World Wide Web (2025) 28:40
https://doi.org/10.1007/s11280-025-01353-z

Multi-label out-of-distribution detection with spectral normalized joint energy

Yihan Mei[1] · Xinyu Wang[1] · Changzhi Sun[2] · Dell Zhang[2] · Xiaoling Wang[1]

Received: 18 February 2025 / Revised: 12 May 2025 / Accepted: 17 May 2025

Abstract

In today's interconnected world, achieving reliable out-of-distribution (OOD) detection poses a significant challenge for machine learning models. While numerous studies have introduced improved approaches for multi-class OOD detection tasks, the investigation into multi-label OOD detection tasks has been notably limited. We introduce Spectral Normalized Joint Energy (SNoJoE), a method that consolidates label-specific information across multiple labels through the theoretically justified concept of an energy-based function. Throughout the training process, we employ spectral normalization to manage the model's feature space, thereby enhancing model efficacy and generalization, in addition to bolstering robustness. Our findings indicate that the application of spectral normalization to joint energy scores notably amplifies the model's capability for OOD detection. We perform OOD detection experiments utilizing PASCAL-VOC or MS-COCO as the in-distribution dataset and ImageNet-22K or Texture as the out-of-distribution datasets. Our experimental results reveal that, in comparison to prior top performances, SNoJoE achieves 11% and 54% relative reductions in FPR95 on ImageNet-22K and Texture, respectively, when using PASCAL-VOC as the in-distribution dataset. Similarly, with MS-COCO as the in-distribution dataset, SNoJoE achieves 11.3% and 42.58% relative reductions on ImageNet-22K and Texture. These improvements establish a new *state of the art* in OOD detection and further validate the effectiveness of incorporating spectral normalization.

This article belongs to the Topical Collection: *Special Issue on APWeb-WAIM 2024*

✉ Dell Zhang
dell.z@ieee.org

✉ Xiaoling Wang
xlwang@cs.ecnu.edu.cn

Yihan Mei
yihanmei@stu.ecnu.edu.cn

Xinyu Wang
xinyu_wang@stu.ecnu.edu.cn

Changzhi Sun
czsun.cs@gmail.com

[1] Department of Information, East China Normal University, Shanghai 200062, P.R. China

[2] Institute of Artificial Intelligence (TeleAI), China Telecom, Beijing, P.R. China

Published online: 02 June 2025

Keywords Out-of-distribution detection · Multi label classification · Spectral normalization

1 Introduction

In the current digital era, the pervasive use of machine learning models is undeniable. However, these models often grapple with data that deviates from their training data, known as out-of-distribution (OOD) data, when deployed in real-world settings. This discrepancy can lead to inaccurate predictions, raising safety concerns and other issues. OOD detection plays a crucial role in identifying unfamiliar data, thereby enhancing model safety and robustness in diverse environments. Thus, assessing OOD uncertainty emerges as a critical challenge for researchers.

Significant advancements have been made in OOD detection research. The Local Outlier Factor (LOF) method [1] and unsupervised outlier detection using globally optimal sample-based Gaussian Mixture Models (GMM) by Yang et al. [2] represent foundational work. G-ODIN [3] builds on ODIN [4] to improve sensitivity to covariate shifts. OpenMax [5] introduces Extreme Value Theory (EVT) to neural networks, calibrating logits with EVT probability models, including the Weibull distribution. Classification-based approaches see innovations like extending One-Class Classification (OCC) through elastic-net regularization for learning decision boundaries [6], and selecting reliable data from unlabeled sources as negative samples for supervised anomaly detection settings [7].

Despite these advancements, OOD detection in multi-label classification contexts remains underexplored. Multi-label classification poses unique challenges due to the necessity of evaluating uncertainty across multiple labels, rather than a single dominant one [8]. Achieving stable model training is essential for accurate multi-label OOD sample identification, with strategies like using free energy for OOD uncertainty assessment proposed by Liu et al. [9].

This paper introduces a novel approach, **S**pectral **No**rmalized **Jo**int **E**nergy (SNoJoE), for assessing OOD uncertainty in multi-label datasets. SNoJoE calculates free energy for each label and combines these energies, overcoming the difficulties generative models face in estimating joint likelihood for multi-label data [10]. Additionally, it demonstrates that aggregating label energies is more effective than summing label scores in OOD detection evaluations [8], highlighting the importance of choosing the right label assessment function.

We also utilize ResNet for feature extraction from in-distribution images, employing an energy function as the metric for OOD assessment. To counter overfitting and enhance model robustness, we apply spectral normalization as a regularization technique. Our findings show that spectral normalization reduces gradient variation ranges during training, minimizing the risk of gradient problems and promoting a well-regulated feature space. This approach helps the model to generalize better to OOD instances by focusing on extracting generalizable features rather than memorizing training data. Applying spectral normalization to OOD detection tasks has been shown to significantly improve model performance, such as achieving a 54% reduction in FPR95 on the Texture dataset with respect to PASCAL-VOC (t-test p-value < 0.01), underscoring the technique's value in OOD detection. Similarly, when using MS-COCO as the in-distribution dataset, spectral normalization leads to a 42.58% reduction in FPR95 on the Texture dataset, further demonstrating its robustness and effectiveness across different in-distribution settings.

Our main contributions include:

- Introducing SNoJoE, an innovative method for OOD uncertainty assessment in multi-label classification that can deliver today's best performance on two real-world datasets.

- Demonstrating through ablation studies that spectral normalization significantly enhances multi-label OOD detection performance.
- Establishing an enhanced baseline incorporating spectral regularization for comparison, providing deeper insights into the role of spectral constraints in OOD detection.
- Making our experimental code and datasets available for reproducible research.[1]

2 Related work

2.1 Multi-label classification

Unlike the simpler scenario of multi-class (single-label) classification, multi-label classification allows each image to be associated with multiple label concepts. Early approaches to multi-label classification treated the presence of each label independently, neglecting the potential correlations among labels [11, 12].

Initial research in multi-label classification demanded significant computational resources. Ghamrawi and McCallum [13] employed Conditional Random Fields (CRF) to create graphical models that identify correlations between labels, and Chen et al. [14] integrated CRF with deep learning techniques to examine the dependencies among output variables. These strategies necessitate the explicit modeling of label correlations, leading to elevated computational demands.

Conversely, deep learning techniques do not inherently require substantial computational resources for multi-class recognition tasks and have shown notable effects [15, 16]. Gong et al. [11] utilized Convolutional Neural Networks (CNN) to label images with 3 or 5 labels in the NUS-WIDE dataset, while Chen et al. [17] applied CNNs to categorize road scene images from a set of 52 potential labels. Thus, efficiently solving multi-label classification challenges is intricately linked to a wide range of applications in the contemporary open world.

