Information Sciences 705 (2025) 121989

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Out-of-distribution detection with non-semantic exploration

Zhen Fang, Jie Lu *, Guangquan Zhang

Australian Artificial Intelligence Institute, University of Technology Sydney, P.O. Box 123, Broadway NSW, Australia

ARTICLE INFO

Keywords:
Out-of-distribution detection
Out-of-distribution generalization
Machine learning

ABSTRACT

Out-of-distribution (OOD) detection is crucial in modern deep learning applications, as it can identify OOD data drawn from distributions differing from those of the in-distribution (ID) data. Advanced OOD detection methods primarily rely on post-hoc strategies, which identify OOD data by analyzing the predictions of a model well-trained on ID data. However, deep models are known to be impacted by spurious features such as backgrounds, causing existing OOD detection methods to fail in identifying OOD data that share the same spurious features as ID data. Therefore, this paper studies how to mitigate spurious features to improve OOD detection. To address this challenge, we propose a novel method called Non-semantic Exploration OOD Detection (NsED), which focuses on exploring and exploiting non-semantic features. In particular, NsED first explores non-semantic features in an OOD generalization manner. These non-semantic features are then used to train deep models to be more robust against spurious features. Through extensive experiments on representative benchmarks, we show that NsED significantly and consistently improves the detection performance of many representative post-hoc OOD detection methods.

1. Introduction

Deep learning has emerged as a prominent approach to address machine learning problems due to its remarkable performance across various applications [1]. However, the success of deep learning heavily relies on an underlying in-distribution (ID) assumption, whereby the training data (training ID data) and test data share the same distribution. In real-world scenarios, the test data might include out-of-distribution (OOD) data, which possess semantic labels that are distinct from those of ID data [2] (see Fig. 1). These OOD observations violate the ID assumption, leading to concerns regarding safety and reliability [3]. To tackle the issue caused by OOD data, which is a realistic problem, researchers have extensively studied a specific task called *out-of-distribution detection* [4]. In OOD detection, the deep model's predictor should accurately predict ID data while also identifying OOD data.

Many representative OOD detection methods are based on the post-hoc strategies [5,6], which employ scoring functions to analyze predictions of deep models that are well-trained on ID data, in order to identify OOD data. Therefore, the effectiveness of the post-hoc strategies largely depends on the quality of the extracted features from the deep model. Previous studies have indicated that deep models are prone to extracting spurious features, which may predict labels without being causally related to them [7]. For instance, neural networks trained on ImageNet have been observed to rely heavily on background features [8] that are often correlated with labels but are not causally significant. This phenomenon can also result in decreased detection performance as the correlation between ID labels and these spurious features grows [9,10], see Fig. 2.

[*] Corresponding author.
E-mail addresses: zhen.fang@uts.edu.au (Z. Fang), jie.lu@uts.edu.au (J. Lu), guangquan.zhang@uts.edu.au (G. Zhang).

https://doi.org/10.1016/j.ins.2025.121989
Received 1 June 2023; Received in revised form 14 February 2025; Accepted 15 February 2025

Available online 19 February 2025

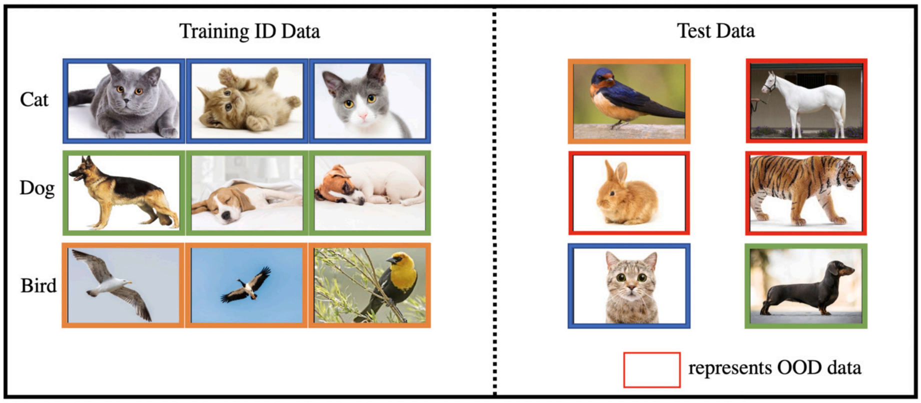

Fig. 1. In OOD detection, the test data contain OOD data that have different semantics compared to the training ID data.

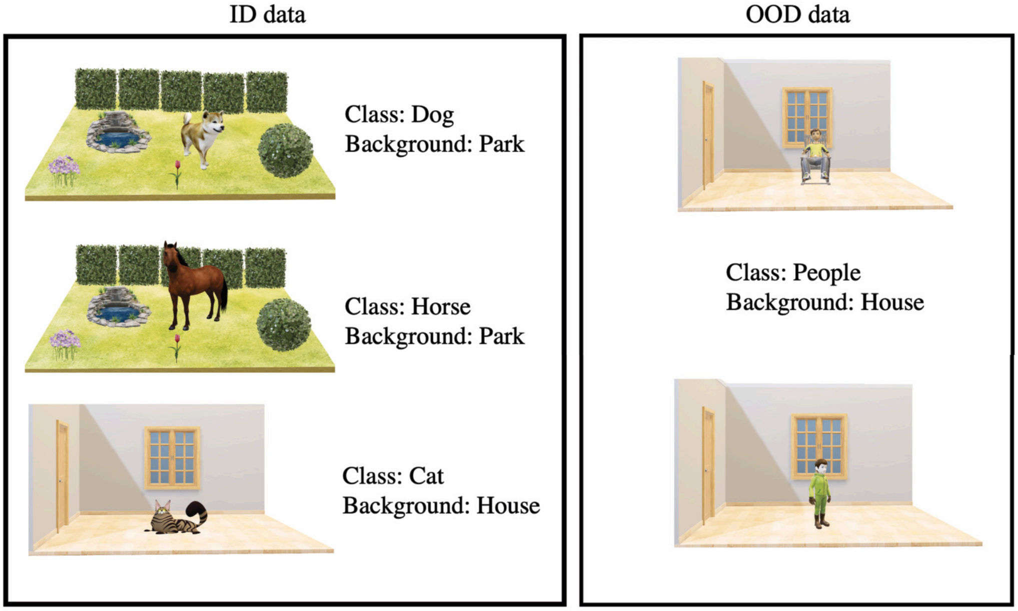

Fig. 2. An illustration on the negative effectiveness of spurious features. Left figure: ID data consist of dog, horse and cat images, whose backgrounds are park and house. Right figure: OOD data consist of people images, whose backgrounds are house. In this case, one can classify dog and cat according to the backgrounds park and house. However, if using the background house as the features to detect OOD data, then the OOD data will be recognized as ID data. The background house is the spurious features to OOD data.

To tackle the spurious correlation issue and improve detection performance, a promising strategy is to learn semantic-invariant representations, inspired by out-of-distribution (OOD) generalization methods [11]. These methods aim to train deep models capable of effectively generalizing to new data that share the same semantics as the training data. By leveraging well-generalized models with robust features, we expect that the existing post-hoc OOD detection methods can be further enhanced.

Accordingly, we propose a novel method called Non-semantic Exploration OOD Detection (NsED) to generate non-semantic features and then train deep models using these features. Ideally, to enhance the robustness of deep models against spurious features, we should employ multiple training datasets with identical semantics but varying styles (i.e., non-semantic factors). However, it is typically difficult to collect such a set of datasets. Instead, we employ a style transfer strategy [12] to generate multiple training ID datasets with different non-semantic factors, based on the original ID dataset. Training with these data encourages the learning of semantic-invariant features, resulting in improved robustness against spurious features and detection performance. Extensive experiments conducted on standard OOD detection benchmarks consistently demonstrate that the extracted features can enhance representative post-hoc OOD detection methods [5,6,13–15]. Main contributions are presented as follows:

- This paper proposes a novel method, termed as *Non-semantic Exploration OOD detection* (NsED), to address the spurious issue by leveraging out-of-distribution generalization techniques. Specifically, NsED explores non-semantic factors to learn semantic-invariant features against spurious features, leading to significant improvements in existing representative OOD detection methods [5,6,13–15].

- This paper is the **first** to explore the relationship between single-domain OOD generalization and OOD detection. By designing novel and effective single-domain OOD generalization methods, we can enhance the generalization ability of deep models. Experiments have indicated that representative post-hoc OOD detection methods can be significantly improved by using more generalized models. These findings verify that OOD generalization can enhance the detection ability of deep models.
- This paper conducts extensive experiments and parameter analysis to analyze the impact of different components of our proposed method NsED. The experiments show that NsED is robust to different types of OOD data. The parameter analysis also demonstrates the effectiveness of each component of NsED in improving OOD detection performance.

The paper is organized as follows: Section 2 reviews related work on post-hoc OOD detection, spurious correction issue in OOD detection and OOD generalization. Section 3 presents the definitions, important notations and our problem. Section 4 lists the basic challenge and key motivation of our research. Section 5 introduces the details of the proposed method. Comprehensive evaluation results and analyses are provided in Section 6. Lastly, Section 7 concludes the paper.

2. Related work

In this section, we primarily discuss post-hoc OOD detection methods, the issue of spurious correlations in OOD detection, and advanced OOD generalization techniques. We will also explore the relationship between our proposed method and OOD generalization.

Post-hoc OOD detection. Maximum Softmax Probability (MSP) is a pioneering method for OOD detection, proposed by Hendrycks et al. [16]. This method utilizes the maximum probability of model prediction to detect OOD data. ODIN [13] enhances MSP by applying temperature scaling to the softmax function and employing input processing techniques, resulting in stronger separability between ID and OOD data. Motivated by energy-based models [17], another method for OOD detection called Energy-Based OOD Detection [5] proposes the use of a free energy scoring function to differentiate ID and OOD data. These methods primarily rely on the predictive information obtained from model outputs. Advanced studies have also focused on exploring the latent ID feature representations and gradient information of deep models. [6] delves into the gradient information of deep models and designs a gradient-based scoring function called GradNorm. React [14], which stands for Rectified Activations, clips noisy activations in the classifier's penultimate layer, thereby strengthening the energy score. K-nearest neighbor-based (KNN) OOD detection [18] utilizes the geometric information of extracted ID features and designs a distance-based score. KL-Matching (KLM) [15] explores the relationship between softmax predictions and mean class-conditional distributions using KL divergence. Overall, the post-hoc OOD detection methods heavily depend on the feature representations.

Spurious correction issue in OOD detection. Previous studies have indicated that deep models are prone to extracting spurious features, which may predict labels without being causally related to them [7]. The spurious collection issue also matters in OOD detection. Recently, [9] discusses the impact of spurious correlation on OOD detection, and presents a new formalization to model the data shifts by considering both the invariant and environmental (spurious) features. Their results suggest that the detection performance significantly worsens when the correlation between spurious features and labels increases in the training set. A related study by [10] explores the impact of the spurious correlation issue on detection performance from a theoretical perspective. It illustrates why models that are robust to spurious correlations can yield better detection performance. Both works [9,10] emphasize the importance of addressing the spurious correlation issue in OOD detection.

OOD generalization. Out-of-distribution generalization focuses on training procedures that ensure the resulting models can generalize well on unknown target data that share the same semantics as the training data. There have been numerous efforts devoted to different directions such as causal learning [19], distribution robustness [20], and invariant representation [21]. [22] can be regarded as one of the pioneering OOD generalization works. They revise the classical empirical risk minimization principle and propose to minimize invariant risk for extracting the semantic features. Although the invariant risk minimization (IRM) framework has elegant theoretical supports, some additional assumptions on the learning models are indispensable, resulting in limited performance [23]. Motivated by IRM [22], [24] further extend the IRM framework by considering risk extrapolation and nonlinear prediction functions. There also exist works that do not follow the IRM framework. For example, the work [25] leverages the gradient of the domain's predictor to perturb input data in the direction of the most significant domain change, while ensuring that the semantics are preserved. Other studies [26] have explored style-transfer strategies to generate training data with different styles but similar semantics. Overall, the style-augmented data enable more effective extraction of semantic features, thus the main focus of our paper.

Connection between our method and OOD generalization. Our method, termed NsED, is inspired by out-of-distribution generalization, aiming to extract semantic-invariant features from datasets sharing the same semantics but differing in style. It is worth noting that the majority of OOD generalization methods [19,27] demand multiple diverse training datasets, which poses a challenge in cases where only one dataset is accessible, as in OOD detection. Hence, it is infeasible to use representative OOD generalization methods to tackle our problem directly. Inspired by style transfer strategy [12], we generate training ID datasets with distinct non-semantic factors. Subsequently, training models using these data would incentivize models to acquire semantic-invariant features, enhancing the model's robustness against spurious features.

