and heterogeneous.

Energy-based learning Recently, OOD detection models using *energy function* have been proposed [12, 14, 24, 27]. The energy function measures the compatibility between a given input and a label [21]. One line of works exploited an energy-based model on top of a standard discriminative classifier. [12] demonstrated that energy-based training of the joint distribution improves OOD detection. [24] proposed energy scores to distinguish in- and OOD samples, showing that this score outperforms the softmax confidence score. Another line of works built their energy-based model on top of deep generative framework. [27] trains energy based model in the latent space to serve as a prior, while [14] jointly trains variational autoencoder and energy based model.

3 Background

We formally define the problem and offer a theoretical interpretation on why contrastive learning's objective does not match with anomaly detection under data contamination.

3.1 Contrastive learning (CL)

The core concept of CL is to train an encoder f by maximizing agreement among similar images (i.e., positive samples) while minimizing agreement among dissimilar images (i.e., negative samples). Let $\mathbf{x}$ be an input query, and a set of positive and negative samples of $\mathbf{x}$

be denoted $\mathcal{X}_+$ and $\mathcal{X}_-$. The contrastive loss is defined as:

$$L_c(\mathbf{x}) = -\frac{1}{|\mathcal{X}_+|} \log \frac{\sum_{\mathbf{x}' \in \mathcal{X}_+} \exp(\text{sim}(f(\mathbf{x}), f(\mathbf{x}'))/\tau)}{\sum_{\mathbf{x}' \in (\mathcal{X}_+ \cup \mathcal{X}_-)} \exp(\text{sim}(f(\mathbf{x}), f(\mathbf{x}'))/\tau)}, \tag{1}$$

$$= -\frac{1}{|\mathcal{X}_+|} (\underbrace{\log \sum_{\mathbf{x}' \in \mathcal{X}_+} \exp(\text{sim}(f(\mathbf{x}), f(\mathbf{x}'))/\tau)}_{L_{\text{align}}(\mathbf{x})} - \underbrace{\log \sum_{\mathbf{x}' \in (\mathcal{X}_+ \cup \mathcal{X}_-)} \exp(\text{sim}(f(\mathbf{x}), f(\mathbf{x}'))/\tau))}_{L_{\text{uniform}}(\mathbf{x})} \tag{2}$$

where τ is the temperature value that controls entropy [17] and sim$(\cdot)$ is the function that computes the similarity between two instances over the latent space.

We decompose the contrastive loss into two terms in Eq. 2 as in [42]. First is the alignment loss (L_{align}), which encourages embeddings of positive samples to be closely positioned. Next is the uniformity loss (L_{uniform}), which matches all samples into the pre-defined prior distribution with high entropy by pushing one another far away.

3.2 Energy-based model

Energy-based models [21] assume that any probability density p_θ can be expressed as $E_\theta(\cdot)$:

$$p_\theta(\mathbf{x}) = \frac{\exp(-E_\theta(\mathbf{x}))}{\int_{\mathbf{x}} \exp(-E_\theta(\mathbf{x}))}. \tag{3}$$

The energy function $E_\theta(\cdot)$ maps each data point $\mathbf{x}$ to a scalar value that represents its fit to given data distribution. For the energy function choice, one can also consider a decision-making model with two variables X and Y. In this scenario, the energy-based model defines the energy function $E_\theta(X,Y)$, and the energy function can be transformed in the form of a conditional probability with temperature τ as in Eq. 4 [24]. Then, $E_\theta(\mathbf{x})$ can be defined by marginalizing the energy function $E_\theta(\mathbf{x}, y')$ over y', as in Eq. 5.

$$p_\theta(y|\mathbf{x}) = \frac{\exp(-E_\theta(\mathbf{x}, y)/\tau)}{\int_{y'} \exp(-E_\theta(\mathbf{x}, y')/\tau)} = \frac{\exp(-E_\theta(\mathbf{x}, y)/\tau)}{\exp(-E_\theta(\mathbf{x})/\tau)} \tag{4}$$

$$E_\theta(\mathbf{x}) = -\tau \cdot \log \int_{y'} \exp(-E_\theta(\mathbf{x}, y')/\tau) \tag{5}$$