2.2 Out-of-distribution detection

In the realm of OOD detection, research has primarily concentrated on four areas: Novelty Detection (ND), Open Set Recognition (OSR), Outlier Detection (OD), and Anomaly Detection (AD).

Initially, methods leaned heavily on confidence estimation and the setting of thresholds, judging inputs' relevance to known categories by the confidence scores produced by the model. However, they often falter when facing complex data distributions. Zhang et al. [18] introduced OpenHybrid, a strategy that combines representation space learning from both an inlier classifier and a density estimator, the latter acting as an outlier detector.

Ayadi et al. [19] outlined twelve diverse interpretations of outliers, highlighting the challenge of defining outliers precisely. This has spurred a wave of innovative approaches for identifying and addressing outliers [20]. Among them, density-based methods for detecting outliers represent some of the earliest strategies. The Local Outlier Factor (LOF) method, introduced by Breunig et al. [1], stands as a pioneering density-based clustering technique for outlier detection, leveraging the concept of loose correlation through k-nearest neighbors (KNN). LOF calculates local reachability density (LRD) within each point's KNN set and compares it to the densities of neighbors within that set.

Yang et al. [2] proposed an unsupervised outlier detection approach using a globally optimal sample-based Gaussian Mixture Model (GMM), employing the Expectation-

[1] https://github.com/Nicholas-Mei/Ood_Detection_SNoJoE

Maximization (EM) algorithm for optimal fitting to the dataset. They define an outlier factor for each data point as the weighted sum of mixture proportions, where weights denote the relationships among data points.

Certain studies have focused on increasing sensitivity to covariate shifts by examining hidden representations in neural networks' intermediate layers. Generalized ODIN [3] builds on ODIN [4] by adopting a specialized training objective, DeConf-C, and choosing hyperparameters like perturbation magnitude for in-distribution data. Wei et al. [21] demonstrated that issues of overconfidence could be alleviated through Logit Normalization (Logit Norm), which counters the typical cross-entropy loss by enforcing a constant vector norm on logits during training, enabling neural networks to distinctly differentiate between in-distribution and OOD data. Other efforts have sought to refine OOD uncertainty estimation via confidence scores based on Mahalanobis distance [22] and gradient-based GradNorm scores [23].

Within classification-based OOD detection methods, One-Class Classification (OCC) uniquely establishes a decision boundary matching the expected normal data distribution density level set [24]. Deep SVDD [25] was the first to adapt classical OCC for deep learning, mapping normal samples to a hypersphere to delineate normality. Deviations from this model are flagged as anomalous. Later efforts expanded this approach through elastic regularization [6] or adaptive descriptions with multi-linear hyperplanes [26]. Additionally, some methods employ Positive-Unlabeled (PU) learning in semi-supervised AD contexts, providing unlabeled data alongside normal data. Mainstream PU strategies either select reliable negative samples for a supervised AD setting, using clustering [7] and density models [27], or treat all unlabeled data as noise negatives for learning with noise labels, employing sample re-weighting [28] and label cleaning [29, 30].

Despite advancements, OOD detection remains a challenging field, predominantly explored within multi-class tasks, with limited work in multi-label classification. A notable exception is YolOOD [31], which draws inspiration from object detection frameworks to address the multi-label OOD detection problem. Instead of modeling the entire image holistically, YolOOD treats each predicted label as a distinct region and evaluates its "objectness" to determine whether it corresponds to an in-distribution or out-of-distribution label. This reformulation allows the model to localize and score anomalous instances at the region level, offering fine-grained anomaly signals particularly suitable for complex multi-label scenarios.

Hence, we introduce a technique that integrates spectral normalization into the network and utilizes energy scores to derive label-wise joint energy scores for OOD detection tasks.

2.3 Energy-based models

Energy-based models (EBMs) in machine learning trace their origins to Boltzmann machines [32]. This approach offers a cohesive framework encompassing a broad spectrum of learning algorithms, both probabilistic and deterministic [33, 34]. Xie et al. [35] showed that the discriminative classifiers within GAN networks can be interpreted through an energy-based lens. Moreover, these methods have been leveraged for structured prediction challenges [36].

Recent studies [9, 37] have advocated for the use of energy scores in detecting OOD instances, grounding their arguments in theoretical perspectives related to likelihood [38]. Here, samples exhibiting lower energy are classified as in-distribution (ID), while those with higher energy are flagged as OOD. Liu et al. [9] pioneered a technique for quantifying OOD uncertainty by utilizing energy scores, showcasing remarkable efficacy in multi-class classification networks. Meanwhile, research by Wang et al. [8] targets multi-label contexts, illustrating the benefits of harnessing the collective power of all label data. Our contribution

merges cross-label energy scores, affirming enhanced performance through the implementation of spectral normalization.

3 Method

In this section, we introduce a novel approach for OOD detection in multi-label scenarios. First, we address multi-label inputs by integrating concepts from the free energy function, assessing OOD uncertainty through the evaluation of joint label energies across labels. Subsequently, we present SNoJoE, a technique that applies spectral normalization to the joint label energy scores. This enhancement not only improves the model's robustness but also facilitates the extraction of features that are more generalizable.

3.1 Preliminaries

3.1.1 Multi-label classification

Multi-label classification is a machine learning task where the goal is to assign input data samples to one or more categories out of a set of predefined labels. Unlike traditional single-label classification tasks, where each sample can only belong to one category, multi-label classification allows a sample to have multiple labels simultaneously. Generally, consider $\mathcal{X}$ (representing the input space) and $\mathcal{Y}$ (representing the output space), with $\mathcal{P}$ denoting a distribution over $\mathcal{X} \times \mathcal{Y}$. Suppose $f : \mathcal{X} \longrightarrow \mathbb{R}^{|\mathcal{Y}|}$ represents a neural network trained on samples drawn from P. Each input can be correlated with a subset of labels in $\mathcal{Y} = 1, 2, \cdots, K$, denoted by a vector $\mathbf{y} = [y_1, y_2, \cdots, y_K]$, where

$$y_i = \begin{cases} 1 \text{ , if } i \text{ is associated with } x \\ 0 \text{ , otherwise} \end{cases} . \tag{1}$$

Utilizing a convolutional neural network (CNN) with a shared feature space, we generate multi-label output predictions. This approach has emerged as the standard training mechanism for multi-label classification tasks, finding widespread application across various domains [39, 40].