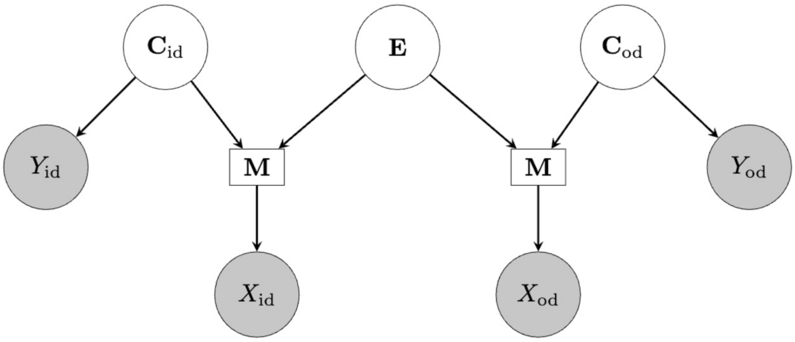

Fig. 3. This figure shows the causal graph of feature random variables $X_{\rm id}, X_{\rm od}$, label variables $Y_{\rm id}, Y_{\rm od}$, semantic variables $\mathbf{C}_{\rm id}, \mathbf{C}_{\rm od}$ and non-semantic variable $\mathbf{E}$, and presents that $X_{\rm id}$ and $X_{\rm od}$ are generated by an invertible causal mechanism $\mathbf{M}$, i.e., $X_{\rm id} = \mathbf{M}(\mathbf{C}_{\rm id}, \mathbf{E})$, $X_{\rm od} = \mathbf{M}(\mathbf{C}_{\rm od}, \mathbf{E})$, $\mathbf{C}_{\rm id}, \mathbf{E} = \mathbf{M}^{-1}(X_{\rm id})$ and $\mathbf{C}_{\rm od}, \mathbf{E} = \mathbf{M}^{-1}(X_{\rm od})$.

3. Preliminaries

In this section, we mainly introduce some necessary concepts. Let $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} = \{1, ..., K\}$ be the feature space and the ID label space, respectively.

Random variables and distributions. Denote $X_{\rm id} \in \mathcal{X}$ and $X_{\rm od} \in \mathcal{X}$ the feature random variables corresponding to ID and OOD data, respectively. $Y_{\rm id} \in \mathcal{Y}$ and $Y_{\rm od} \notin \mathcal{Y}$ are the label random variables corresponding to ID and OOD data, respectively. We use $P_{X_{\rm id}, Y_{\rm id}}(\mathbf{x}, y)$ to represent the ID joint distribution and use $P_{X_{\rm od}, Y_{\rm od}}(\mathbf{x}, y)$ to represent the OOD joint distribution. Then $P_{X_{\rm id}}(\mathbf{x})$ is the ID marginal distribution and $P_{X_{\rm od}}(\mathbf{x})$ is the OOD marginal distribution.

Out-of-distribution detection. In OOD detection [4], we target on learning OOD detector G, such that for any test data $\mathbf{x}$: 1) if $\mathbf{x}$ is drawn from $P_{X_{\rm id}}(\mathbf{x})$, then G can classify $\mathbf{x}$ into correct ID classes; 2) if $\mathbf{x}$ is drawn from $P_{X_{\rm od}}(\mathbf{x})$, then G can detect $\mathbf{x}$ as OOD data.

Post-hoc strategies. Many representative OOD detection methods [6,5,13] adopt the post-hoc strategies. Therein, given a threshold λ, a pre-trained ID model $\mathbf{f}$ and a scoring function S, then $\mathbf{x}$ is detected as ID data if and only if $S(\mathbf{x}; \mathbf{f}) \geq \lambda$:

$$G_\lambda(\mathbf{x}) = \text{ID}, \text{ if } S(\mathbf{x}; \mathbf{f}) \geq \lambda; \text{ otherwise, } G_\lambda(\mathbf{x}) = \text{OOD}. \tag{1}$$

The effectiveness of post-hoc OOD detection is largely dependent on the design of S and the ID model $\mathbf{f}$ such that the scores assigned to OOD data are lower than those of the ID data.

Causal assumption. Fang et al., [4] has developed learning theory to prove that OOD detection cannot be addressed well without further assumptions. Especially, when the ID and OOD data share the spurious features, [9] provide provable evidence for the existence of OOD data with high confidence such that it cannot be differentiated from ID data. Therefore, to address the OOD detection issue, it is imperative to rely on practical assumptions. Following OOD generalization works [28], we assume that the random variables $X_{\rm id}$ and $X_{\rm od}$ are generated by an *invertible* causal mechanism $\mathbf{M}$ with three causes: ID semantic variable $\mathbf{C}_{\rm id}$, OOD semantic variable $\mathbf{C}_{\rm od}$ and non-semantic variable $\mathbf{E}$:

$$X_{\rm id} = \mathbf{M}(\mathbf{C}_{\rm id}, \mathbf{E}) \text{ and } X_{\rm od} = \mathbf{M}(\mathbf{C}_{\rm od}, \mathbf{E}),$$

and ID label variable $Y_{\rm id}$ and OOD label variable $Y_{\rm od}$ are the effects of the variables $\mathbf{C}_{\rm id}$ and $\mathbf{C}_{\rm od}$, respectively: $Y_{\rm id} \leftarrow \mathbf{C}_{\rm id}$ and $Y_{\rm od} \leftarrow \mathbf{C}_{\rm od}$. See Fig. 3 for the causal graph.

4. Challenge and motivation

In this section, we present the main challenge in the classical OOD detection learning framework. Additionally, we discuss our underlying motivation for explaining how we address and mitigate this challenge.

Classical learning strategy. The representative post-hoc OOD detection methods [5] learn the pre-trained deep models based on the risk minimization principle: for any model $\mathbf{g}_{\mathbf{w}} : \mathcal{X} \to \mathbb{R}^K$ with parameter $\mathbf{w} \in \mathcal{W}$,

$$\mathbf{w}^* \in \mathop{\arg\min}_{\mathbf{w} \in \mathcal{W}} \mathbb{E}_{(\mathbf{x}, y) \sim (X_{\rm id}, Y_{\rm id})} \left[\ell(\mathbf{g}_{\mathbf{w}}(\mathbf{x}), y) \right]. \tag{2}$$

Due to finite training data, the empirical risk minimization (ERM) principle [29] is employed as a substitute for the risk minimization: given ID training data $\mathcal{D}_{\rm id} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, for any model $\mathbf{g}_{\mathbf{w}} : \mathcal{X} \to \mathbb{R}^K$ with parameter $\mathbf{w} \in \mathcal{W}$,

Algorithm 1: Semantic exploration OOD detection.

```
Input  : Training ID data $\mathcal{D}_{\rm id}$, inner iterative step $L$, inner learning rate $l$, hyper-parameters $\alpha, \gamma$, number of styles $m$, batch size $q$, causal mechanism
         $\mathbf{M}(\mathbf{P}(\cdot), \mathbf{A}(\cdot; \cdot, \gamma))$, learning model $\mathbf{g}_{\mathbf{w}}$, post-hoc OOD detection method $\mathcal{A}$ with scoring function $S$.
Output: Learned model $\mathbf{g}_{\mathbf{w}}$ and OOD detector $G$.
Initialize model $\mathbf{g}_{\mathbf{w}}$;
Repeat
    Sampling mini-batch $\mathcal{D}_b = \{(\mathbf{x}_{b_1}, y_{b_1}), ..., (\mathbf{x}_{b_q}, y_{b_q})\}$ from $\mathcal{D}_{\rm id}$;
    For $i = 1$ to $q$
        Initialize $\boldsymbol{\beta}$ as $\boldsymbol{\beta}_0$ and compute $\mathbf{x}_{b_i}^{\boldsymbol{\beta}_0, \gamma}$ by Eq. (9);
        For $k = 1$ to $L$
            Update $\boldsymbol{\beta}_k$ by $\boldsymbol{\beta}_{k-1}$ and Eq. (10);
            Update $\mathbf{x}_{b_i}^{\boldsymbol{\beta}_k, \gamma}$ by $\boldsymbol{\beta}_k$ and Eq. (9);
        End for
    End for
    Compute $\mathcal{L}(\mathbf{g}_{\mathbf{w}}; \mathcal{D}_b, \alpha, \gamma, L)$ by Eq. (12);
    Update $\mathbf{w} \leftarrow \mathbf{w} - {\rm lr}\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{g}_{\mathbf{w}}; \mathcal{D}_b, \alpha, \gamma, L)$;
Until convergence;
Obtain the learned model $\mathbf{g}_{\mathbf{w}}$ and learn OOD detector $G$ by Eq. (1).
```

$$\widehat{\mathbf{w}} \in \mathop{\arg\min}_{\mathbf{w} \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{g}_{\mathbf{w}}(\mathbf{x}_i), y_i). \tag{3}$$

In Eq. (2), $\frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{g}_{\mathbf{w}}(\mathbf{x}_i), y_i)$ is known as the empirical risk w.r.t. the loss function ℓ. Statistical learning theory [29] has indicated that under suitable assumptions, the risk w.r.t. empirical optimal solution $\widehat{\mathbf{w}}$ will get close to the risk w.r.t. optimal solution $\mathbf{w}^*$. However, the empirical optimal solution $\widehat{\mathbf{w}}$ does not mean that it is the optimal solution to OOD detection, due to the interference of non-semantic variable $\mathbf{E}$.

Challenge in classical learning strategy. The efficacy of post-hoc strategies is closely tied to the quality of features extracted by deep models trained using ERM principle, i.e., Eq. (3). Existing literature suggests that these deep models are susceptible to capturing spurious features—features that can predict labels but lack causal relevance [7]. For example, studies on models trained with the ImageNet dataset reveal an over-reliance on background features, which, although correlated with labels, don't necessarily have causal significance [8]. This tendency towards spurious features can adversely impact model performance, particularly in terms of detection capabilities, as the correlation between ID labels and these irrelevant features grows [9].

Motivation. To address the issue of spurious correlation, we adopt the causal learning framework (see Fig. 3). Within this framework, our main objective is to mitigate the negative impact of non-semantic variables. To accomplish this, we draw inspiration from OOD generalization methods that aim to train deep models capable of effectively generalizing to new data with the same semantics as the training data. By employing a well-generalized model with robust features, we anticipate that existing post-hoc OOD detection methods can be further improved. Ideally, to equip deep models with resilience against spurious features, we should utilize multiple training datasets with identical semantics but varying non-semantic factors. However, it is often challenging to collect such datasets. As an alternative, we employ a style transfer strategy [12] to generate multiple training ID datasets with different non-semantic factors, based on the original ID dataset. By training with these datasets, we encourage the learning of semantic-invariant features, thereby enhancing robustness against spurious features and improving detection performance. In Section 5, we present novel method NsED to migitate the spurious correlation issue.

5. Proposed methodology

In this section, we introduce the proposed method Non-semantic Exploration OOD Detection (NsED). The pseudo code is summarized in Algorithm 1.

5.1. *Proposed learning strategy*

Here, we mainly introduce the details of our proposed learning strategy according to Section 4.

Learning strategy to semantic robustness. Motivated by distribution robustness [30], we explore the risk minimization under the worst-case non-semantic factor $\mathbf{e} \sim \mathbf{E}$: for any model $\mathbf{g}_{\mathbf{w}} : \mathcal{X} \to \mathbb{R}^K$ with parameter $\mathbf{w} \in \mathcal{W}$,

$$\mathbf{w}^* \in \mathop{\arg\min}_{\mathbf{w} \in \mathcal{W}} \mathbb{E}_{(\mathbf{c}, y) \sim (\mathbf{C}_{\rm id}, Y_{\rm id})} \max_{\mathbf{e} \sim \mathbf{E}} \left[\ell(\mathbf{g}_{\mathbf{w}} \circ \mathbf{M}(\mathbf{c}, \mathbf{e}), y)\right], \tag{4}$$

where $\mathbf{M}$ is the causal mechanism and $\mathbf{C}_{\rm id}$ is the ID semantic variable introduced in Section 3. Eq. (4) indicates that the optimal classifier $\mathbf{g}_{\mathbf{w}^*}$ can maintain semantic robustness even when there are changes in the non-semantic factor $\mathbf{e} \sim \mathbf{E}$.