3.3 Rethinking the use of CL for anomaly detection

The objective of CL can be theoretically interpreted from the perspective of the energy function. For this, a discriminative classifier h_ψ can be considered, which maps each data input to a logit vector and estimates the categorical distribution with the following softmax function:

$$p_\psi(y|\mathbf{x}) = \frac{\exp(h_\psi(\mathbf{x})[y])}{\sum_{y'} \exp(h_\psi(\mathbf{x})[y'])}, \tag{6}$$

where $h_\psi(\mathbf{x})[y]$ indicates the y^{th} index of $h_\psi(\mathbf{x})$, i.e. the logit corresponding to the y^{th} class label. Combining Eq. 4 and Eq. 6 leads a model to optimize with the following energy function $E_\psi(\mathbf{x})$. This marginalizes the $E_\psi(\mathbf{x}, y)$ over y [12].

$$E_\psi(\mathbf{x}) = -\text{LogSumExp}_y(h_\psi(\mathbf{x})[y]) = -\log \sum_y \exp(h_\psi(\mathbf{x})[y]) \tag{7}$$

Similarly, we translate CL's objective as a classification task via considering each image instance to become a class of its own. Training a classifier by assigning the same pseudo-label to positive samples and a different label to negative samples will solve the objective of CL. If we denote $\hat{y}$ as a class label, the contrastive loss in Eq. 1 can be re-defined in the form of the cross-entropy loss:

$$L_c(\mathbf{x}) = -\frac{1}{|\mathcal{X}_+|}\log p(\hat{y}\in\mathcal{Y}_+|\mathbf{x})$$

$$p(\hat{y}\in\mathcal{Y}_+|\mathbf{x}) = \frac{\sum_{\mathbf{x}'\in\mathcal{X}_+}\exp(\mathrm{sim}(f(\mathbf{x}),f(\mathbf{x}'))/\tau)}{\sum_{\mathbf{x}'\in(\mathcal{X}_+\cup\mathcal{X}_-)}\exp(\mathrm{sim}(f(\mathbf{x}),f(\mathbf{x}'))/\tau)}, \tag{8}$$

where $\mathcal{Y}_+$ is the set of pseudo-labels corresponding to $\mathcal{X}_+$ (i.e., positive samples of $\mathbf{x}$). Since the re-defined contrastive loss has the same form as in Eq. 6, we obtain the energy function by marginalizing $E(\mathbf{x},\hat{y})$ over $\hat{y}$ in Eq. 9. Intuitively, this energy function represents how far an instance is placed from every training sample.

$$E(\mathbf{x}) = -\log\sum_{\mathbf{x}'\in(\mathcal{X}_+\cup\mathcal{X}_-)}\exp(\mathrm{sim}(f(\mathbf{x}),f(\mathbf{x}'))/\tau) \propto -L_{\text{uniform}}(\mathbf{x}) \tag{9}$$

The energy function from the contrastive objective need to be negatively proportional to the uniformity loss (L_{uniform}), based on Eq. 2 and Eq. 9. This means minimizing the contrastive loss leads to smaller uniformity loss as well as smaller energy scores — an observation that contradicts the original definition of the energy function, where a high energy value should correspond to the incompatible data configurations or anomalies.

4 Energy-based Learning for Semi-supervised Anomaly Detection (Elsa)

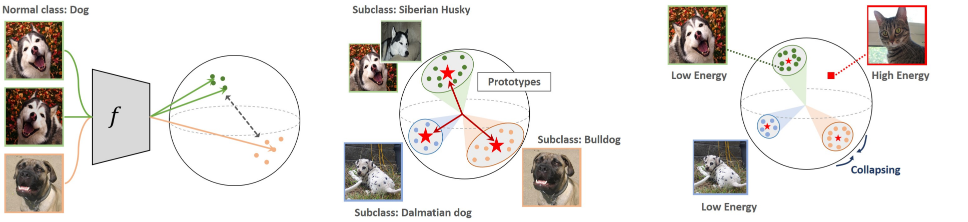
We tackle the energy maximization problem via a fine-tuning step that combines an energy function with unsupervised contrastive pre-training. Figure 2 illustrates these steps.