3.1.2 Out-of-distribution detection

Similar to the concept presented in [8], we define the problem of OOD detection for multi-label classification as follows. Let D_{in} denote the marginal distribution $\mathcal{P}$ over the label set $\mathcal{X}$, representing the distribution of in-distribution data. During testing, the environment may generate out-of-distribution data D_{out} on $\mathcal{X}$. The goal of OOD detection is to define a decision function D such that:

$$D(x; f) = \begin{cases} 1 \text{ , if x} \sim \mathcal{D}_{in} \\ 0 \text{ , if x} \sim \mathcal{D}_{out} \end{cases} . \tag{2}$$

3.1.3 Energy function

The definition of the energy equation was first proposed by Liu et al. They introduced the free energy as the scoring function for OOD uncertainty assessment in a multi-class setting.

Given a classifier $f(\mathrm{x}) : \mathcal{X} \to \mathbb{R}^K$ mapping the input x to K real numbers as logits, the class distribution is represented through softmax:

$$p(y_i = 1|x) = \frac{e^{f_{y_i}(x)}}{\Sigma_{j=1}^{K} e^{f_{y_j}(x)}}. \tag{3}$$

Then, the transformation from logits to probability distribution is achieved through the Boltzmann distribution:

$$p(y_i = 1|x) = \frac{e^{-E(x,y_i)}}{\int_{y'} e^{-E(x,y')}} = \frac{e^{-E(x,y_i)}}{e^{-E(x)}}. \tag{4}$$

Thus, the initially defined classifier can be interpreted from an energy-based perspective. Viewing the logits $f_{y_i}(x)$ as an energy function $E(x, y_i)$, we can obtain the free energy function $E(x)$ for any given input x:

$$E(x) = -\log \sum_{i=1}^{K} e^{f_{y_i}(x)}. \tag{5}$$

3.1.4 Spectral regularization

Deep neural networks often suffer from overfitting and instability due to excessively large singular values in their weight matrices. Spectral regularization provides a solution by constraining these singular values, thereby improving model generalization and robustness. This technique has been widely used in adversarial robustness, structured sparsity, and general deep learning applications where model stability is crucial.

The core idea of spectral regularization is to control the spectral norm or other related measures (e.g., Frobenius norm or nuclear norm) of the weight matrices in the network. By reducing the dominance of large singular values, the model learns a more balanced and compact representation, leading to improved generalization and better resistance to input perturbations.

In our implementation, we adopt spectral norm regularization, which directly penalizes the largest singular value $\sigma_{max}(W)$ of a weight matrix W. The additional loss term is formulated as:

$$\mathcal{L}_{\text{spec}} = \lambda_{\text{reg}} \|\sigma_{max}(W)\| \tag{6}$$

where λ_{reg} is the regularization strength that controls the impact of the spectral constraint.

This method provides different levels of constraint on the weight matrices, and their effects are analyzed in Section 4.

3.2 Label-wise joint energy

We first consider the problem of OOD uncertainty detection on a standard multi-label classifier. For a given input x, its output for the i-th class is:

$$f_{y_i}(x) = h_{l-1}(x) \cdot w_{cls}^{i}, \tag{7}$$

where $h_{l-1}(x)$ is the feature vector of the penultimate layer of the network, and w_{cls}^{i} is the weight matrix corresponding to i-th class. The predictive probability of label y_i is then implemented through a variant of a binary logistic classifier:

$$p(y_i = 1 \mid x) = \frac{e^{f_{y_i}(x)}}{1 + e^{f_{y_i}(x)}}. \tag{8}$$

For the logistic form in (8), we can consider it as a softmax form with only 0 and $e^{f_{y_i}(x)}$ as the logits:

$$p(y_i = 1 \mid x) = \frac{e^{f_{y_i}(x)}}{e^0 + e^{f_{y_i}(x)}}. \tag{9}$$

Through the softmax form of (9), for each $i \in \{1, 2, ..., K\}$, the *energy function* of class y_i can be expressed as follows:

$$E_{y_i}(x) = -\ln(1 + e^{f_{y_i}(x)}). \tag{10}$$

Therefore, for each class $\{y_i\}_{i=1}^{K}$, we can derive a *label-wise joint energy function* as follows:

$$E_{joint}(x) = \sum_{i=1}^{K} -E_{y_i}(x) \tag{11}$$

In Equation (10), we consider the joint uncertainty among labels. Wang et al. [8] provided a theoretical foundation based on joint likelihood. Subsequent work by Zhang and Taneva-Popova [41], however, found that while Wang et al.'s approach assumed label independence, contrary to the initial beliefs of leveraging label independencies, joint energy indeed provides the optimal probabilistic approach to address the multi-label OOD problems. Moreover, Wang et al. [8] confirmed that utilizing multiple dominant labels to signal in-distribution inputs effectively captures data features, thus bypassing the need for direct computation and optimization in multi-label datasets. This approach also sidesteps the complexities associated with estimating joint likelihood through generative models, a notably challenging endeavor.

After deriving the *label-wise joint energy* in (11), we can utilize this method to detect the OOD uncertainty:

$$D(x; \tau) = \begin{cases} \text{out} & \text{if } E_{joint}(x) \leq \tau \\ \text{in} & \text{if } E_{joint}(x) > \tau \end{cases}, \tag{12}$$

where τ is the energy threshold. In our experimental setup, we defined $\tau = 95\%$ to ensure that $D(x; \tau)$ can correctly classify the majority of in-distribution data.

3.3 Spectral normalized joint energy

Based on the foundation laid by Section 3.2, we present **S**pectral **No**rmalized **Jo**int **E**nergy (SNoJoE). As part of the feature vector extraction process, spectral normalization is applied to the initial layers of the model. Through power iteration, we evaluate the spectral norm, guaranteeing that the weight matrices of the model adhere to *bi-Lipschitz constraint*.

Firstly, we need to ensure that the spectral norm of the weight matrices $g_l(x) = \sigma(W_l x + b)$ in the non-linear residual blocks of the network is less than 1, thereby ensuring:

$$\|g_l\|_{Lipschitz} \leq \|W_l x + b\|_{Lipschitz} \leq \|W_l\|_2 \leq 1. \tag{13}$$

To achieve this, we apply *spectral normalization* to constrain the weight matrices of the first L layers in the network:

$$W_l = \begin{cases} W_l/\sigma & 1 \leq l \leq L \\ W_l & l > L \end{cases}, \tag{14}$$

where σ is the spectral norm of the weight matrix, defined as the maximum singular value of the weight matrix. This singular value is obtained through singular value decomposition(SVD) of the weight matrix. As recommended in [42], *spectral normalization* is used to

enforce the weight matrices $\{W_l\}_{l=1}^{L}$ in (13) to be *Lipschitz-constrained*, ensuring that the hidden layer parameters $h_i(x)$ "*distance preserving*".