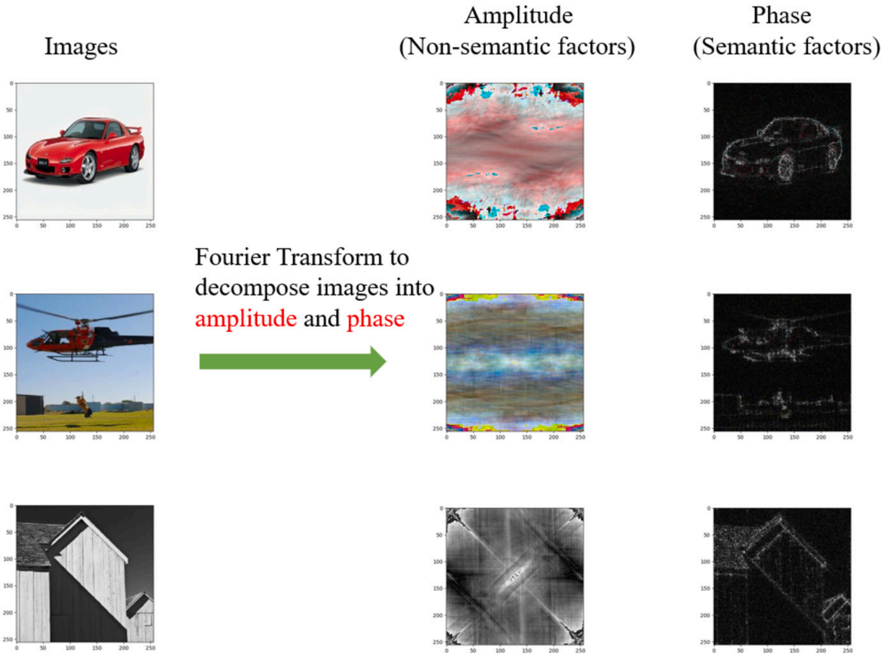

Fig. 4. Using the discrete Fourier transform to decompose an image $\mathbf{x}$ into amplitude $\mathbf{A}(\mathbf{x})$ (non-semantic factor) and phase $\mathbf{P}(\mathbf{x})$ (semantic factor).

Since the standard benchmarks for OOD detection only provide ID training data with limited non-semantic factors $\mathbf{e}$, our focus is on generating novel and diverse non-semantic factors to reconstruct data. Given ID training data $\mathcal{D}_{\text{id}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, we use $\mathcal{E}_i$ (the details of $\mathcal{E}_i$ are presented in Eq. (8)) to present the generated non-semantic space w.r.t. $\mathbf{x}_i$, which consists of different non-semantic factors $\mathbf{e}$, and we also use $\mathbf{c}_i$ to present the corresponding semantic factor of $\mathbf{x}_i$. Then, the empirical form of Eq. (4) can be replaced with the following optimization problem:

$$\widehat{\mathbf{w}} \in \arg\min_{\mathbf{w}\in\mathcal{W}} \frac{1}{n}\sum_{i=1}^{n} \max_{\mathbf{e}\in\mathcal{E}_i} \left[\ell(\mathbf{g}_{\mathbf{w}} \circ \mathbf{M}(\mathbf{c}_i,\mathbf{e}), y)\right]. \tag{5}$$

Now, we introduce ERM in Eq. (3) to Eq. (5) as a regularization to balance the semantic robustness and ID classification performance:

$$\widehat{\mathbf{w}} \in \arg\min_{\mathbf{w}\in\mathcal{W}} \Big[\frac{1-\alpha}{n}\sum_{i=1}^{n} \max_{\mathbf{e}\in\mathcal{E}_i} \left[\ell(\mathbf{g}_{\mathbf{w}} \circ \mathbf{M}(\mathbf{c}_i,\mathbf{e}), y_i)\right] + \frac{\alpha}{n}\sum_{i=1}^{n} \ell(\mathbf{g}_{\mathbf{w}}(\mathbf{x}_i), y_i)\Big], \tag{6}$$

where $\alpha \in [0,1]$ is the hyperparameter. By optimizing Eq. (6), we can learn the classifier $\mathbf{g}_{\widehat{\mathbf{w}}}$, which not only accurately predicts the ID data but also maintains semantic robustness despite the shift of the non-semantic factor $\mathbf{e}$ from $\mathcal{E}_i$. Next, we introduce how to construct the non-semantic space $\mathcal{E}_i$.

Exploring non-semantic space and causal mechanism. To generate diverse non-semantic factors and simulate the causal mechanism, we utilize the Fourier-based style transfer strategy [12]. This method compels the model to capture both phase information (semantic information) and amplitude information, which enables the generation of non-semantic factors by linearly interpolating among the amplitude spectra of different images. Specifically, [12] use the discrete Fourier transform to decompose an image $\mathbf{x}$ into amplitude $\mathbf{A}(\mathbf{x})$ and phase $\mathbf{P}(\mathbf{x})$ (see Fig. 4). The amplitude $\mathbf{A}(\mathbf{x})$ and the phase $\mathbf{P}(\mathbf{x})$ can be regarded as the non-semantic factor $\mathbf{e}$ and the semantic factor $\mathbf{c}$ of data point $\mathbf{x}$. By given m data points $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_m$ from training data $\mathcal{D}_{\text{id}}$, the novel style can be created:

$$\mathbf{A}(\mathbf{x}_i; \boldsymbol{\beta}, \gamma) = \gamma\left[\beta_0\mathbf{A}(\mathbf{x}_i) + \beta_1\mathbf{A}(\tilde{\mathbf{x}}_1) + \ldots + \beta_m\mathbf{A}(\tilde{\mathbf{x}}_m)\right] + (1-\gamma)\mathbf{A}(\mathbf{x}_i), \tag{7}$$

where $0 \le \gamma \le 1$, and $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_m]$ is an element of $(m+1)$-dimensional simplex, i.e., $\sum_{i=0}^{m} \beta_i = 1, \beta_i \le 1$. Eq. (7) indicates that a novel non-semantic factor can be generated by taking a convex combination of the amplitudes from $\mathbf{x}_i$ and m given data points $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_m$. By selecting different $\boldsymbol{\beta}$ from the $(m+1)$-dimensional simplex, we can create different non-semantic factors for $\mathbf{x}_i$. Therefore, the non-semantic space $\mathcal{E}_i$ can be presented as follows:

$$\mathcal{E}_i = \{\mathbf{A}(\mathbf{x}_i; \boldsymbol{\beta}, \gamma) : \ \forall\ \boldsymbol{\beta} \text{ from the } (m+1)\text{-dimensional simplex}\}. \tag{8}$$

Lastly, the amplitude $\mathbf{A}(\mathbf{x}_i; \boldsymbol{\beta}, \gamma)$ is combined with the original phase $\mathbf{P}(\mathbf{x}_i)$ to construct new image with same semantic but different style:

$$\mathbf{x}_i^{\beta,\gamma} = \mathbf{M}(\mathbf{P}(\mathbf{x}_i), \mathbf{A}(\mathbf{x}_i;\boldsymbol{\beta},\gamma)) = \mathbf{F}^{-1}\left(\mathbf{A}(\mathbf{x}_i;\boldsymbol{\beta},\gamma) * \exp(-j * \mathbf{P}(\mathbf{x}_i))\right), \tag{9}$$

where $\mathbf{F}^{-1}$ is the inverse Fourier transformation, j is the imaginary unit, and $*$ is the convolution operation. Here $\mathbf{M}(\mathbf{P}(\mathbf{x}_i), \mathbf{A}(\mathbf{x}_i;\boldsymbol{\beta},\gamma))$ is regard as causal mechanism. For more information related to the inverse Fourier transformation, please refer to [12].

5.2. Learning details

This section mainly introduces how to achieve the learning strategy proposed in Section 5.1.

Loss function and algorithm design. Following [16,12], we adopt the cross entropy loss to realize ℓ in Eq. (6). Additionally, according to the discussion in Section 5.1, we mainly focus on designing the optimization strategy to address Eq. (6). Note that the implementation of the causal mechanism $\mathbf{M}$ is crucial in algorithm design. In Section 5.1, we mention utilizing Fourier-based style transfer [12] to construct $\mathbf{M}$ in practice.

Exploring worst-case non-semantic factors. To optimize Eq. (5), the key challenge is how to select the worst-case non-semantic factor from $\mathcal{E}_i$ given by Eq. (8) in each iteration. According to Eq. (8), this is equal to select the optimal $\boldsymbol{\beta}$ from the $(m+1)$-dimensional simplex. Motivated by [31], we use the gradient direction to update $\boldsymbol{\beta}$: given inner learning rate l and inner iterative step L,

$$\boldsymbol{\beta}_k = \frac{\boldsymbol{\beta}_{k-1} + l\,\mathrm{sign}\left[\nabla_{\boldsymbol{\beta}}\ell(\mathbf{g}_{\mathbf{w}}(\mathbf{x}_i^{\boldsymbol{\beta}_{k-1},\gamma}), y_i)\right]}{\|\boldsymbol{\beta}_{k-1} + l\,\mathrm{sign}\left[\nabla_{\boldsymbol{\beta}}\ell(\mathbf{g}_{\mathbf{w}}(\mathbf{x}_i^{\boldsymbol{\beta}_{k-1},\gamma}), y_i)\right]\|_2}, \tag{10}$$

where $\mathbf{x}_i^{\boldsymbol{\beta}_{k-1},\gamma} = \mathbf{M}(\mathbf{P}(\mathbf{x}_i), \mathbf{A}(\mathbf{x}_i;\boldsymbol{\beta}_{k-1},\gamma))$. Then the final augmented data would be presented as $\mathbf{x}_i^{\boldsymbol{\beta}_L,\gamma}$.

Optimization objective. After the worst-case non-semantic factor $\mathbf{e}$ from $\mathcal{E}_i$ for each data is obtained, we generate the worst-case augmented data $\mathbf{x}_i^{\boldsymbol{\beta}_L,\gamma}$ and use gradient descent to minimize the following problem:

$$\widehat{\mathbf{w}} \in \mathop{\arg\min}_{\mathbf{w}\in\mathcal{W}} \mathcal{L}(\mathbf{g}_{\mathbf{w}};\mathcal{D}_{\mathrm{id}},\alpha,\gamma,L), \tag{11}$$

where $\mathcal{L}(\mathbf{g}_{\mathbf{w}};\mathcal{D}_{\mathrm{id}},\alpha,\gamma,L)$ is presented in Eq. (5) and can be rewritten as

$$\frac{1-\alpha}{n}\sum_{i=1}^{n}\ell(\mathbf{g}_{\mathbf{w}}(\mathbf{x}_i^{\boldsymbol{\beta}_L,\gamma}), y_i) + \frac{\alpha}{n}\sum_{i=1}^{n}\ell(\mathbf{g}_{\mathbf{w}}(\mathbf{x}_i), y_i). \tag{12}$$

The pseudo code is summarized in Algorithm 1.

6. Experiments

This section conducts extensive experiments for NsED in standard OOD detection benchmarks [32]. The experimental setup is described in Section 6.1, and Section 6.2 shows the main outcomes of our method, integrated with representative post-hoc OOD detection methods. Additionally, in Section 6.4, we conduct parameter analysis to assess our method thoroughly.

6.1. Experimental setup

Baselines and datasets. In this work, we mainly integrate NsED with 6 representative post-hoc OOD detection methods, namely, MSP [16], ODIN [13], Energy [5], GradNorm [6], React [14], and KLM [15]. The details of these methods have been introduced in Section 2. Following [32], experiments are conducted on standard benchmarks introduced as follows:

- **CIFAR-10** [1] is a dataset that contains 10 distinct classes, consisting of $50,000$ training images and $10,000$ test images for classification tasks. We utilize CIFAR-10 as the ID dataset and investigate two different OOD scenarios: near-OOD and far-OOD. More specifically, we utilize CIFAR-100 [1] and TinyImageNet [33] as the near-OOD datasets, while MNIST [34], SVHN [35], Textures [36], and Places365 [37] as the far-OOD datasets. We exclude any data whose labels match those of the ID cases.
- **CIFAR-100** [1] is a dataset consisting of 50,000 training images and 10,000 test images, divided into 100 distinct classes. Regarding OOD datasets, near-OOD cases comprise CIFAR-10 and TinyImageNet [1]. Far-OOD cases contain MNIST [34], SVHN [35], Textures [36], and Places365 [37]. We eliminate any data whose labels coincide with ID cases.
- **MNIST** [34] comprises 70,000 grayscale images of handwritten digits from 0 to 9, divided into a training set of 60,000 images and a test set of 10,000 images. Regarding OOD datasets, near-OOD cases include NotMNIST [38] and FashionMNIST [39], while far-OOD cases include CIFAR-10 [1], TinyImageNet [1], Textures [36], and Places365 [37]. We exclude any data with labels that match those in the ID cases.