Problem statement: Let $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$ denotes a set of training images $\mathbf{x}_i$. In a semi-supervised problem setting, $\mathcal{D}$ can be divided into three disjoint sets $\mathcal{D} = \mathcal{X}_u \cup \mathcal{X}_n \cup \mathcal{X}_a$, where each stands for a set of unlabeled samples, labeled normal samples, and labeled anomaly samples. The main objective of anomaly detection is to train a normality score function $S(\mathbf{x})$ that represents the likelihood of a given instance $\mathbf{x}$ sampled from the normal data distribution. We assume that the majority of unlabeled samples $\mathcal{X}_u$ are normal, and thus let the model learn the normal data distribution from $\mathcal{X}_u \cup \mathcal{X}_n$, while distinguishing the labeled anomaly set $\mathcal{X}_a$. Elsa involves the following three steps:

(Step-1) Unsupervised contrastive pre-training

This step initializes the encoder f to learn general features from the normal data distribution. It only uses the unlabeled set $\mathcal{X}_u$ and the labeled normal set $\mathcal{X}_n$, and pre-trains the encoder f with an unsupervised CL approach, such as SimCLR [6]. Let $\hat{\mathbf{x}}^{(1)}$ and $\hat{\mathbf{x}}^{(2)}$ be two independent views of $\mathbf{x}$ from a pre-defined augmentation family $\mathcal{T}_a$ (i.e., $\hat{\mathbf{x}}^{(1)} = t_1(\mathbf{x})$, $\hat{\mathbf{x}}^{(2)} = t_2(\mathbf{x})$ where $t_1, t_2 \sim \mathcal{T}_a$). The unsupervised CL loss on the given pair of images is defined as:

$$L_{\text{CL}}(\hat{\mathbf{x}}^{(1)},\hat{\mathbf{x}}^{(2)}) = -\log\frac{\exp(\mathrm{sim}(f(\hat{\mathbf{x}}^{(1)}),f(\hat{\mathbf{x}}^{(2)}))/\tau)}{\sum_{\mathbf{x}'\in\hat{\mathcal{B}}^{(-)}}\exp(\mathrm{sim}(f(\hat{\mathbf{x}}^{(1)}),f(\mathbf{x}'))/\tau)}, \tag{10}$$



(a) Unsupervised pre-training (b) Choosing prototypes (c) Semi-supervised fine-tuning

Figure 2: Illustration of Elsa. The model is pre-trained via unsupervised contrastive learning. The model next chooses prototype vectors representing each subclass. The model is then fine-tuned with the energy function derived from the prototype vectors.

where $\hat{\mathcal{B}}^{(1)} = \{\hat{\mathbf{x}}_i^{(1)}\}_{i=1}^m$, $\hat{\mathcal{B}}^{(2)} = \{\hat{\mathbf{x}}_i^{(2)}\}_{i=1}^m$ denotes the batches with batch size m, and $\hat{\mathcal{B}}^{(-)} = \hat{\mathcal{B}}^{(1)} \cup \hat{\mathcal{B}}^{(2)} \setminus \hat{\mathbf{x}}^{(1)}$. τ and $\mathrm{sim}(\cdot)$ are defined as before.

By assigning the augmented variations of each training data instance in $\mathcal{X}_u \cup \mathcal{X}_n$ as positive samples, the model can maximize agreement among those samples (i.e., the numerator in Eq. 10). Simultaneously, all other instances in the same batch are treated as negative samples and pushed far away from one another in the latent space (i.e., the denominator in Eq. 10).

(Step-2) Prototypes selection

Conventional unsupervised CL maximizes the energy scores of all training samples and distinguishes every instance in the latent space. Toward this, one may try to minimize the energy score of normal samples, which is identical to maximizing the uniformity loss (Eq. 9). This approach, however, will pull every embedded point into a single position. Instead, we propose a new energy function by assigning a pseudo-label y_p to every training instance on the pre-trained embedding.