Bartlett et al. [43] demonstrates that consider a *hidden mapping* $h : \mathcal{X} \longrightarrow \mathcal{Y}$ with residual architecture $h = h_{l-1} \circ \cdots \circ h_2 \circ h_1(x)$ where $h_l(x) = x + g_l(x)$. If for $0 < \alpha \leq 1$, all g_l's are α-*Lipschitz*, *i.e.*, $\|g_l(x) - g_l(x')\|_Y \leq \alpha \|x - x'\|_X \quad \forall (x, x') \in \mathcal{X}$. Then:

$$Lips_{lower} * \|x - x'\|_X \leq \|h(x) - h(x')\|_Y \leq Lips_{upper} * \|x - x'\|_X , \tag{15}$$

where $Lips_{lower} = (1 - \alpha)^{L-1}$ and $Lips_{upper} = (1 + \alpha)^{L-1}$ are respectively the lower and upper bounds of *Lipschitz continuity*. Through the *bi-Lipschitz constraint*, the upper bound prevents overfitting during model gradient updates, ensuring the generalization and robustness of the model. The lower bound ensures that there is a certain distance maintained between input feature vectors, *i.e.*, $h(x)$ is *distance preserving*, thereby enabling the extraction of more generalizable features.

Combining the approach from Section 3.2, we now update the expression for $h_i(x)$ in (7):

$$h_i(x) = \begin{cases} \frac{W_{i-1}}{\sigma} \cdot h_{i-1}(x) & 2 \leq i \leq L \\ W_{i-1} \cdot h_{i-1}(x) & \text{otherwise} \end{cases} \tag{16}$$

Through the transformation of $h_i(x)$ in (16), the feature vectors can possess the property of "*distance preserving*" and replace $h_i(x)$ in (7) to complete the subsequent OOD uncertainty detection.

4 Experiments

In this section, we expound upon our experimental configuration (Section 4.1) and showcase the effectiveness of our approach across various out-of-distribution (OOD) evaluation tasks (Section 4.2). Furthermore, we delve into ablation studies and conduct comparative analyses, thereby fostering a deeper comprehension of distinct methodologies and ultimately contributing to an enhanced understanding of the field.

4.1 Setup

4.1.1 In-distribution datasets

We consider the PASCAL-VOC [44] and MS-COCO [45] as the in distribution multi-label dataset. MS-COCO comprises 82,783 images for training, 40,504 images for validation, and 40,775 images for testing, encompassing 80 commonly encountered object categories. PASCAL-VOC comprises 22,531 images of objects from 20 different categories such as *people, dogs, cars, etc.*, with detailed annotations provided. In this paper, We conduct the OOD detection task to evaluate the performance of our proposed method on this dataset.

4.1.2 Training details

In this study, the multi-label classifier trained is based on the ResNet-101 backbone architecture. The classifier is pretrained on ImageNet-1K [47], and the last layer is replaced by two fully connected layers. Spectral normalization is applied to the first 9 layers of the model. We utilize the Adam optimizer with standard parameters($\beta_1 = 0.9, \beta_2 = 0.999$), and the

initial learning rate during training is set to 1×10^{-4}. Data augmentation techniques such as random cropping and random flipping are employed during training to enhance the dataset, resulting in color images of size 256×256. After training, the mean Average Precision (mAP) on PASCAL-VOC is 89.19%, while on MS-COCO, it is 76.52%. The entire experimental process is conducted on NVIDIA RTX A6000.

Besides, to evaluate the impact of spectral regularization, we integrate it into our baseline model and compare its performance with our proposed method SNoJoE.

The model is trained with a standard Binary Cross-Entropy (BCE) loss, defined as:

$$\mathcal{L}_{\text{original}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log f(x_i;\theta) + (1-y_i)\log(1-f(x_i;\theta))] \tag{17}$$

We incorporate spectral regularization into the total loss function as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{original}} + \mathcal{L}_{\text{spec}} \tag{18}$$

The value of λ_{reg} in $\mathcal{L}_{\text{spec}}$ is chosen from $\{5 \times 10^{-5}, 1 \times 10^{-4}, 5 \times 10^{-4}\}$ based on a grid search strategy to find the best trade-off between regularization and model accuracy (Table 1).

4.1.3 Out-of-distribution datasets

To evaluate the performance of the model trained on the in-distribution dataset, we designate 20 classes from ImageNet-22K [47] and employ Texture dataset [48] as out-of-distribution (OOD) datasets. Following the evaluation protocol outlined in [8], we configure the ImageNet-22K dataset in a identical manner for evaluating the PASCAL-VOC pretrained model. The selected classes for evaluation encompass a diverse range, including *dolphin, deer, bat, rhino, raccoon, octopus, giant clam, leech, venus flytrap, cherry tree, Japanese cherry blossoms, redwood, sunflower, croissant, stick cinnamon, cotton, rice, sugar cane, bamboo, turmeric*.

4.1.4 Evaluation metrics

In our experiments, we employ commonly used evaluation metrics for OOD detection under multi-label settings: (i) the false positive rate (FPR95) of OOD examples is calculated when the true positive rate (TPR) of in-distribution examples is held constant at 95%; (ii) the area under the receiver operating characteristic curve (AUROC); (iii) the area under the precision-recall curve (AUPR).

Table 1 The dataset configuration in experiments

Dataset	Role	#Classes	#Instances
PASCAL-VOC [44]	In-Distribution (ID)	20	22,531
MS-COCO [45]	In-Distribution (ID)	80	82,783
ImageNet-22K [47]	Out-of-Distribution (OOD)	20 (out of 21841)	18,835
Texture [48]	Out-of-Distribution (OOD)	47	5,640

Table 2 The comparison of OOD detection performance using spectral normalized joint energy vs. competitive baselines

$\mathcal{D}_{\text{in}}$	PASCAL-VOC [44]			MS-COCO [45]		
OOD Score	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑
MaxLogit [50]	36.32	91.04	82.68	34.54	90.93	94.30
MSP [51]	69.85	78.24	67.93	77.92	72.43	83.34
ODIN [4]	36.32	91.04	82.68	34.58	90.26	93.69
Mahalanobis [22]	78.02	70.93	59.84	94.04	49.49	70.71
LOF [1]	76.71	67.54	55.35	74.30	74.87	85.82
Isolation Forest [52]	98.64	41.94	33.50	99.06	37.59	63.43
JointEnergy [8]	31.96	92.32	86.87	31.51	92.68	96.15
JointEnergy‡	30.29	93.19	87.93	29.24	93.38	96.59
SNoJoE(ours)	**28.49**	**93.48**	**88.11**	**27.97**	**93.91**	**96.92**