Implementation details. We adopt the unified settings with common hyperparameters and architectural selections for each implemented method, as suggested by [32], to ensure fair comparisons among different methods.

Table 1
Comparison between our method NsED and baseline methods on the benchmark with CIFAR-10 as ID dataset. ↑ indicates larger AUROC values are preferred. ↓ indicates smaller FPR95 values are preferred. Δ estimates the improvement of NsED integrated with other baselines.

Methods	Near-OOD Dataset				Far-OOD Dataset								Average	
	CIFAR-100		TinyImageNet		MNIST		SVHN		Textures		Places365			
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
MSP	85.74	67.00	87.33	64.16	87.34	63.55	87.37	69.94	88.34	62.31	87.99	62.98	87.35	64.99
NsED+MSP	**87.91**	**64.16**	**89.73**	**59.74**	**90.27**	**60.34**	**93.32**	**44.49**	**92.19**	**48.92**	**89.63**	**59.67**	**90.51**	**56.22**
Δ	+2.17	-2.84	+2.40	-4.42	+2.93	-3.21	+5.95	-25.45	+3.85	-13.39	+1.64	-3.31	+3.16	-8.77
ODIN	83.38	60.06	85.77	55.41	87.09	50.69	81.94	70.88	86.66	53.81	87.54	51.30	85.40	57.03
NsED+ODIN	**87.01**	**57.99**	**89.84**	**50.00**	**93.04**	**39.38**	**94.26**	**29.43**	**92.65**	**36.05**	**90.60**	**47.39**	**91.23**	**43.37**
Δ	+3.63	-2.07	+4.07	-5.41	+5.95	-11.31	+12.32	-41.45	+5.99	-17.76	+3.06	-3.91	+5.83	-13.66
Energy	84.95	57.83	87.36	52.17	87.69	50.38	83.63	71.09	87.58	53.08	88.77	48.35	86.66	55.48
NsED+Energy	**87.90**	**56.13**	**90.72**	**47.16**	**92.70**	**41.74**	**95.84**	**21.68**	**93.40**	**32.93**	**91.37**	**44.37**	**91.99**	**40.67**
Δ	+2.95	-1.70	+3.36	-5.01	+5.01	-8.64	+12.21	-49.41	+5.82	-20.15	+2.60	-3.98	+5.33	-14.81
GradNorm	63.13	81.65	63.46	80.01	61.36	82.78	53.30	87.03	63.70	79.69	68.09	77.23	62.17	81.40
NsED+GradNorm	**67.83**	**74.93**	**71.57**	**69.38**	**73.18**	**70.74**	**92.67**	**24.62**	**81.56**	**43.83**	**74.39**	**67.32**	**76.87**	**58.47**
Δ	+4.70	-6.72	+8.11	-10.63	+11.82	-12.04	+39.37	-62.41	+17.86	-35.86	+6.30	-9.91	+14.70	-22.93
React	84.90	58.01	87.28	52.35	87.41	50.75	83.66	70.66	88.08	52.43	88.57	48.86	86.65	55.51
NsED+React	**87.98**	**56.06**	**90.77**	**47.01**	**92.58**	**42.34**	**95.86**	**22.13**	**93.79**	**32.18**	**91.35**	**44.50**	**92.06**	**40.70**
Δ	+3.08	-1.95	+3.49	-5.34	+5.17	-8.41	+12.20	-48.53	+5.71	-20.25	+2.78	-4.36	+5.41	-14.81
KLM	76.99	66.97	78.26	63.43	82.63	**62.91**	83.89	72.40	81.16	62.32	75.95	60.94	79.81	64.83
NsED+KLM	**80.40**	**64.14**	**82.94**	**59.43**	**82.91**	63.46	**87.93**	**54.28**	**87.60**	**49.85**	**81.02**	**57.46**	**83.80**	**58.10**
Δ	+3.41	-2.83	+4.68	-4.00	+0.28	+0.55	+4.04	-18.12	+6.44	-12.47	+5.07	-3.48	+3.99	-6.73

Table 2
Comparison between our method NsED and baseline methods on the benchmark with CIFAR-100 as ID dataset. ↑ indicates larger AUROC values are preferred. ↓ indicates smaller FPR95 values are preferred. Δ estimates the improvement of NsED integrated with other baselines.

Methods	Near-OOD Dataset				Far-OOD Dataset								Average	
	CIFAR-10		TinyImageNet		MNIST		SVHN		Textures		Places365			
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
MSP	75.26	83.11	78.80	78.87	**70.94**	**89.21**	**80.06**	75.76	72.79	85.35	76.56	81.89	75.73	82.36
NsED+MSP	**75.87**	**82.72**	**79.59**	**78.35**	66.93	91.66	79.98	**75.03**	**77.78**	**76.91**	**77.35**	**81.76**	**76.25**	**81.07**
Δ	+0.61	-0.39	+0.79	-0.52	-4.01	+2.45	-0.08	-0.73	+4.99	-8.44	+0.79	-0.13	+0.52	-1.29
ODIN	75.50	83.50	79.74	78.63	**77.55**	**86.81**	83.29	**74.83**	74.65	84.35	**77.95**	**81.81**	78.11	81.66
NsED+ODIN	**76.03**	**83.18**	**80.45**	**77.88**	73.51	87.96	**83.56**	75.19	**80.32**	**73.22**	77.94	82.14	**78.64**	**79.93**
Δ	+0.53	-0.32	+0.71	-0.75	-4.04	+1.15	+0.27	+0.36	+5.67	-11.13	-0.01	+0.33	+0.53	-1.73
Energy	75.48	**83.94**	79.77	79.21	**75.76**	**90.59**	84.48	72.40	74.60	84.61	**77.91**	**82.39**	78.00	82.19
NsED+Energy	**75.66**	84.00	**80.27**	**78.43**	70.27	91.71	**86.82**	**64.47**	**81.32**	**69.85**	77.68	82.67	**78.67**	**78.52**
Δ	+0.18	+0.06	+0.50	-0.78	-5.49	+1.12	+2.34	-7.93	+6.72	-14.76	-0.23	+0.28	+0.67	-3.67
GradNorm	62.14	89.63	65.05	88.01	**65.68**	**92.10**	77.37	71.63	65.61	84.93	**65.97**	**87.12**	66.97	85.57
NsED+GradNorm	**67.27**	**85.32**	**66.37**	**83.83**	52.07	94.11	**85.26**	**55.41**	**73.71**	**64.90**	63.26	87.15	**67.99**	**78.45**
Δ	+5.13	-4.31	+1.32	-4.18	-13.61	+2.01	+7.89	-16.22	+8.10	-20.03	-2.71	+0.03	+1.02	-7.12
React	74.92	84.21	79.89	79.23	**75.25**	**90.94**	86.70	69.74	78.28	83.16	**78.41**	**82.36**	78.91	81.61
NsED+React	**75.67**	**83.99**	**80.57**	**78.24**	70.36	91.64	**87.30**	**63.75**	**82.78**	**68.84**	78.11	82.42	**79.13**	**78.15**
Δ	+0.75	-0.22	+0.68	-0.99	-4.89	+0.70	+0.60	-5.99	+4.50	-14.32	-0.30	+0.06	+0.22	-3.46
KLM	72.49	**76.99**	77.55	**70.76**	**67.89**	**79.22**	79.72	60.48	71.82	70.21	73.52	72.02	73.83	71.61
NsED+KLM	**73.40**	78.54	**78.73**	71.10	65.70	82.71	**80.97**	**58.65**	**78.54**	**63.57**	**76.24**	**72.00**	**75.60**	**71.09**
Δ	+0.91	+1.55	+1.18	+0.34	-2.19	+3.49	+1.25	-1.83	+6.72	-6.64	+2.72	-0.02	+1.77	-0.52

- **Pre-training setup.** Following [32], we utilize ResNet-18 [40] for scenarios involving CIFAR datasets as the ID datasets, and LeNet [41] for scenarios involving MNIST as the ID dataset. To train the deep models, we utilize the Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.1, a momentum of 0.9, and a weight decay rate of 0.0005 for 100 epochs to avoid overfitting.
- **Hyperparameter setup.** For our method NsED, we utilize the SGD optimizer with a learning rate of 0.1, a momentum of 0.9, and a weight decay rate of 0.0005 for 100 epochs with a batch size of 128. We follow the standard augmentation protocol suggested in [12]. The learning rate is decayed by 0.5 every 10 epochs. In our experiments, we set $\alpha = 0.5$, $\gamma = 1$, $m = 9$, the inner step size as $l = 0.05$, and the number of inner step iterations as $L = 10$. Parameter analysis about important parameters α, γ, m and L are conducted in Section 6.4.

Evaluations. Following the standard evaluation strategy [16], the detection performance is assessed using representative metrics: the area under the receiver operating characteristic curve (AUROC), FPR95, and the area under the precision recall curve (AUPR). These metrics are threshold-independent. FPR95 is the false positive rate of OOD data when the true positive rate of ID data is at 95%. AUROC represents the probability of the ID case having a higher score than that of the OOD case, as determined by the area under the receiver operating characteristic curve. The AUPR, on the other hand, quantifies the overall performance of a binary classifier. It is calculated as the area under the precision-recall curve, where precision is plotted against recall. In cases where datasets exhibit a significant imbalance between positive and negative cases, the AUPR provides more informative insights. Similar to [16], we report the AUPR-In and AUPR-Out values.

Table 3
Comparison between our method NsED and baseline methods on the benchmark with MNIST as ID dataset. ↑ indicates larger AUROC values are preferred. ↓ indicates smaller FPR95 values are preferred. Δ estimates the improvement of NsED integrated with other baselines.

Methods	Near-OOD Dataset				Far-OOD Dataset								Average	
	NotMNIST		FashionMNIST		CIFAR-10		TinyImageNet		Textures		Places365			
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
MSP	93.64	32.00	95.61	20.94	99.38	1.99	98.87	5.07	99.01	4.57	98.71	**6.23**	97.54	11.80
NsED+MSP	**94.03**	**29.69**	**97.58**	**14.39**	**99.65**	**0.98**	**99.19**	**3.66**	**99.02**	**4.51**	**98.73**	6.26	**98.03**	**9.91**
Δ	+0.39	-2.31	+1.97	-6.55	+0.27	-1.01	+0.32	-1.41	+0.01	-0.06	+0.02	+0.03	+0.49	-1.89
ODIN	**94.08**	29.67	97.16	12.11	99.74	0.80	99.33	2.87	99.45	2.40	99.28	3.32	98.17	8.53
NsED+ODIN	93.56	**27.05**	**98.83**	**5.80**	**99.95**	**0.17**	**99.80**	**0.89**	**99.72**	**1.33**	**99.56**	**2.05**	**98.57**	**6.21**
Δ	-0.52	-2.62	+1.67	-6.31	+0.21	-0.63	+0.47	-1.98	+0.27	-1.07	+0.28	-1.27	+0.40	-2.32
Energy	**94.00**	29.90	97.21	11.66	99.76	0.80	99.34	2.91	99.46	2.35	99.29	3.30	98.18	8.49
NsED+Energy	93.01	**29.39**	**98.85**	**5.76**	**99.96**	**0.18**	**99.80**	**0.89**	**99.72**	**1.33**	**99.54**	**2.15**	**98.48**	**6.62**
Δ	-0.99	-0.51	+1.64	-5.90	+0.20	-0.62	+0.46	-2.02	+0.26	-1.02	+0.25	-1.15	+0.30	-1.87
GradNorm	84.76	53.29	95.48	19.83	99.65	0.90	99.07	3.90	99.21	3.32	99.10	3.93	96.21	14.20
NsED+GradNorm	**87.65**	**37.58**	**96.43**	**14.88**	**99.97**	**0.14**	**99.90**	**0.49**	**99.78**	**0.96**	**99.62**	**1.79**	**97.23**	**9.31**
Δ	+2.89	-15.71	+0.95	-4.95	+0.32	-0.76	+0.83	-3.41	+0.57	-2.36	+0.52	-2.14	+1.02	-4.89
React	**95.02**	**27.93**	97.76	11.62	99.76	0.78	99.38	2.82	99.48	2.28	99.30	3.21	98.45	8.11
NsED+React	93.48	29.19	**98.84**	**5.97**	**99.96**	**0.18**	**99.79**	**0.95**	**99.71**	**1.35**	**99.54**	**2.22**	**98.55**	**6.64**
Δ	-1.54	+1.26	+1.08	-5.65	+0.20	-0.60	+0.41	-1.87	+0.23	-0.93	+0.24	-0.99	+0.10	-1.47
KLM	85.20	28.44	88.90	18.54	98.79	1.88	97.46	5.08	97.81	**4.41**	96.82	**6.75**	94.16	10.85
NsED+KLM	**89.20**	**28.01**	**95.59**	**14.16**	**99.39**	**1.30**	**98.40**	**4.29**	**98.00**	5.10	**97.25**	7.22	**96.31**	**10.01**
Δ	+4.00	-0.43	+6.69	-4.38	+0.60	-0.58	+0.94	-0.79	+0.19	+0.69	+0.43	+0.47	+2.15	-0.84

6.2. Main results

The detection performance on different datasets (CIFAR-10, CIFAR-100, MNIST as ID datasets) is shown in Tables 1, 2, 3, 4, and 5. ↓ (or ↑) indicates that smaller (or larger) values are preferred. Δ estimates the improvement of NsED when integrated with other baselines. From our analysis of Tables 1-5, we can summarize the following results.