We define a set of prototypes $\mathcal{P}$ representing their subclasses in the normalized latent space. These prototypes can conceptually indicate heterogeneity of the training dataset. Then, every training sample should be mapped to the single nearest prototype $\mathbf{p} \in \mathcal{P}$ based on the cosine similarity. Following the form of the discriminative model, we may regard the encoder f as a classifier that maps each data point to a prototype, where the categorical distribution of each pseudo-label is computed via the following softmax function:

$$p(y_p|\mathbf{x}) = \frac{\exp(\mathrm{sim}(f(\mathbf{x}),\mathbf{p}))}{\sum_{\mathbf{p}' \in \mathcal{P}} \exp(\mathrm{sim}(f(\mathbf{x}),\mathbf{p}'))}. \tag{11}$$

This problem transformation lets us compute the energy function by marginalizing $E(\mathbf{x}, y_p)$ over y_p, similar to Eq. 9. The normality score function is then defined as a negation of the computed energy function (Eq. 12) with the temperature value τ.

$$S(\mathbf{x}) = -E(\mathbf{x}) = \log \sum_{\mathbf{p} \in \mathcal{P}} \exp(\mathrm{sim}(f(\mathbf{x}),\mathbf{p})/\tau) \tag{12}$$

To ensure prototypes are well dispersed, we choose centroids of each cluster to be prototypes. Spherical k-means clustering algorithm over the embeddings of training samples is considered in this work, with the cosine similarity as a distance metric.

(Step-3) Fine-tuning with prototypes

Finally, the model is fine-tuned via the following loss (L_e):

$$L_e = \sum_{\mathbf{x}\in\mathcal{X}_a} \frac{1}{C - S(\mathbf{x})} + \sum_{\mathbf{x}\in\mathcal{X}_u\cup\mathcal{X}_n} \frac{1}{S(\mathbf{x})}, \tag{13}$$

where C is a constant. This loss minimizes the normality scores of abnormal samples $\mathcal{X}_a$ and maximizes the scores of the mostly normal samples $\mathcal{X}_u \cup \mathcal{X}_n$. To stabilize this training process, we use an inverse form as the learning objective. According to the gradient of L_e with respect to $\mathbf{x}$ (Eq. 14), the inverse of the quadratic term is multiplied on the gradient of the score function, $\nabla_{\mathbf{x}} S(\mathbf{x})$. This multiplier helps reduce the gradient signal when the score becomes sufficiently small for $\mathcal{X}_a$ or large for $\mathcal{X}_u \cup \mathcal{X}_n$. To ensure denominators remain positive, we set the constant C as the largest possible value; given the input instance has the maximum similarity with all prototypes (i.e., $\mathrm{sim}(f(\mathbf{x}),\mathbf{p}) = 1$ for all prototypes).

$$\nabla_{\mathbf{x}} L_e = \sum_{\mathbf{x}\in\mathcal{X}_a} \nabla_{\mathbf{x}} S(\mathbf{x}) \left(\frac{1}{C - S(\mathbf{x})}\right)^2 - \sum_{\mathbf{x}\in\mathcal{X}_u\cup\mathcal{X}_n} \nabla_{\mathbf{x}} S(\mathbf{x}) \left(\frac{1}{S(\mathbf{x})}\right)^2 \tag{14}$$

Since we embed the data instances in $\mathcal{X}_u \cup \mathcal{X}_n$ nearby the chosen prototypes in the latent space, this process can be interpreted as the minimum volume estimation [37] over the normalized latent space. It is connected to two works, Deep-SVDD [33] and Deep-SAD [34], where the training instances are collapsed into a single centroid in the latent space. However, we utilize *multiple* centroids (called 'prototypes' in this research) to account for heterogeneity in data. Multiple centroids are better suited to learning distinct features from heterogeneous data; the same learning capability is hard for a single centroid. For example, Bulldogs and Siberian Huskies have distinctive visual features, yet they belong to the same class of dogs. Our model will assign these two dog types to different prototypes as in Fig. 2b.

As fine-tuning continuously changes the distribution of data instances in the latent space, one can no longer ensure the previous step's prototypes to be valid. Thus, we update the prototypes to fit the fine-tuned distribution periodically. This is done by repeating step-2 every few epochs in step-3.

Next, we introduce a novel strategy on *early stopping* to avoid overfitting and guide the model to determine when to stop. The strategy is based on the observation on strongly augmented images and their potential use as a validation indicator. Strong augmentations such as AutoContrast, Shear, and Cutout [9] can be regarded as tentative anomalies due to massive content-wise distortions [8, 38]. RandAugment algorithm [8] is used as the strong augmentation and separated the validation set from the unlabeled training set with a ratio of 5-to-95. The AUROC scores between the original and augmented images determine the final model, denoted as the *earlystop* score.