Note: JointEnergy‡ represents the variant of JointEnergy with spectral regularization applied
We use ResNet [49] to train on the in-distribution dataset and use ImageNet-22K (20 classes) as OOD dataset. All values are percentages. **Bold** numbers are superior results. ↑ indicates larger values are better, and ↓ indicates smaller values are better

4.2 Results

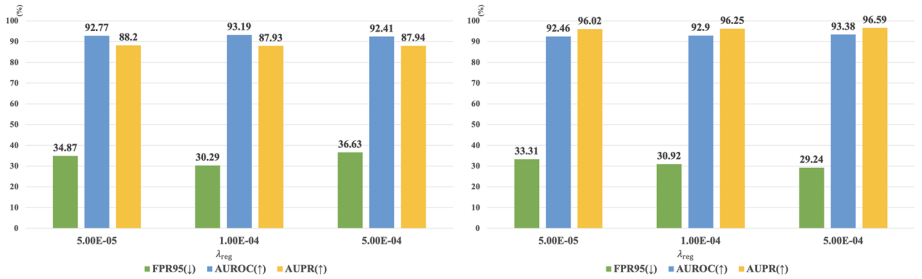
In Table 2, we compare our approach with leading OOD detection methods from the literature, showcasing SNoJoE as the new *state-of-the-art* benchmark. Our experimental design carefully selects methods based on pre-trained models to maintain fair comparison standards. Following the guidelines set forth in [8], we evaluated all metrics using the ImageNet dataset for OOD detection.

Additionally, as detailed in Section 4.1, we conducted further evaluations using the Texture dataset for OOD detection, with results presented in Table 3. Noteworthy, baseline methods like MaxLogit [50], Maximum Softmax Probability (MSP) [51], ODIN [4], and Mahalanobis [22] utilize statistics from the highest values across labels to calculate OOD scores. The Local Outlier Factor (LOF) [1] uses K-nearest neighbors (KNN) to assess local densities, identifying

Table 3 OOD detection performance using spectral normalized joint energy vs. competitive baselines on Texture [48] as the OOD dataset

$\mathcal{D}_{\text{in}}$	PASCAL-VOC [44]			MS-COCO [45]		
OOD Score	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑
MaxLogit [50]	12.36	96.22	96.97	14.63	96.10	99.32
MSP [51]	41.81	89.76	93.00	60.82	83.70	97.05
ODIN [4]	12.36	96.22	96.97	12.22	96.18	99.29
Mahalanobis [22]	19.17	96.23	97.90	44.61	85.71	97.41
LOF [1]	89.49	60.37	76.70	70.16	74.73	94.96
Isolation Forest [52]	99.59	20.89	50.11	95.55	53.21	90.45
JointEnergy [8]	10.87	96.78	97.87	12.82	96.84	99.54
JointEnergy‡	6.05	98.15	98.91	8.78	97.55	99.60
SNoJoE(ours)	**5.02**†	**98.48**†	**99.00**†	**7.36**†	**97.80**†	**99.64**†

Besides, † denotes that SNoJoE is statistically better (t-test with p-value < 0.01) than JointEnergy



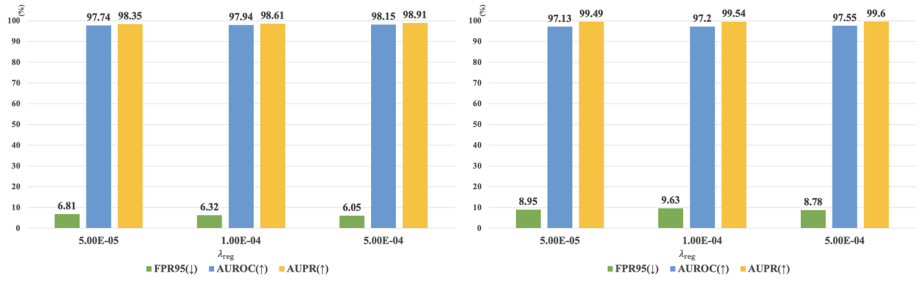
(a) PASCAL as the in-distribution dataset (b) MS-COCO as the in-distribution dataset

Figure 1 OOD detection performance of JointEnergy‡ while using different λ_{reg} on ImageNet-22K(20 classes) as the OOD dataset

OOD samples through their relatively lower densities compared to neighbors. The Isolation Forest method [52], a tree-based strategy, identifies anomalies by the path lengths from root to terminal nodes. JointEnergy [8] is an energy-based approach that detects OOD instances by evaluating the joint uncertainty among labels.

Moreover, we introduce an enhanced baseline by incorporating spectral regularization into the JointEnergy framework to further improve OOD detection performance. This approach applies spectral constraints to the weight matrices while computing joint energy, limiting the growth of the largest singular values to reduce overfitting to the ID data and enhance generalization to unseen distributions. By integrating spectral regularization, this method retains the core structure of JointEnergy while leveraging additional regularization effects to improve the separation between in-distribution and OOD samples. This enhanced baseline serves as a comparative method against SNoJoE, allowing us to assess the impact of spectral regularization across different OOD detection frameworks. As shown in Tables 2 and 3, experimental results demonstrate that this approach provides performance improvements, further validating the potential of spectral regularization in OOD detection tasks.

To further analyze the impact of spectral regularization, we evaluate its effectiveness under varying λ_{reg} values. As shown in Figures 1 and 2, results indicate that the choice of λ_{reg} significantly influences OOD detection performance. For both PASCAL-VOC and MS-COCO as in-distribution datasets, moderate regularization strengths ($\lambda_{reg} = 1 \times 10^{-4}$ or 5×10^{-4}) generally lead to the best performance, achieving lower FPR95 and higher AUROC/AUPR scores compared to weaker (5×10^{-5}) or stronger regularization settings.



(a) PASCAL as the in-distribution dataset (b) MS-COCO as the in-distribution dataset

Figure 2 OOD detection performance of JointEnergy‡ while using different λ_{reg} on Texture as the OOD dataset

Specifically, in the ImageNet-22K OOD evaluation, when using PASCAL-VOC as the in-distribution dataset, setting $\lambda_{\text{reg}} = 1 \times 10^{-4}$ achieves an FPR95 of 30.29%, outperforming both the weaker $\lambda_{\text{reg}} = 1 \times 10^{-4}$(34.87%) and the stronger $\lambda_{\text{reg}} = 5 \times 10^{-4}$(36.63%). Similarly, for MS-COCO, the optimal λ_{reg} setting yields an FPR95 of 29.24%, demonstrating a relative improvement over other configurations (33.31% and 30.92%). In terms of AUROC, the best-performing spectral regularization setup achieves 93.19% and 93.38% for PASCAL-VOC and MS-COCO. A similar trend is observed in the Texture OOD evaluation, where an appropriately tuned λ_{reg} yields a considerable reduction in false positive rates while maintaining robust detection performance. For PASCAL-VOC, the best λ_{reg} results in an FPR95 of 6.05%. In MS-COCO, the best setting leads to an FPR95 of 8.78%, improving upon the ablation baseline (9.22%). Additionally, AUROC and AUPR scores consistently improve with spectral regularization, peaking at 97.55% and 99.60% under optimal configurations. These results reinforce the importance of carefully selecting spectral regularization parameters to balance generalization and over-constraint effects, demonstrating that spectral regularization effectively enhances the separability of OOD samples across diverse datasets.