- **CIFAR-10.** The OOD detection performance of methods NsED+MSP, NsED+ODIN, NsED+Energy, NsED+GradNorm, NsED+React and NsED+KLM is significantly better compared to those of MSP, ODIN, Energy, GradNorm, React, and KLM in all OOD cases. Specifically, the average AUROC improvements are 3.16, 5.83, 5.33, 6.70, 5.38, and 3.99. Meanwhile, the average FPR95 improvements are 8.76, 13.64, 14.81, 22.93, 10.87, 14.71, and 6.73. The average AUPR-In improvements are 7.92, 13.43, 12.08, 15.00, 12.97, and 3.16, and the average AUPR-Out improvements are 2.70, 3.86, 3.87, 11.50, 3.86, and 3.18. Therefore, using NsED to train deep models can significantly and consistently enhance the detection performance of MSP, ODIN, Energy, GradNorm, React, and KLM on the CIFAR-10 benchmark when used as the ID dataset.
- **CIFAR-100.** NsED+MSP, NsED+ODIN, NsED+Energy, NsED+KLM, NsED+React, and NsED+GradNorm exhibit better OOD detection performance than MSP, ODIN, Energy, GradNorm, React, and KLM in almost all OOD cases (6/6, 3/6, 4/6, 6/6, 5/6, 3/6). The average AUROC improvements are 0.52, 0.53, 0.67, 1.02, 0.35, and 1.77, and the average FPR95 improvements are 1.31, 1.72, 4.26, 7.12, 1.12, and 0.52, respectively. Although the average AUPR-In values are slightly lower than those of the baselines (the overall average AUPR-In improvement is 0.26), the average AUPR-Out values show significant improvement, with gains of 1.57, 1.83, 2.37, 7.88, 2.00, and 1.11. Therefore, NsED consistently improves the detection performance of MSP, ODIN, Energy, GradNorm, React, and KLM on the CIFAR-100 benchmark. It's worth noting that the improvement in detection performance on the CIFAR-100 benchmark is less than that on the CIFAR-10 benchmark. One possible reason for this difference is the increased difficulty in extracting semantic features due to the larger number of classes.
- **MNIST.** NsED, when combined with MSP, ODIN, Energy, GradNorm, React, and KLM, exhibits better performance in detecting OOD cases compared to the individual methods MSP, ODIN, Energy, GradNorm, React, and KLM. The improvement is noticeable in almost all OOD cases (5/6, 5/6, 5/6, 6/6, 5/6, 4/6), with an average AUROC improvement of 0.49, 0.40, 0.30, 1.02, 0.10, and 2.15, and an average FPR95 improvement of 1.89, 2.32, 1.87, 4.89, 1.47, and 0.84, respectively. The average AUPR-In improvements are 0.49, 0.24, 0.10, 0.27, −0.32, and 5.46, while the average AUPR-Out improvements are 0.41, 0.43, 0.37, 0.85, 0.26, and 0.97. Hence, NsED enhances the detection performance of MSP, ODIN, Energy, GradNorm, React, and KLM on the MNIST benchmark. It's worth noting that the improvement in detection performance on the MNIST benchmark is not as significant as on the CIFAR-10 benchmark. One possible explanation for this difference is that the original methods (without NsED) have already achieved a high performance level, making it challenging to significantly improve. For example, React exhibits an AUROC and FPR95 of 99.76 and 0.78, respectively.
- **Near-OOD tasks.** The similarity between the non-semantic factors in ID and OOD data [4] makes near-OOD tasks more challenging than far-OOD tasks. However, our experiments show that the enhancement in performance, denoted as Δ, is comparable to what is achieved in most far-OOD tasks. This suggests that NsED is effective both in extracting semantic features and in mitigating issues related to spurious correlations.
- **ID accuracy.** We conducted experiments to assess the accuracy of the ID in Table 5. The results, presented in Table 5, demonstrate that overall, NsED achieved slightly better performance in ID classification. This finding suggests that NsED can ensure comparable performance in ID classifi

Table 4
Comparison between our method NsED and baseline methods on the benchmark with CIFAR-10, CIFAR-100 and MNIST as ID datasets. ↑ indicates larger AUPR values are preferred. Δ estimates the improvement of NsED integrated with other baselines.

ID Dataset: CIFAR-10

Methods	Near-OOD Dataset				Far-OOD Dataset								Average	
	CIFAR-100		TinyImageNet		MNIST		SVHN		Textures		Places365			
	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑
MSP	84.93	83.27	88.29	83.24	55.14	97.39	78.87	92.64	93.01	77.70	71.92	95.21	78.69	88.24
NsED+MSP	**88.88**	**84.93**	**91.99**	**85.56**	**74.51**	**97.86**	**89.19**	**96.56**	**95.56**	**84.92**	**79.51**	**95.78**	**86.61**	**90.94**
Δ	+3.95	+1.66	+3.70	+2.32	+19.37	+0.47	+10.32	+3.92	+2.55	+7.22	+7.59	+0.57	+7.92	+2.70
ODIN	80.23	83.15	84.97	83.77	48.08	97.67	64.57	90.69	90.37	78.95	65.65	95.61	72.31	88.31
NsED+ODIN	**86.47**	**85.40**	**90.93**	**87.07**	**76.17**	**98.64**	**88.30**	**97.42**	**94.98**	**87.97**	**77.59**	**96.49**	**85.74**	**92.17**
Δ	+6.24	+2.25	+5.96	+3.30	+28.09	+0.97	+23.73	+6.73	+4.61	+9.02	+11.94	+0.88	+13.43	+3.86
Energy	82.34	84.43	86.89	85.31	50.26	97.74	69.04	91.16	91.28	79.69	69.02	96.05	74.81	89.06
NsED+Energy	**87.60**	**86.35**	**91.88**	**88.21**	**75.77**	**98.54**	**91.21**	**98.23**	**95.48**	**89.41**	**79.40**	**96.83**	**86.89**	**92.93**
Δ	+5.26	+1.92	+4.99	+2.90	+25.51	+0.80	+22.17	+7.07	+4.20	+9.72	+10.38	+0.78	+12.08	+3.87
GradNorm	60.43	64.88	63.47	62.89	20.43	91.78	33.63	76.27	72.86	53.17	36.58	88.08	47.90	72.84
NsED+GradNorm	**64.09**	**70.61**	**70.53**	**72.24**	**33.68**	**94.71**	**81.37**	**97.26**	**83.51**	**80.03**	**44.21**	**91.19**	**62.90**	**84.34**
Δ	+3.66	+5.73	+7.06	+9.35	+13.25	+2.93	+47.74	+20.99	+10.65	+26.86	+7.63	+3.11	+15.00	+11.50
ReAct	82.03	84.42	86.55	85.28	47.90	97.70	68.26	91.22	91.66	80.23	68.07	95.99	74.08	89.14
NsED+ReAct	**87.73**	**86.40**	**91.93**	**88.25**	**75.70**	**98.51**	**91.52**	**98.23**	**95.93**	**89.80**	**79.49**	**96.82**	**87.05**	**93.00**
Δ	+5.70	+1.98	+5.38	+2.97	+27.80	+0.81	+23.26	+7.01	+4.27	+9.57	+11.42	+0.83	+12.97	+3.86
KLM	68.82	78.62	72.49	78.38	**40.83**	96.70	71.40	91.17	83.40	73.20	35.62	92.42	62.09	85.08
NsED+KLM	**73.05**	**81.17**	**78.39**	**82.12**	36.94	**96.73**	**71.70**	**94.11**	**89.13**	**81.48**	**42.29**	**93.94**	**65.25**	**88.26**
Δ	+4.23	+2.55	+5.90	+3.74	-3.89	+0.03	+0.30	+2.94	+5.73	+8.28	+6.67	+1.52	+3.16	+3.18

ID Dataset: CIFAR-100

Methods	Near-OOD Dataset				Far-OOD Dataset								Average	
	CIFAR-10		TinyImageNet		MNIST		SVHN		Textures		Places365			
	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑
MSP	77.93	70.78	85.02	68.80	**43.75**	**92.71**	**69.65**	89.11	83.21	55.11	59.13	89.51	**69.78**	77.67
NsED+MSP	**78.62**	**71.40**	**85.65**	**69.78**	37.80	91.44	68.85	**89.14**	**86.07**	**63.84**	**60.86**	**89.84**	69.64	**79.24**
Δ	+0.69	+0.62	+0.63	+0.98	-5.95	-1.27	-0.80	+0.03	+2.86	+8.73	+1.73	+0.33	-0.14	+1.57
ODIN	77.20	70.92	85.12	69.80	**53.17**	**94.30**	74.24	**90.28**	84.29	57.40	59.65	90.03	**72.28**	78.79
NsED+ODIN	**78.19**	**71.46**	**85.89**	**70.76**	44.95	93.39	**75.34**	90.20	**87.66**	**67.89**	**59.98**	**90.03**	72.00	**80.62**
Δ	+0.99	+0.54	+0.77	+0.96	-8.22	-0.91	+1.10	-0.08	+3.37	+10.49	+0.33	+0.00	-0.28	+1.83
Energy	77.21	70.75	85.23	69.68	**50.77**	**93.46**	76.12	90.83	84.24	57.24	60.00	**89.92**	**72.26**	78.65
NsED+Energy	**77.93**	**70.83**	**85.82**	**70.36**	41.02	92.22	**79.20**	**92.66**	**88.15**	**70.24**	**60.15**	89.81	72.05	**81.02**
Δ	+0.72	+0.08	+0.59	+0.68	-9.75	-1.24	+3.08	+1.83	+3.91	+13.00	+0.15	-0.11	-0.21	+2.37
GradNorm	62.34	60.15	**71.25**	53.48	**34.77**	**90.92**	51.60	88.45	76.03	49.63	**38.48**	68.53	**55.75**	68.53
NsED+GradNorm	**65.91**	**65.45**	70.07	**59.37**	18.18	87.72	**70.07**	**93.10**	**78.48**	**68.23**	31.69	**84.57**	55.73	**76.41**
Δ	+3.57	+5.30	-1.18	+5.89	-16.59	-3.20	+18.47	+4.65	+2.45	+18.60	-6.79	+16.04	-0.02	+7.88
ReAct	76.29	70.40	85.32	69.82	**49.94**	**93.33**	**80.72**	91.84	87.59	60.67	60.94	**90.10**	**73.47**	79.36
NsED+ReAct	**77.94**	**70.82**	**86.16**	**70.60**	42.30	92.23	80.22	**92.90**	**89.49**	**71.67**	**61.25**	89.96	72.89	**81.36**
Δ	+1.65	+0.42	+0.84	+0.78	-7.64	-1.10	-0.50	+1.06	+1.90	+11.00	+0.31	-0.14	-0.58	+2.00
KLM	67.88	**72.90**	79.57	73.01	**29.92**	**93.47**	56.16	91.38	76.91	65.19	38.09	90.46	58.09	81.07
NsED+KLM	**68.08**	72.87	**80.98**	**73.34**	29.32	92.54	**59.37**	**91.79**	**83.40**	**71.38**	**44.28**	**91.14**	**60.90**	**82.18**
Δ	+0.20	-0.03	+1.41	+0.33	-0.60	-0.93	+3.21	+0.41	+6.49	+6.19	+6.19	+0.68	+2.81	+1.11