Extension with contrasting shifted instances

Elsa is next optimized in two ways. First, the training instances can be augmented via rotations. Enlarging the data size helps learn features more effectively, and enables to increase the number of prototypes (i.e., allowing data heterogeneity). Second, ensemble technique can be adopted during the inference. We iteratively calculated the normality scores from multiple views of the same image via random weak augmentation. Then, the ensembled score was computed by averaging the normality scores. Implementing the above techniques on Elsa, we propose an extended model Elsa+. Algorithm details are described in the supplementary material.

5 Experiments

5.1 OOD detection result

We consider three representative scenarios proposed from existing works on three datasets: CIFAR-10, ImageNet-10, and Places-5. The latter two are random 10 and 5 class subsets of ImageNet and Places-365 datasets. Details of each scenario are described below.

(Scenario-1) Semi-supervised classification [2, 34]. Here we assume having access to a small subset of labeled normal $\mathcal{X}_n$ and anomalies $\mathcal{X}_a$ during training. One of the data classes in CIFAR-10 is set as in-distribution and let the remaining nine classes represent an anomaly. This means to sample $\mathcal{X}_a$ from the nine anomaly classes. Let the ratio of $\mathcal{X}_n$ and $\mathcal{X}_a$ both be denoted as γ_l. We then report the averaged AUROC scores, following the standards of previous literature [33, 34, 40], on the test set over 90 experiments (10 normal $\times$ 9 anomaly) for a given γ_l.

(Scenario-2) Contaminated one-class classification [34]. The next scenario tests the model robustness under contamination in the training set, which is our utmost interest. This data contamination scenario is widely applicable to any scenarios that involve datasets obtained from crawling or crowdsourcing [32]. It starts with the same setting as in Scenario-1. We assume the training data is polluted with a fixed ratio γ_p. This is done by sampling images from every anomaly class and adding them into the unlabeled set $\mathcal{X}_u$. We report the averaged AUROC scores over 90 experiments for each pollution ratio γ_p. The labeling ratio γ_l is fixed to 0.05 for all experiments.

(Scenario-3) Auxiliary anomaly set [24]. In real-world settings, anomaly samples may be hard to obtain, as discussed in the task of identifying malicious users, credit card frauds, and crowd surveillance [26, 45]. This scenario tests whether the proposed model can leverage a large-scaled external dataset as an auxiliary anomaly set, which is an easy alternative for anomalies. We set all images in CIFAR-10 as in-distribution and let images from other datasets as anomalies. Then, we train the model with an auxiliary dataset (i.e., down-sampled ImageNet) as labeled anomalies and evaluate the detection performance on five other datasets with AUROC metrics.

Results. We discuss the results of each experiment. First, in the semi-supervised one-class classification in Scenario-1, Elsa+ achieves the state-of-the-art performance against all baselines (Table 1). The model works well even with a small set of labeled samples ($\gamma_l = 0.01$). The next experiment for Scenario-2 tests the model's performance against data contamination (Table 2), which shows both Elsa and Elsa+ to be stable against contamination in the training data. Table 3 reports the anomaly detection results for both scenario-1 and 2 over ImageNet-10 and Places-5, which demonstrates the proposed model's applicability to large-scale dataset. A significant performance drop seen for CSI supports our claim that the CL objective alone fails to handle data contamination. In contrast, the energy-based fine-tuning step can alleviate this problem and achieve outstanding performance.

Lastly, Table 4 shows the result for Scenario-3 on the model's ability to leverage the external dataset as an auxiliary outlier. Compared to other strong baselines, Elsa+ shows the highest or comparable detection performance for all datasets.

5.2 Component analyses

We also test the contribution of each component in Elsa+ in four critical analyses. We regard the plane class in CIFAR-10 as the normal data and fix the labeling ratio γ_l and pollution ratio γ_p to 0.05 (Scenario-2) for the remainder of this section.

γ_l	Latent-EBM [27]	VAE-EBM [14]	OC-SVM [36]	IF [23]	KDE [33]	DeepSVDD [33]	GOAD [1]	CSI [40]	SS-DGM [18]	SSAD [11]	DeepSAD [34]	Elsa	Elsa+
.00	56.6	65.3	62.0	60.0	59.9	60.9	88.2	**94.3**	-	62.0	60.9	-	-
.01									49.7	73.0	72.6	80.0	**94.3**
.05									50.8	71.5	77.9	85.7	**95.2**
.10									52.0	70.1	79.8	87.1	**95.5**

Table 1: Experiment results on anomaly detection Scenario-1 over CIFAR-10.