When conducting OOD detection on different ID and OOD datasets, SNoJoE outperforms several baseline methods across three evaluation metrics. Compared to JointEnergy, which performs OOD detection by utilizing label-wise joint energy, SNoJoE consistently outperforms baseline methods across different in-distribution datasets. When using PASCAL-VOC as the in-distribution dataset, SNoJoE achieves an 10.86% relative reduction of FPR95 on the subset of ImageNet-22K and a 53.82% relative reduction on the Texture dataset. Similarly, when using MS-COCO as the in-distribution dataset, SNoJoE achieves an 11.3% relative reduction on ImageNet-22K and a 42.58% reduction on Texture. Furthermore, compared to the spectral regularization-enhanced JointEnergy, which introduces spectral constraints to improve generalization, SNoJoE further enhances OOD detection performance. Specifically, with PASCAL-VOC as in-distribution, SNoJoE achieves a 5.94% reduction in FPR95 on ImageNet-22K and a 17.02% reduction on Texture. With MS-COCO, SNoJoE achieves a 4.34% reduction on ImageNet-22K and a 16.17% reduction on Texture. These improvements demonstrate the robustness of SNoJoE across different datasets, further validating the effectiveness of incorporating spectral normalization in OOD detection.

These improvements highlight the effectiveness of incorporating spectral normalization in a more structured manner, leading to better separation between in-distribution and OOD samples. Additionally, the enhancement in AUROC and AUPR further demonstrates the robustness of SNoJoE against diverse OOD datasets. These results suggest that spectral normalization, when integrated effectively, provides a substantial advantage in distinguishing OOD instances, surpassing both the standard JointEnergy and its spectral-regularized variant.

4.3 Ablation studies

In this section, we delve into a series of ablation experiments to further affirm that neural networks, when subjected to spectral normalization, exhibit a highly regularized feature space. This regularization, in turn, empowers them to identify generalizable features within the data more effectively, thereby enhancing their capability to accurately distinguish out-of-distribution (OOD) data. The observed performance improvement of SNoJoE over JointEnergy, as highlighted in Tables 2 and 3, underscores that spectral normalization plays a pivotal role in enabling the extraction of more generalizable features from image input space vectors. This enhancement bolsters the model's proficiency in recognizing OOD samples with greater effectiveness.

Table 4 Ablation study on the impact of the numbers of layers applied spectral normalization using ImageNet-22K (20 classes) as OOD dataset

$\mathcal{D}_{\text{in}}$	PASCAL-VOC			MS-COCO		
#layers	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑
0	31.37	93.37	**89.29**	29.29	93.09	96.31
7	48.19	89.59	84.19	30.74	93.10	96.42
8	30.85	92.98	87.24	**26.14**	93.74	96.64
9	**28.49**	**93.48**	88.11	27.97	**93.91**	**96.92**

#**layers** are numbers of layers applied spectral normalization

In conducting our ablation studies, we persist in utilizing JointEnergy [8] as the benchmark for comparison against our method, SNoJoE. This choice is motivated by the findings presented in Section 4.2, where JointEnergy emerged as the most proficient among competing methods, excluding ours. It's noteworthy that both JointEnergy and SNoJoE capitalize on the joint uncertainty between labels to facilitate OOD detection. For the training configurations and parameters, we adhere to the specifications outlined in Section 4.1.

Additionally, we explored the application of spectral normalization across various layers of the network structure to gauge its influence on multi-label OOD detection tasks. Our experimental findings, detailed in Tables 4 and 5, involved implementing spectral normalization at different levels within the ResNet framework [49] to assess its effect on OOD detection. The results suggest that indiscriminate use of spectral normalization could, in some cases, impair the model's ability to perform multi-label OOD detection effectively. Specifically, when spectral normalization is limited to the first seven layers of the network (refer to the second row of Tables 4 and 5), the model's efficacy may decline compared to a non-normalized version. This deterioration in performance might stem from the application of spectral normalization solely to the network's more superficial layers. Given that these initial layers process simpler representations, imposing stringent constraints on them could diminish the network's capacity for expressive representation, thereby undermining its performance. Conversely, extending spectral normalization to the model's deeper layers (as illustrated in the last two rows of Tables 4 and 5) appears to enhance the model's proficiency in learning and capturing intricate input vector features. This improvement is likely due to the advanced abstraction abilities of the deeper layers.

To evaluate the generality of this observation across architectures, we further conduct ablation studies on alternative backbone networks such as DenseNet. The results, presented in Appendix, demonstrate similar trends and reinforce our conclusion that spectral normalization, when applied to carefully chosen layers, offers a favorable balance between performance

Table 5 Ablation study on the impact of the numbers of layers applied spectral normalization using Texture [48] as OOD dataset

$\mathcal{D}_{\text{in}}$	PASCAL-VOC			MS-COCO		
#layers	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑
0	6.21	97.87	98.58	9.22	96.94	99.45
7	6.72	98.20	98.94	9.22	97.57	99.62
8	**4.91**	**98.49**	98.94	8.62	97.33	99.55
9	5.02	98.48	**99.00**	**7.36**	**97.80**	**99.64**

#**layers** are numbers of layers applied spectral normalization

and efficiency. However, this finding should not be misconstrued to suggest that greater application of spectral normalization invariably results in superior performance, as it also increases computational demands. Furthermore, the practicality of applying spectral normalization to certain layers (such as those involved in average pooling to decrease spatial dimensions of feature maps) remains questionable, given the negligible benefits it may offer.

In summary, our experiments reveal that indiscriminate use of spectral normalization across the network does not invariably enhance the model's performance and might even impair it. Nevertheless, if spectral normalization is judiciously applied to enable the network to more effectively learn complex and generalizable features from the input vectors, the model's performance surpasses that of models without spectral normalization. The performance discrepancy can reach as high as 2.88% and 1.30% in FPR95, with the OOD dataset being ImageNet-22K and Texture, respectively. Similarly, when using MS-COCO as the in-distribution dataset, the discrepancy increases to 3.15% on ImageNet-22K and 1.86% on Texture.