ID Dataset: MNIST

Methods	Near-OOD Dataset				Far-OOD Dataset								Average	
	NotMNIST		FashionMNIST		CIFAR-10		TinyImageNet		Textures		Places365			
	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑
MSP	**89.77**	96.10	95.21	95.93	99.62	99.02	98.90	98.90	**99.03**	99.05	**96.59**	99.64	96.52	98.11
NsED+MSP	89.67	**96.42**	**97.89**	**97.29**	**99.79**	**99.43**	**99.19**	**99.24**	99.02	**99.08**	96.50	**99.65**	**97.01**	**98.52**
Δ	-0.10	+0.32	+2.68	+1.36	+0.17	+0.41	+0.29	+0.34	-0.01	+0.03	-0.09	+0.01	+0.49	+0.41
ODIN	**90.20**	96.43	96.72	97.55	99.83	99.63	99.29	99.38	99.42	99.49	97.80	99.81	97.21	98.71
NsED+ODIN	88.07	**96.58**	**98.85**	**98.84**	**99.97**	**99.93**	**99.77**	**99.83**	**99.67**	**99.76**	**98.36**	**99.89**	**97.45**	**99.14**
Δ	-2.13	+0.15	+2.13	+1.29	+0.14	+0.30	+0.48	+0.45	+0.25	+0.27	+0.56	+0.08	+0.24	+0.43
Energy	**90.16**	**96.35**	96.76	97.61	99.84	99.67	99.29	99.40	99.42	99.51	97.79	99.81	97.21	98.72
NsED+Energy	87.29	96.26	**98.88**	**98.86**	**99.97**	**99.94**	**99.77**	**99.83**	**99.67**	**99.77**	**98.28**	**99.89**	**97.31**	**99.09**
Δ	-2.87	-0.09	+2.12	+1.25	+0.13	+0.27	+0.48	+0.43	+0.25	+0.26	+0.49	+0.08	+0.10	+0.37
GradNorm	71.14	91.39	94.83	96.16	99.68	99.59	98.84	99.24	99.02	99.35	96.59	99.77	93.35	97.58
NsED+GradNorm	**73.79**	**93.85**	**95.83**	**97.11**	**99.98**	**99.96**	**99.88**	**99.91**	**99.73**	**99.83**	**98.48**	**99.91**	**94.62**	**98.43**
Δ	+2.65	+2.46	+1.00	+0.95	+0.30	+0.37	+1.04	+0.67	+0.71	+0.48	+1.89	+0.14	+1.27	+0.85
ReAct	**92.37**	**96.82**	97.69	97.92	99.84	99.67	99.36	99.43	99.45	99.52	97.82	99.81	**97.75**	98.86
NsED+ReAct	88.03	96.47	**98.88**	**98.85**	**99.97**	**99.94**	**99.76**	**99.83**	**99.66**	**99.76**	**98.29**	**99.89**	97.43	**99.12**
Δ	-4.34	-0.35	+1.19	+0.93	+0.13	+0.27	+0.40	+0.40	+0.21	+0.24	+0.47	+0.08	-0.32	+0.26
KLM	60.22	93.44	76.08	93.17	98.38	98.72	94.52	98.18	95.21	98.43	81.42	99.24	84.30	96.86
NsED+KLM	**69.77**	**94.90**	**92.20**	**96.36**	**99.46**	**99.12**	**96.99**	**98.77**	**95.96**	**98.51**	**84.19**	**99.33**	**89.76**	**97.83**
Δ	+9.55	+1.46	+16.12	+3.19	+1.08	+0.40	+2.47	+0.59	+0.75	+0.08	+2.77	+0.09	+5.46	+0.97

Table 5
Comparison between our method NsED and baseline methods on the benchmark with CIFAR-10, CIFAR-100 and MNIST as ID datasets. We report the average ID accuracy (%) of our method and baseline methods.

Dataset	CIFAR-10		CIFAR-100		MNIST	
	Baselines	NsED+Baselines	Baselines	NsED+Baselines	Baselines	NsED+Baselines
Average Acc.	93.29	**93.39**	71.77	**73.06**	**98.88**	98.79

Table 6

Comparison between our method and several GAN-based OOD detection methods on the benchmark with CIFAR-10, CIFAR-100 and MNIST as ID datasets. ↑ indicates larger AUROC values are preferred. ↓ indicates smaller FPR95 values are preferred.

ID Dataset: CIFAR-10

Methods	Near-OOD Dataset				Far-OOD Dataset								Average	
	CIFAR-100		TinyImageNet		MNIST		SVHN		Textures		Places365			
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
ConfGAN+MSP	71.31	70.46	71.27	67.93	62.47	76.44	71.73	62.83	67.61	72.9	75.06	65.42	69.91	69.33
BoundaryGAN+MSP	72.1	70.46	74.16	64.92	68.97	80.12	70.98	67.11	72.07	68.24	72.85	67.31	71.86	69.69
CMG+MSP	75.08	100.0	76.81	100.0	67.9	100.0	72.76	100.0	73.18	100.0	73.04	100.0	73.13	100.00
NsED+MSP	**87.91**	**64.16**	**89.73**	**59.74**	**90.27**	**60.34**	**93.32**	**44.49**	**92.19**	**48.92**	**89.63**	**59.67**	**90.51**	**56.22**
ConfGAN+Energy	69.02	71.47	70.75	69.67	51.83	79.45	57.96	67.71	64.01	71.3	71.47	66.39	64.17	71.00
BoundaryGAN+Energy	69.54	72.24	70.69	68.07	61.7	81.4	65.58	67.79	71.8	65.98	68.38	71.53	67.95	71.17
CMG+Energy	53.75	85.31	52.39	84.93	54.02	90.63	74.96	59.56	62.64	80.73	56.83	85.42	59.10	81.10
NsED+Energy	**87.90**	**56.13**	**90.72**	**47.16**	**92.70**	**41.74**	**95.84**	**21.68**	**93.40**	**32.93**	**91.37**	**44.37**	**91.99**	**40.67**

ID Dataset: CIFAR-100

Methods	Near-OOD Dataset				Far-OOD Dataset								Average	
	CIFAR-10		TinyImageNet		MNIST		SVHN		Textures		Places365			
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
ConfGAN+MSP	61.57	83.44	65.8	78.46	**68.88**	**75.33**	59.69	82.05	58.35	84.56	54.23	90.3	61.42	82.36
BoundaryGAN+MSP	61.7	82.84	66.3	**76.81**	55.38	80.54	57.34	86.78	57.33	83.43	55.63	90.3	58.95	83.45
CMG+MSP	60.13	**81.64**	62.01	81.39	56.13	85.58	61.39	78.32	53.09	100.0	52.99	100.0	57.62	87.82
NsED+MSP	**75.87**	82.72	**79.59**	78.35	66.93	91.66	**79.98**	**75.03**	**77.78**	**76.91**	**77.35**	**81.76**	**76.25**	**81.07**
ConfGAN+Energy	64.61	**81.24**	63.57	75.54	44.39	**78.87**	57.42	80.92	53.78	84.41	58.7	89.21	57.08	81.70
BoundaryGAN+Energy	62.83	81.36	66.85	**74.06**	43.17	80.94	56.99	83.93	54.26	85.12	56.19	91.54	56.72	82.83
CMG+Energy	55.76	89.52	52.22	90.91	43.43	88.15	72.29	76.35	76.07	78.06	52.86	93.08	58.77	86.01
NsED+Energy	**75.66**	84.00	**80.27**	78.43	**70.27**	91.71	**86.82**	**64.47**	**81.32**	**69.85**	**77.68**	**82.67**	**78.67**	**78.52**

ID Dataset: MNIST

Methods	Near-OOD Dataset				Far-OOD Dataset								Average	
	NotMNIST		FashionMNIST		CIFAR-10		TinyImageNet		Textures		Places365			
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
ConfGAN+MSP	91.15	32.81	86.03	45.1	83.08	53.4	79.25	59.65	78.79	59.97	80.9	55.38	83.20	51.05
BoundaryGAN+MSP	89.47	38.91	82.51	50.86	87.58	41.13	84.49	44.66	87.14	48.64	86.59	41.05	86.30	44.21
CMG+MSP	69.31	100.0	92.16	100.0	94.33	100.0	95.72	100.0	98.43	**1.21**	95.49	100.0	90.91	83.53
NsED+MSP	**94.03**	**29.69**	**97.58**	**14.39**	**99.65**	**0.98**	**99.19**	**3.66**	**99.02**	4.51	**98.73**	**6.26**	**98.03**	**9.91**
ConfGAN+Energy	73.96	48.61	85.48	42.73	83.21	52.82	81.18	59.02	78.04	60.99	82.22	54.57	80.68	53.12
BoundaryGAN+Energy	67.81	50.56	80.14	49.09	79.21	43.31	80.21	47.21	76.14	64.62	79.03	44.84	77.09	49.94
CMG+Energy	**97.11**	**12.33**	98.78	**4.28**	99.44	2.3	99.47	2.21	99.15	5.18	99.5	2.22	**98.91**	**4.75**
NsED+Energy	93.01	29.39	**98.85**	5.76	**99.96**	**0.18**	**99.80**	**0.89**	**99.72**	**1.33**	**99.54**	**2.15**	98.48	6.62

Table 7
Comparison between our method and several GAN-based OOD detection methods on the benchmark with CIFAR-10, CIFAR-100 and MNIST as ID datasets. ↑ indicates larger AUPR values are preferred.

ID Dataset: CIFAR-10

Methods	Near-OOD Dataset				Far-OOD Dataset								Average	
	CIFAR-100		TinyImageNet		MNIST		SVHN		Textures		Places365			
	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑
ConfGAN+MSP	66.36	74.24	65.79	75.29	55.45	68.02	81.95	61.89	47.43	80.96	70.74	77.92	64.62	73.05
BoundaryGAN+MSP	67.44	74.59	68.95	77.32	64.31	69.74	82.5	58.43	53.21	83.75	68.27	75.93	67.45	73.29
CMG+MSP	70.18	79.09	71.59	80.62	63.73	73.12	84.26	61.39	54.61	85.05	68.07	77.39	68.74	76.11
NsED+MSP	**88.88**	**84.93**	**91.99**	**85.56**	**74.51**	**97.86**	**89.19**	**96.56**	**95.56**	**84.92**	**79.51**	**95.78**	**86.61**	**90.94**
ConfGAN+Energy	64.11	72.6	66.54	74.28	47.0	61.49	70.94	52.99	43.28	79.59	64.53	75.92	59.40	69.48
BoundaryGAN+Energy	62.47	72.78	62.63	74.94	53.8	65.89	76.15	56.24	50.19	84.05	60.99	72.68	61.04	71.10
CMG+Energy	48.55	59.19	47.38	59.03	49.61	56.54	80.75	64.74	41.4	76.95	51.05	60.68	53.12	62.85
NsED+Energy	**87.60**	**86.35**	**91.88**	**88.21**	**75.77**	**98.54**	**91.21**	**98.23**	**95.48**	**89.41**	**79.40**	**96.83**	**86.89**	**92.93**

ID Dataset: CIFAR-100

Methods	Near-OOD Dataset				Far-OOD Dataset								Average	
	CIFAR-10		TinyImageNet		MNIST		SVHN		Textures		Places365			
	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑
ConfGAN+MSP	58.2	63.9	62.17	68.39	**66.7**	71.4	76.37	43.02	40.39	**73.29**	52.97	56.02	59.47	62.67
BoundaryGAN+MSP	58.14	64.01	61.8	69.65	49.51	62.76	75.56	37.51	39.84	73.25	54.51	56.57	56.56	60.62
CMG+MSP	56.86	63.87	58.87	64.93	52.86	60.27	**77.57**	47.03	37.37	68.41	51.62	55.3	55.86	59.97
NsED+MSP	**78.62**	**71.40**	**85.65**	**69.78**	37.80	**91.44**	68.85	**89.14**	**86.07**	63.84	**60.86**	**89.84**	**69.64**	**79.24**
ConfGAN+Energy	61.09	66.57	55.5	68.67	42.4	58.98	73.18	42.72	35.25	71.53	57.88	58.92	54.22	61.23
BoundaryGAN+Energy	58.78	65.36	61.38	**70.78**	42.03	56.41	74.6	39.31	38.38	70.93	55.84	56.16	55.17	59.82
CMG+Energy	53.78	57.51	50.89	54.27	**42.55**	52.53	**85.01**	53.31	65.74	**82.85**	52.24	53.4	58.37	58.98
NsED+Energy	**77.93**	**70.83**	**85.82**	70.36	41.02	**92.22**	79.20	**92.66**	**88.15**	70.24	**60.15**	**89.81**	**72.05**	**81.02**