γ_p	Latent-EBM [27]	VAE-EBM [14]	OC-SVM [36]	IF [23]	KDE [33]	DeepSVDD [33]	GOAD [1]	CSI [40]	SS-DGM [18]	SSAD [11]	DeepSAD [34]	Elsa	Elsa+
.00	56.6	65.3	62.0	60.0	59.9	60.9	88.2	94.3	50.8	73.8	77.9	85.7	**95.2**
.05	61.2	63.5	61.4	59.6	58.1	59.6	85.2	88.2	50.1	71.5	74.0	83.5	**93.0**
.10	60.3	64.8	60.8	58.8	57.3	58.6	83.0	84.5	50.5	69.8	71.8	81.6	**91.1**

Table 2: Experiment results on anomaly detection Scenario-2 over CIFAR-10.

ImageNet-10	CSI	Elsa+	Places-5	CSI	Elsa+
$\gamma_p = 0.0$	0.95	0.93	$\gamma_p = 0.0$	0.81	0.88
$\gamma_p = 0.1$	0.86	0.90	$\gamma_p = 0.1$	0.69	0.86

Table 3: Experiment results on anomaly detection Scenario-1 and 2 over ImageNet-10 and Places-5. ($\gamma_l = 0.1$)

Datasets		GOAD	CSI	Elsa+
CIFAR-10 [19] →	SVHN [25]	96.3	**99.8**	**99.4**
	LSUN [22]	89.3	97.5	**99.9**
	LSUN (FIX) [40]	78.8	90.3	**95.0**
	ImageNet (FIX) [40]	83.3	93.3	**96.4**
	CIFAR-100 [19]	77.2	**89.2**	86.3

Table 4: Experiment results on anomaly detection in Scenario-3.

Ablations		AUROC (%)
Score function	Cosine similarity: $S_{cos}(\mathbf{x})$	89.1
	Energy from CL objective: $S_{cont}(\mathbf{x})$	81.6
Loss objective	Naive loss: L_{naive}	82.6
	DeepSAD loss: L_{sad}	90.3
Elsa+ (Ours)		**91.4**

Table 5: Ablation study results on the score function and loss objective over CIFAR-10.

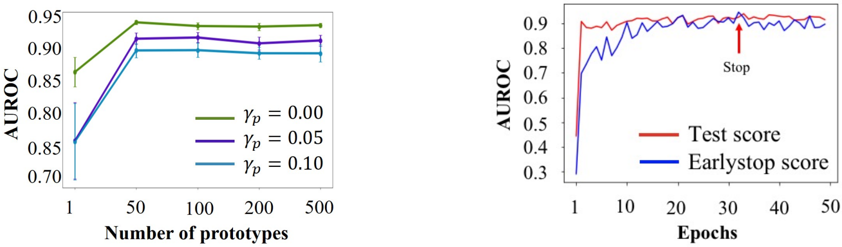
Ablation study on the score and loss function. We explore several possible score functions and objectives as alternatives, thereby measuring each component's contribution. The description of each ablation is described below. The first two are ablations on alternative score functions. The next two are ablations on alternative loss objectives.

- Cosine similarity. A normality score can be obtained by measuring the cosine similarity of the given sample and the nearest prototype vector: $S_{cos}(\mathbf{x}) = max_{\mathbf{p}\in\mathcal{P}}(f(\mathbf{x}), \mathbf{p})$.
- Energy from CL objective. The uniformity loss in CL objective (Eq. 9) can be adopted for score function. With the given augmented batch $\mathcal{B}$ from the training set (i.e., $\mathcal{X}_n \cup \mathcal{X}_u$), the normality score is defined as: $S_{cont}(\mathbf{x}) = \log \sum\limits_{\mathbf{x}'\in\mathcal{B}\setminus\{\mathbf{x}\}} \exp(\text{sim}(f(\mathbf{x}), f(\mathbf{x}')))$.
- Naive loss. The naive loss is the simplest form that maximizes the score for normal samples as a negative form of score function $(-S(\mathbf{x}))$ while minimizing for anomalies as a positive form of score function $(S(\mathbf{x}))$: $L_{naive} = \sum_{\mathbf{x}\in\mathcal{X}_a} S(\mathbf{x}) + \sum_{\mathbf{x}\in\mathcal{X}_n\cup\mathcal{X}_u} -S(\mathbf{x})$
- DeepSAD loss. This form of loss is introduced in DeepSAD [34]. The score for anomalous samples are maximized as the inverse form of the score function: $L_{sad} = \sum_{\mathbf{x}\in\mathcal{X}_a} 1/(C - S(\mathbf{x})) + \sum_{\mathbf{x}\in\mathcal{X}_n\cup\mathcal{X}_u} -S(\mathbf{x})$