5 Conclusion

In this study, we introduce a cutting-edge method for OOD detection named Spectral Normalized Joint Energy (SNoJoE) in the context of multi-label classification.

Our findings reveal that spectral normalization applied to the initial layers of a pre-trained model's network significantly enhances model robustness, improves generalization capabilities, and more effectively distinguishes between in-distribution and out-of-distribution inputs. To further investigate the impact of spectral constraints, we introduce an enhanced baseline incorporating spectral regularization, which provides additional insights into the role of spectral properties in OOD detection. Experimental results demonstrate that SNoJoE consistently outperforms this enhanced baseline and other state-of-the-art approaches across multiple datasets, establishing itself as a new *state of the art* in this domain, while not substantially increasing computational demands.

We anticipate that our contribution will spark further exploration into multi-label OOD detection and encourage the expansion of this research area into wider applications.

Appendix: Extended ablation studies on spectral normalization scope

To supplement the results reported in Section 4.3 (Tables 4 and 5), we conducted further ablation studies to assess the generalizability of spectral normalization (SN) across architectures

Table 6 Ablation study on the impact of the numbers of layers applied spectral normalization

$\mathcal{D}_{\text{in}}$	PASCAL-VOC			MS-COCO		
#layers	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑
10	37.00	91.89	86.87	29.48	93.74	96.83
30	32.00	93.09	88.14	34.07	92.81	96.57
60	35.20	92.70	88.23	32.58	93.16	96.57
90	36.24	92.12	87.35	34.07	92.25	95.90
all	34.38	93.01	88.74	29.78	93.56	96.70

We use DenseNet121 to train on the in-distribution dataset and use ImageNet-22K(20 classes) as OOD dataset. **#layers** are numbers of layers applied spectral normalization

Table 7 Ablation study on the impact of the numbers of layers applied spectral normalization

$\mathcal{D}_{\text{in}}$	PASCAL-VOC			MS-COCO		
#layers	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑
10	6.19	98.20	98.91	7.59	97.67	99.60
30	5.34	98.25	98.93	7.41	97.73	99.63
60	7.34	97.89	98.82	7.91	97.73	99.62
90	8.23	97.63	98.60	9.13	97.41	99.57
all	5.07	98.42	99.06	7.84	97.73	99.63

We use DenseNet121 to train on the in-distribution dataset and use Texture as OOD dataset. **#layers** are numbers of layers applied spectral normalization

and configurations. These extended experiments focus on the scope of spectral normalization application, aiming to determine whether applying SN to more or all layers yields consistent benefits, and whether the insights from ResNet101 can be transferred to other backbones such as DenseNet.

Specifically, we evaluate three configurations for SN application: (1) applying SN to a larger number of layers than our original design; and (2) applying SN to all eligible layers throughout the network. Importantly, all other hyperparameters are kept fixed across experiments to ensure fair and controlled comparison.

Table 6 presents results on DenseNet121 with varying SN configurations. Table 7 shows similar experiments using Texture as OOD dataset, validating that the conclusions drawn from ResNet-based experiments hold more broadly. These results reinforce our core claim that targeted application of SN to early or structurally critical layers achieves a favorable trade-off between OOD performance and computational efficiency, whereas aggressive normalization across all layers may lead to marginal gains at best, and sometimes even degrade performance due to over-regularization.

Acknowledgements This work was supported by National Key R&D Program of China (No. 2021YFC3340700), NSFC grant (No. 62136002), Ministry of Education Research Joint Fund Project (8091B042239), Shanghai Knowledge Service Platform Project (No. ZF1213), and Shanghai Trusted Industry Internet Software Collaborative Innovation Center.

Author Contributions M.Y.H. and D.Z. designed the study and proposed the methodology. M.Y.H. conducted the experiments, implemented the SNoJoE model, and analyzed the results. M.Y.H. performed the comparative analysis with existing methods and drafted the initial manuscript. W.X.Y. and S.C.Z. designed the experiments and experimental framework for the enhanced baseline. W.X.L., D.Z., S.C.Z., and M.Y.H. designed the experiments for spectral regularization. All authors reviewed and revised the manuscript critically for important intellectual content and approved the final version for submission.

Funding National Key R&D Program of China (No. 2021YFC3340700), NSFC grant (No. 62136002), Ministry of Education Research Joint Fund Project (8091B042239), Shanghai Knowledge Service Platform Project (No. ZF1213), and Shanghai Trusted Industry Internet Software Collaborative Innovation Center.

Data Availability No datasets were generated or analysed during the current study.

Materials Availability Not applicable

Code Availability https://github.com/Nicholas-Mei/Ood_Detection_SNoJoE

Declarations

Ethics approval and consent to participate Not applicable

Consent for publication Not applicable

Conflict of interest Not applicable

Competing interests The authors declare no competing interests.