ID Dataset: MNIST

Methods	Near-OOD Dataset				Far-OOD Dataset								Average	
	CIFAR-10		TinyImageNet		MNIST		SVHN		Textures		Places365			
	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑	AUPR-In↑	AUPR-Out↑
ConfGAN+MSP	79.48	94.15	79.2	87.9	77.16	84.74	72.71	81.24	59.96	85.29	74.36	82.87	73.81	86.03
BoundaryGAN+MSP	76.88	93.4	76.03	85.14	83.61	89.3	78.61	87.01	74.78	91.97	81.86	88.68	78.63	89.25
CMG+MSP	77.52	85.76	95.31	93.19	96.63	94.9	97.38	96.07	98.43	**99.14**	97.31	95.87	93.76	94.16
NsED+MSP	**89.67**	**96.42**	**97.89**	**97.29**	**99.79**	**99.43**	**99.19**	**99.24**	**99.02**	99.08	**96.50**	**99.65**	**97.01**	**98.52**
ConfGAN+Energy	55.08	85.18	82.34	87.58	79.16	84.85	78.38	82.14	59.64	84.62	78.29	83.46	72.15	84.64
BoundaryGAN+Energy	46.25	82.73	74.57	84.03	68.9	84.44	70.88	84.26	52.46	85.38	67.68	84.11	63.46	84.16
CMG+Energy	**95.71**	**98.06**	98.29	**99.07**	99.25	99.55	99.27	99.57	98.21	99.57	**99.33**	99.59	**98.34**	**99.23**
NsED+Energy	87.29	96.26	**98.88**	98.86	**99.97**	**99.94**	**99.77**	**99.83**	**99.67**	**99.77**	98.28	**99.89**	97.31	99.09

Table 8
Ablation studies on the benchmark with CIFAR-10 as ID data.

Ablation Study		$\alpha = 0$	$\alpha = 1$	$\gamma = 0$	$m = 0$	NsED
Average Performance	AUROC↑	83.28	81.34	79.75	79.75	**87.74**
	FPR95↓	54.40	63.21	64.28	64.28	**49.59**
	AUPR-In↑	71.87	68.31	65.91	65.91	**79.07**
	AUPR-Out↑	87.86	85.45	84.69	84.69	**90.27**

6.3. Comparison with generative adversarial networks-based OOD detection

Generative adversarial network-based OOD detection, known as GAN-based OOD detection, is a significant aspect of OOD detection. In this study, we conducted experiments to compare our methods NsED+MSP and NsED+Energy with representative GAN-based OOD detection methods, namely ConfGAN [42], BoundaryGAN [43] and CMG [44]. ConfGAN utilizes GAN to explicitly generate OOD data that the classifier is confident about, i.e., data with low entropy. Additionally, the classifier maximizes the entropy for this generated data. BoundaryGAN proposes the inclusion of two additional terms to the original loss function, such as cross entropy. The first term aims to reduce the confidence of the classifier when dealing with OOD data, while the second term focuses on generating more effective training data for the first term. Essentially, BoundaryGAN simultaneously trains both the classification and generative neural networks to handle OOD scenarios. CMG generates pseudo OOD data by incorporating abnormal conditions as mixed class embeddings into a conditional variational auto-encoder. This data is subsequently utilized to fine-tune a classifier constructed with the provided ID data.

Tables 6 and 7 present the comparison results. In general, it can be observed that NsED+MSP or NsED+Energy consistently outperform ConfGAN, BoundaryGAN, and CMG. One possible explanation for this is that the generated OOD data, which is based on ID data, is inadequate for addressing OOD detection, since ID data lacks any information about real OOD data.

6.4. Parameter analysis

In this study, we conduct experiments on the CIFAR-10 benchmark to demonstrate that a wide range of parameter values can yield satisfactory performance. We specifically evaluate five main parameters: α, γ, the number of styles m, and the inner iterative step L. We then report the average results for each of baseline methods—MSP, ODIN, Energy, GradNorm, React, and KLM. For these tests, we use CIFAR-10 as the ID dataset and employ CIFAR-100, TinyImageNet, MNIST, SVHN, Textures, and Places365 as OOD datasets. Figs. 5, 6, and 7 provide an analysis of the parameters α, γ, m, and L.

- α and γ. The parameter α balances the empirical risk minimization strategy (Eq. (3)) and semantic robustness minimization strategy (Eq. (5)). γ, introduced in Eq. (7), adjusts the degree of change on non-semantic factors. We conduct parameter analysis with varying $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and $\gamma \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$. Experiments illustrated in Fig. 5 imply that higher values of α and γ lead to improved detection performance. Generally, NsED consistently enhances the performance of each baseline method, given that $\alpha \in \{0.3, 0.5, 0.7, 0.9\}$ and $\gamma \in \{0.8, 1.0\}$.
- m. The parameter m presents the number of styles. A larger m implies that more styles are used to create new ones. We conduct experiments with varying m from set $\{3, 6, 9, 12, 15, 18\}$. The experiments depicted in Fig. 6 suggest that larger values of m do not necessarily result in better detection performance. In general, optimal performance may be achieved when $m = 6$, while the worst performance may be observed at $m = 15$ and 18. Moreover, with m values within the set $\{3, 6, 9, 12, 15, 18\}$, NsED consistently enhances the performance of each baseline method.
- L. The parameter L is the inner iterative step and represents the number of steps for exploring worst-case non-semantic factors in Eq. (10). We perform experiments with different values of L selected from the set $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 15\}$. A larger value of L enables us to more effectively explore the worst-case non-semantic factors. From Fig. 7, it can be observed that there is a stable increase in the detection performance as L is increased from 1 to 15.

6.5. Ablation studies

In this section, we conduct ablation studies on the CIFAR-10 benchmark to demonstrate the contribution of the individual components in NsED, as shown in Table 8. We consider the following baselines: 1) $\alpha = 0$: train models without employing the empirical risk $\frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{g}_{\mathbf{w}}(\mathbf{x}_i), y_i)$ in Eq. (6) in NsED; 2) $\alpha = 1$: train models without employing the empirical risk of semantic robustness $\frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{g}_{\mathbf{w}}(\mathbf{x}_i^{\beta_{L,\gamma}}), y_i)$ in Eq. (6) in NsED; 3) $\gamma = 0$: train models without using convex combination with non-semantic factors in Eq. (7) in NsED; and 4) $m = 0$: generate non-semantic factors without introducing additional novel non-semantic factors in NsED.

- When setting α to 0, the average AUROC/FPR95/AUPR-In/AUPR-Out performance of different OOD detection methods (MSP, ODIN, Energy, GradNorm, React, and KLM) drops from 87.74/49.59/79.07/90.27 to 83.28/54.40/71.87/87.86. This decrease indicates the significance of the empirical risk $\frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{g}_{\mathbf{w}}(\mathbf{x}_i), y_i)$ presented in Eq. (6).
- When α is set to 1, the average AUROC/FPR95/AUPR-In/AUPR-Out performance of different OOD detection methods decreases from 87.74/49.59/79.07/90.27 to 81.34/63.21/68.31/85.45. This decline emphasizes the significance of the empirical risk associated with semantic robustness, as presented in Eq. (6).

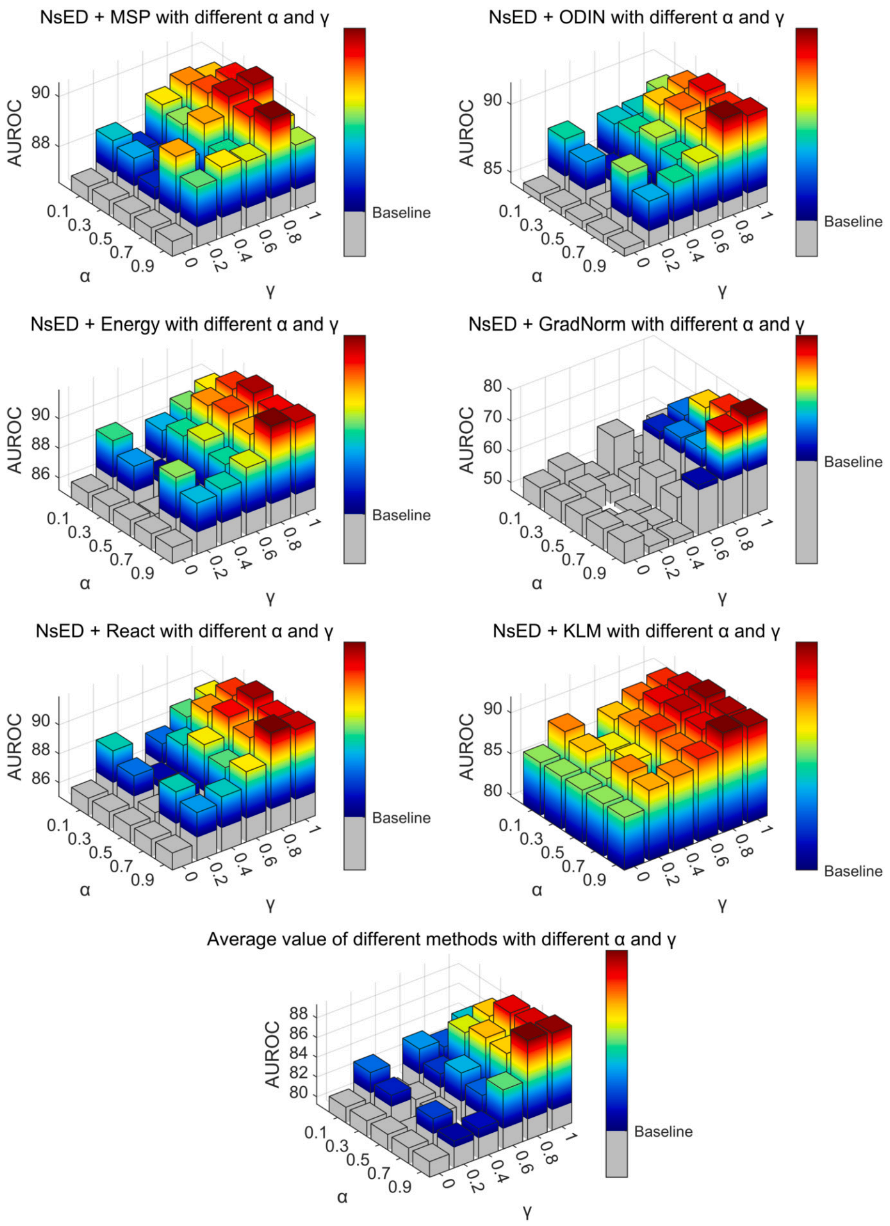

Fig. 5. The parameter analysis for proposed method NsED on α and γ.

- By setting γ to 0, the average AUROC/FPR95/AUPR-In/AUPR-Out performance of different OOD detection methods significantly decreases to 79.75/64.28/65.91/84.69, compared to the average AUROC/FPR95/AUPR-In/AUPR-Out of 87.74/49.59/79.07/90.27 achieved by NsED. This result emphasizes the substantial performance improvement of NsED through the convex combination with non-semantic factors presented in Eq. (7).
- When $m = 0$, the average AUROC/FPR95/AUPR-In/AUPR-Out performance of NsED decreases from 87.74/49.59/79.07/90.27 to 79.75/64.28/65.91/84.69. This outcome highlights the importance of incorporating additional novel non-semantic factors.