Table 5 shows the results, where all ablations lead to a substantial performance drop. Specifically, changing the CL objective's energy function led to the most extensive degradation, reinforcing our theoretical interpretation presented in §3.3. Our loss objective design, which exploits the inverse form of the energy function for both anomaly and normal samples, achieves the best performance compared to all ablations.

Analyses on prototype count. We investigate the dependency between the prototype count and Elsa+ performance. The prototype count decides the model's capacity for handling heterogeneous data types within the normal distribution. It also directly impacts the overall performance, as too small or too large count leads to underfitting or overfitting. We analyze the prototype count's effect by varying it to 1, 50, 100, 200, and 500. Figure 3a shows the mean AUROC score with the standard error over three different contamination ratios (γ_p = 0.00, 0.05, 0.10). The model fails to converge for the count of 1, and the early stop strategy does not work in order. In contrast, the model converges with a successful result for the larger counts, e.g., 100. Given the prototype count is set to a reasonably large value, Elsa+ consistently produces high-performance results.

Analysis on the early stopping strategy. The early stopping strategy is another important factor to be examined. The earlystop score is computed by the AUROC score of the task that distinguishes between the input image and its strongly augmented versions. This analysis reveals a highly positive correlation between the earlystop score and the actual model performance (Pearson correlation 0.912±0.038), implying that the proposed score can guide the actual performance on the test set. Furthermore, Figure 3b shows that the earlystop score eventually converges, and it gives an appropriate timing for earlystop with high detection performance on the test-set.



(a) Performance by the number of prototypes (b) Test-set and early-stop score across epochs

Figure 3: Analyses on the number of prototypes and early stopping.

6 Conclusion

We presented a unified energy-based approach for semi-supervised anomaly detection. We demonstrated that the contrastive learning objective is potentially fragile in the contamination scenarios interpreted from the energy-based perspective. We suggested a new energy function concerning prototypes and introduced the fine-tuning process in this light. With these components, the proposed model Elsa and Elsa+ could successfully distinguish anomalies from normal samples while leveraging the high representation power of unsupervised contrastive learning. Elsa+ achieves SOTA performance among other baselines and shows strong robustness against the contamination of unknown anomalies. We believe our effort will renew the interest in energy-based learning of OOD detection.

Acknowledgments. This work was supported by the Institute for Basic Science (IBS-R029-C2) and the National Research Foundation (No. NRF-2017R1E1A1A01076400) funded by the Ministry of Science and ICT in Korea.

References

[1] Davide Abati, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Latent space autoregression for novelty detection. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 481–490, 2019.

[2] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Proc. of the Asian Conference on Computer Vision (ACCV)*, pages 622–637. Springer, 2018.

[3] Liron Bergman and Yedid Hoshen. Classification-based anomaly detection for general data. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2020.

[4] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. 41(3), 2009. ISSN 0360-0300.

[5] Ting Chen and Lala Li. Intriguing properties of contrastive losses. *arXiv preprint arXiv:2011.02803*, 2020.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 1597–1607, 2020.

[7] Sungik Choi and Sae-Young Chung. Novelty detection via blurring. In *Proc. of the International Conference on Learning Representations (ICLR)*. International Conference on Learning Representations, 2020.

[8] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 702–703, 2020.

[9] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[10] Ruiqi Gao, Erik Nijkamp, Diederik P Kingma, Zhen Xu, Andrew M Dai, and Ying Nian Wu. Flow contrastive estimation of energy-based models. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7518–7528, 2020.

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

[12] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*, 2019.

[13] Sungwon Han, Sungwon Park, Sungkyu Park, Sundong Kim, and Meeyoung Cha. Mitigating embedding and class assignment mismatch in unsupervised image classification. In *Proc. of the 16th European Conference on Computer Vision (ECCV)*. Springer, 2020.