References

1. Breunig, M.M., Kriegel, H.-P., Ng, R.T., Sander, J.: Lof: identifying density-based local outliers. SIGMOD Rec. **29**(2), 93–104 (2000). https://doi.org/10.1145/335191.335388
2. Yang, X., Latecki, L.J., Pokrajac, D.: Outlier detection with globally optimal exemplar-based gmm. In: Proceedings of the 2009 SIAM international conference on data mining, SIAM, pp. 145–154 (2009)
3. Hsu, Y.-C., Shen, Y., Jin, H., Kira, Z.: Generalized ODIN: Detecting Out-of-distribution Image without Learning from Out-of-distribution Data (2020)
4. Liang, S., Li, Y., Srikant, R.: Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks (2020)
5. Bendale, A., Boult, T.: Towards Open Set Deep Networks (2015)
6. Reiss, T., Cohen, N., Bergman, L., Hoshen, Y.: PANDA: Adapting Pretrained Features for Anomaly Detection and Segmentation (2021)
7. Chaudhari, S., Shevade, S.: Learning from positive and unlabelled examples using maximum margin clustering. In: Neural Information Processing: 19th International Conference, ICONIP 2012, Doha, Qatar, November 12-15, 2012, Proceedings, Part III 19, Springer, pp. 465–473 (2012)
8. Wang, H., Liu, W., Bocchieri, A., Li, Y.: Can multi-label classification networks know what they don't know? (2021)
9. Liu, W., Wang, X., Owens, J.D., Li, Y.: Energy-based Out-of-distribution Detection (2021)
10. Hinz, T., Heinrich, S., Wermter, S.: Generating Multiple Objects at Spatially Distinct Locations (2019)
11. Gong, Y., Jia, Y., Leung, T., Toshev, A., Ioffe, S.: Deep Convolutional Ranking for Multilabel Image Annotation (2014)
12. Wei, Y., Xia, W., Lin, M., Huang, J., Ni, B., Dong, J., Zhao, Y., Yan, S.: Hcp: a flexible cnn framework for multi-label image classification. IEEE Trans. Pattern Anal. Mach. Intell. **38**(9), 1901–1907 (2016). https://doi.org/10.1109/TPAMI.2015.2491929
13. Ghamrawi, N., McCallum, A.: Collective multi-label classification. In: Proceedings of the 14th ACM international conference on information and knowledge management, pp. 195–200 (2005)
14. Chen, L.-C., Schwing, A.G., Yuille, A.L., Urtasun, R.: Learning Deep Structured Models (2015)
15. Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., Xu, W.: CNN-RNN: A Unified Framework for Multi-label Image Classification (2016)
16. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing and Mining (IJDWM) **3**(3), 1–13 (2007)
17. Chen, L., Zhan, W., Tian, W., He, Y., Zou, Q.: Deep integration: a multi-label architecture for road scene recognition. IEEE Trans. Image Process. **28**(10), 4883–4898 (2019). https://doi.org/10.1109/TIP.2019.2913079
18. Zhang, H., Li, A., Guo, J., Guo, Y.: Hybrid Models for Open Set Recognition (2020)
19. Ayadi, A., Ghorbel, O., Obeid, A.M., Abid, M.: Outlier detection approaches for wireless sensor networks: A survey. Comput. Netw. **129**, 319–333 (2017)
20. Ranshous, S., Shen, S., Koutra, D., Harenberg, S., Faloutsos, C., Samatova, N.F.: Anomaly detection in dynamic networks: a survey. Wiley Interdisciplinary Rev.: Comput. Stat. **7**(3), 223–247 (2015)
21. Wei, H., Xie, R., Cheng, H., Feng, L., An, B., Li, Y.: Mitigating Neural Network Overconfidence with Logit Normalization (2022)
22. Lee, K., Lee, K., Lee, H., Shin, J.: A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks (2018)
23. Huang, R., Geng, A., Li, Y.: On the Importance of Gradients for Detecting Distributional Shifts in the Wild (2021)
24. Tax, D.M.J.: One-class classification: Concept learning in the absence of counter-examples. (2002)

25. Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S.A., Binder, A., Müller, E., Kloft, M.: Deep one-class classification. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 4393–4402. PMLR, (2018). https://proceedings.mlr.press/v80/ruff18a.html
26. Wang, J., Cherian, A.: GODS: Generalized One-class Discriminative Subspaces for Anomaly Detection (2019)
27. He, F., Liu, T., Webb, G.I., Tao, D.: Instance-Dependent PU Learning by Bayesian Optimal Relabeling (2020)
28. Menon, A., Rooyen, B.V., Ong, C.S., Williamson, B.: Learning from corrupted binary labels via class-probability estimation. In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 37, pp. 125–134. PMLR, Lille, France (2015). https://proceedings.mlr.press/v37/menon15.html
29. Scott, C.: A Rate of Convergence for Mixture Proportion Estimation, with Application to Learning from Noisy Labels. In: Lebanon, G., Vishwanathan, S.V.N. (eds.) Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 38, pp. 838–846. PMLR, San Diego, California, USA (2015). https://proceedings.mlr.press/v38/scott15.html
30. Zhong, J.-X., Li, N., Kong, W., Liu, S., Li, T.H., Li, G.: Graph Convolutional Label Noise Cleaner: Train a Plug-and-play Action Classifier for Anomaly Detection (2019)
31. Zolfi, A., Amit, G., Baras, A., Koda, S., Morikawa, I., Elovici, Y., Shabtai, A.: YolOOD: Utilizing Object Detection Concepts for Multi-Label Out-of-Distribution Detection (2023). https://arxiv.org/abs/2212.02081
32. Ackley, D.H., Hinton, G.E., Sejnowski, T.J.: A learning algorithm for boltzmann machines. Cogn. Sci. **9**(1), 147–169 (1985)
33. LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., Huang, F.: A tutorial on energy-based learning. Predicting Structured Data **1**(0) (2006)
34. Ranzato, M., Poultney, C., Chopra, S., Cun, Y.: Efficient learning of sparse representations with an energy-based model. Advan. Neural inform. Process. Syst. **19** (2006)
35. Xie, J., Lu, Y., Zhu, S.-C., Wu, Y.N.: A Theory of Generative ConvNet (2016)
36. Tu, L., Gimpel, K.: Learning Approximate Inference Networks for Structured Prediction (2018)
37. Lin, Z., Roy, S.D., Li, Y.: MOOD: Multi-level Out-of-distribution Detection (2021)
38. Morteza, P., Li, Y.: Provable Guarantees for Understanding Out-of-distribution Detection (2021)
39. Zhang, W., Yan, J., Wang, X., Zha, H.: Deep Extreme Multi-label Learning (2018)
40. Liu, S.M., Chen, J.-H.: A multi-label classification based approach for sentiment classification. Expert Syst. Appl. **42**(3), 1083–1093 (2015)
41. Zhang, D., Taneva-Popova, B.: A theoretical analysis of out-of-distribution detection in multi-label classification. In: Proceedings of the 2023 ACM SIGIR international conference on theory of information retrieval, pp. 275–282 (2023)
42. Behrmann, J., Grathwohl, W., Chen, R.T.Q., Duvenaud, D., Jacobsen, J.-H.: Invertible Residual Networks (2019)
43. Bartlett, P.L., Evans, S.N., Long, P.M.: Representing smooth functions as compositions of near-identity functions with implications for deep network optimization (2018)
44. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. Int. J. Comput. Vision **111**, 98–136 (2015)
45. Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft COCO: Common Objects in Context (2015). arXiv:1405.0312
46. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017). https://doi.org/10.1145/3065386
47. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on computer vision and pattern recognition, IEEE, pp. 248–255 (2009)
48. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing Textures in the Wild (2013)
49. He, K., Zhang, X., Ren, S., Sun, J.: Identity Mappings in Deep Residual Networks (2016)
50. Chan, R., Lis, K., Uhlemeyer, S., Blum, H., Honari, S., Siegwart, R., Fua, P., Salzmann, M., Rottmann, M.: SegmentMeIfYouCan: A Benchmark for Anomaly Segmentation (2021)
51. Hendrycks, D., Gimpel, K.: A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks (2018)
52. Liu, F.T., Ting, K.M., Zhou, Z.-H.: Isolation forest. In: 2008 Eighth IEEE international conference on data mining, pp. 413–422 (2008). https://doi.org/10.1109/ICDM.2008.17