[14] Tian Han, Erik Nijkamp, Linqi Zhou, Bo Pang, Song-Chun Zhu, and Ying Nian Wu. Joint training of variational auto-encoder and latent energy-based model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7978–7987, 2020.

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020.

[16] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *The Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[17] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[18] Diederik P Kingma, Danilo J Rezende, Shakir Mohamed, and Max Welling. Semi-supervised learning with deep generative models. In *Proc. of the 27th International Conference on Neural Information Processing Systems (NeurIPS) Volume 2*, pages 3581–3589, 2014.

[19] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

[20] Gukyeong Kwon, Mohit Prabhushankar, Dogancan Temel, and Ghassan AlRegib. Backpropagated gradient representations for anomaly detection. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 206–226. Springer, 2020.

[21] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

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

[23] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *Proc. of the Eighth IEEE International conference on data mining (ICDM)*, pages 413–422. IEEE, 2008.

[24] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 2020.

[25] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *Proc. of the Neural Information Processing Systems (NeurIPS) Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[26] Phuc Cuong Ngo, Amadeus Aristo Winarto, Connie Khor Li Kou, Sojeong Park, Farhan Akram, and Hwee Kuan Lee. Fence gan: Towards better anomaly detection. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 141–148. IEEE, 2019.

[27] Bo Pang, Tian Han, Erik Nijkamp, Song-Chun Zhu, and Ying Nian Wu. Learning latent space energy-based prior model. *Advances in Neural Information Processing Systems*, 33, 2020.

[28] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton van den Hengel. Deep learning for anomaly detection: A review. *arXiv preprint arXiv:2007.02500*, 2020.

[29] Sungwon Park, Sungwon Han, Sundong Kim, Danu Kim, Sungkyu Park, Seunghoon Hong, and Meeyoung Cha. Improving unsupervised image clustering with robust learning. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12278–12287, 2021.

[30] Pramuditha Perera, Ramesh Nallapati, and Bing Xiang. Ocgan: One-class novelty detection using gans with constrained latent representations. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[31] Marco A.F. Pimentel, David A. Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014. ISSN 0165-1684. doi: https://doi.org/10.1016/j.sigpro.2013.12.026.

[32] Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q Weinberger. Identifying mislabeled data using the area under the margin ranking. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, pages 17044–17056. Curran Associates, Inc.

[33] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 4393–4402. PMLR, 2018.

[34] Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2020.

[35] 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 *Proc. of the International conference on information processing in medical imaging (ICIPMI)*, pages 146–157. Springer, 2017.

[36] 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*, 13(7):1443–1471, 2001.

[37] Clayton D Scott and Robert D Nowak. Learning minimum volume sets. *The Journal of Machine Learning Research*, 7:665–704, 2006.

[38] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

[39] Gowthami Somepalli, Yexin Wu, Yogesh Balaji, Bhanukiran Vinzamuri, and Soheil Feizi. Unsupervised anomaly detection with adversarial mirrored autoencoders, 2021.

[40] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. In *Proc. of the Neural Information Processing Systems (NeurIPS)*, 2020.

[41] Siqi Wang, Yijie Zeng, Xinwang Liu, En Zhu, Jianping Yin, Chuanfu Xu, and Marius Kloft. Effective end-to-end unsupervised outlier detection via inlier priority of discriminative network. In *Proc. of the Neural Information Processing Systems (NeurIPS)*, pages 5960–5973, 2019.

[42] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proc. of the International Conference on Machine Learning (ICLR)*, pages 9929–9939. PMLR, 2020.

[43] Muhammad Zaigham Zaheer, Jin-ha Lee, Marcella Astrid, and Seung-Ik Lee. Old is gold: Redefining the adversarially learned one-class classifier training paradigm. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14183–14193, 2020.

[44] Houssam Zenati, Manon Romain, Chuan-Sheng Foo, Bruno Lecouat, and Vijay Chandrasekhar. Adversarially learned anomaly detection. In *Proc. of the IEEE International conference on data mining (ICDM)*, pages 727–736. IEEE, 2018.

[45] Panpan Zheng, Shuhan Yuan, Xintao Wu, Jun Li, and Aidong Lu. One-class adversarial nets for fraud detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1286–1293, 2019.