7. Conclusion

The paper proposes a novel method, called NsED, to mitigate the spurious correlation issue and improve the performance of OOD detection. NsED enhances post-hoc OOD detection methods by leveraging non-semantic factors to learn semantic-invariant features and distinguish them from spurious features. Additionally, the paper investigates the relationship between single-domain OOD generalization and OOD detection. By using NsED, the generalization ability of deep models can be enhanced. The effectiveness of NsED is evaluated through extensive experiments, demonstrating its robustness to different types of OOD data. Furthermore,

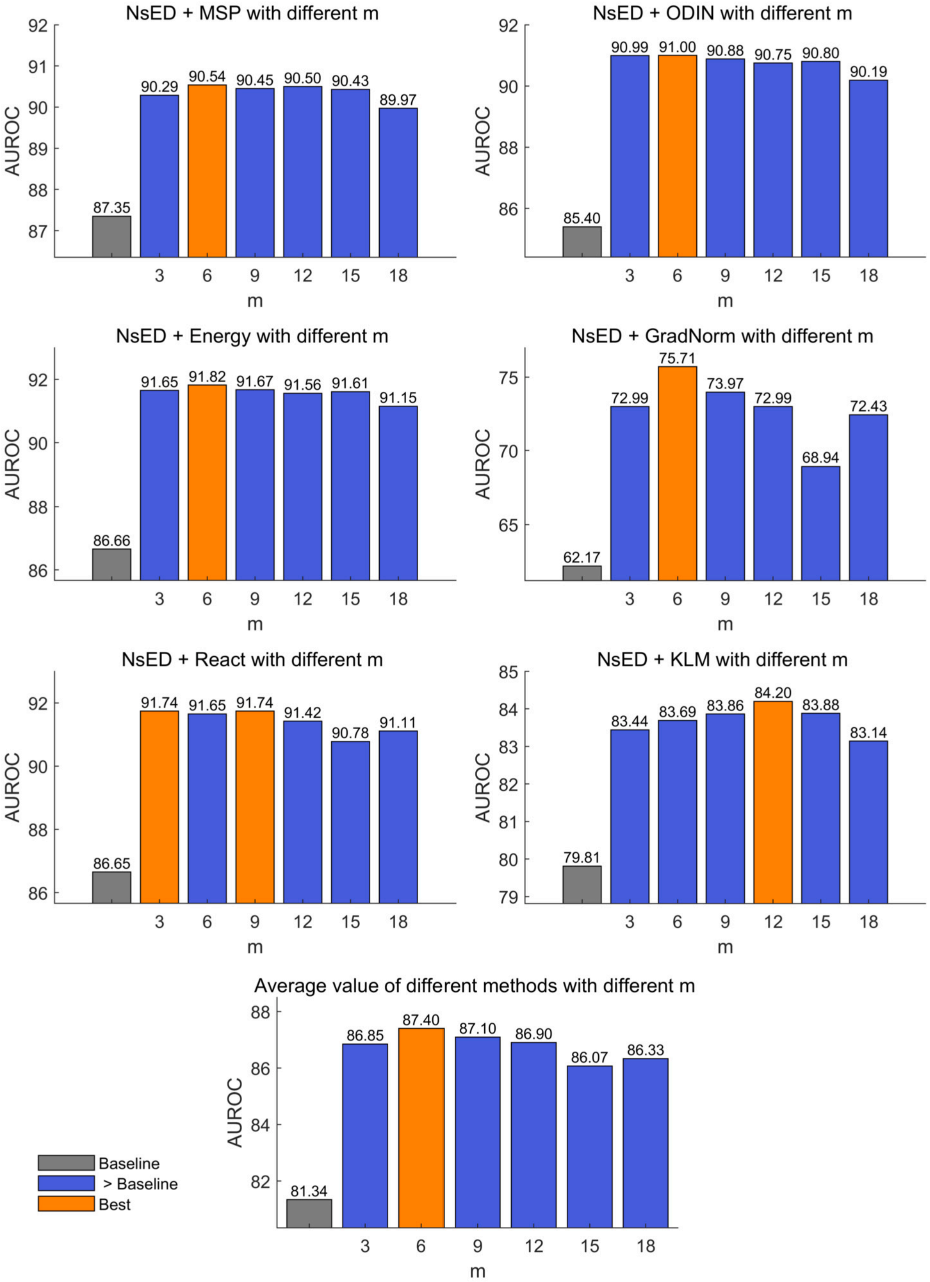

Fig. 6. The parameter analysis for proposed method NsED on the number of styles m.

parameter analysis and ablation studies highlight the contribution of each component of NsED to the improvement of OOD detection performance.

Looking ahead, the research presents several avenues for further study. Firstly, while NsED has demonstrated remarkable success in mitigating the effects of the spurious correlation issue in OOD detection, there are many real-world cases where semantic and non-semantic factors are intricately linked, rendering their strict separation impossible. Therefore, it is essential to employ fuzzy theory to represent the relationship between these factors [45]. In our future work, we intend to extract the fuzzy structure and further refine our method, while also exploring theories that bridge the gap between OOD detection and generalization. Secondly, this paper serves as an initial exploration into the relationship between OOD detection and OOD generalization. However, a more in-depth investigation is needed to provide additional theoretical grounding and empirical findings. Moreover, it is intriguing and promising to theoretically and empirically explore the detection ability of foundation models such as GPT, which is known for its strong generalization ability. Lastly, as data in real-world scenarios often arrives in a streaming form, there is potential in exploring OOD detection using streaming ID and OOD data [46,47].

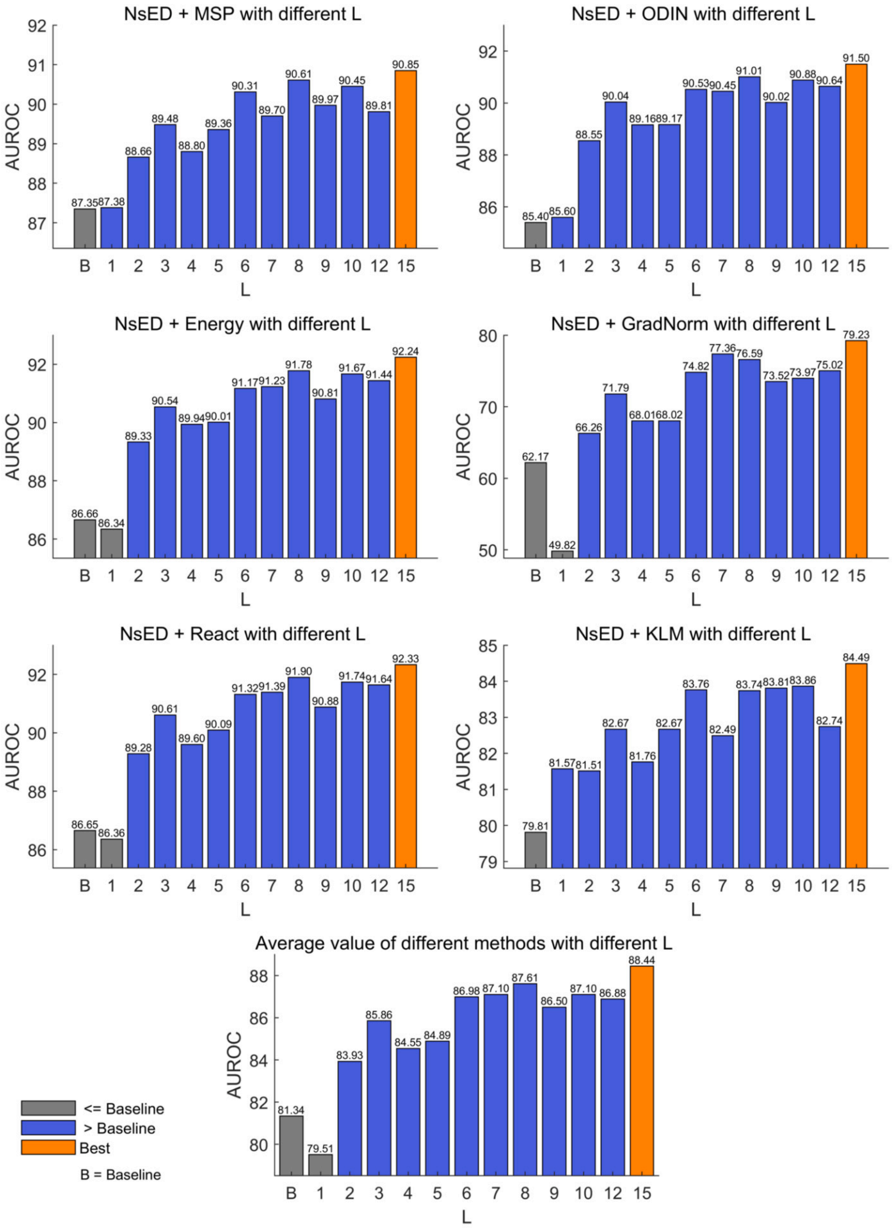

Fig. 7. The parameter analysis for proposed method NsED on inner iterative step L.

CRediT authorship contribution statement

Zhen Fang: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Jie Lu:** Supervision, Writing – review & editing. **Guangquan Zhang:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The work presented in this paper was supported by the Australian Research Council (ARC) FL190100149.

Data availability

The data that has been used is confidential.

References

[1] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images, 2009.

[2] J. Yang, P. Wang, D. Zou, Z. Zhou, K. Ding, W. Peng, H. Wang, G. Chen, B. Li, Y. Sun, et al., Openood: benchmarking generalized out-of-distribution detection, in: Advances in Neural Information Processing Systems Datasets and Benchmarks Track, 2022.

[3] P. Oberdiek, M. Rottmann, H. Gottschalk, Classification uncertainty of deep neural networks based on gradient information, in: Artificial Neural Networks in Pattern Recognition, vol. 11081, Springer, 2018, pp. 113–125.

[4] Z. Fang, Y. Li, J. Lu, J. Dong, B. Han, F. Liu, Is Out-of-Distribution Detection Learnable?, Advances in Neural Information Processing Systems, 2022.

[5] W. Liu, X. Wang, J. Owens, Y. Li, Energy-Based Out-of-Distribution Detection, Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 21464–21475.

[6] R. Huang, A. Geng, Y. Li, On the Importance of Gradients for Detecting Distributional Shifts in the Wild, Advances in Neural Information Processing Systems, vol. 34, 2021, pp. 677–689.

[7] P. Izmailov, P. Kirichenko, N. Gruver, A.G. Wilson, On feature learning in the presence of spurious correlations, in: Advances in Neural Information Processing Systems, 2022.

[8] K.Y. Xiao, L. Engstrom, A. Ilyas, A. Madry, Noise or signal: the role of image backgrounds in object recognition, in: International Conference on Learning Representations, 2021.

[9] Y. Ming, H. Yin, Y. Li, On the impact of spurious correlation for out-of-distribution detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021.

[10] L.H. Zhang, R. Ranganath, Robustness to spurious correlations improves semantic out-of-distribution detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2023.

[11] Z. Shen, J. Liu, Y. He, X. Zhang, R. Xu, H. Yu, P. Cui, Towards out-of-distribution generalization: a survey, CoRR, arXiv:2108.13624, 2021.

[12] Q. Xu, R. Zhang, Y. Zhang, Y. Wang, Q. Tian, A Fourier-based framework for domain generalization, in: Conference on Computer Vision and Pattern Recognition, 2021, pp. 14378–14387.

[13] S. Liang, Y. Li, R. Srikant, Enhancing the reliability of out-of-distribution image detection in neural networks, in: International Conference on Learning Representations, 2018.

[14] Y. Sun, C. Guo, Y. Li, React: Out-of-Distribution Detection with Rectified Activations, Advances in Neural Information Processing Systems, vol. 34, 2021, pp. 144–157.

[15] D. Hendrycks, S. Basart, M. Mazeika, A. Zou, J. Kwon, M. Mostajabi, J. Steinhardt, D. Song, Scaling out-of-distribution detection for real-world settings, in: International Conference on Machine Learning, 2022, pp. 8759–8773.

[16] D. Hendrycks, K. Gimpel, A baseline for detecting misclassified and out-of-distribution examples in neural networks, in: International Conference on Learning Representations, 2017.

[17] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, F. Huang, A tutorial on energy-based learning, 2006.

[18] Y. Sun, Y. Ming, X. Zhu, Y. Li, Out-of-distribution detection with deep nearest neighbors, in: International Conference on Machine Learning, 2022, pp. 20827–20840.

[19] D. Mahajan, S. Tople, A. Sharma, Domain generalization using causal matching, in: International Conference on Machine Learning, 2021, pp. 7313–7324.

[20] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, A. Courville, Out-of-distribution generalization via risk extrapolation (rex), in: International Conference on Machine Learning, 2021, pp. 5815–5826.

[21] A.T. Nguyen, T. Tran, Y. Gal, A.G. Baydin, Domain Invariant Representation Learning with Domain Density Transformations, Advances in Neural Information Processing Systems, vol. 34, 2021, pp. 5264–5275.

[22] M. Arjovsky, L. Bottou, I. Gulrajani, D. Lopez-Paz, Invariant risk minimization, CoRR, arXiv:1907.02893, 2019.

[23] I. Gulrajani, D. Lopez-Paz, In search of lost domain generalization, in: International Conference on Learning Representations, 2021.

[24] C. Lu, Y. Wu, J.M. Hernández-Lobato, B. Schölkopf, Nonlinear invariant risk minimization: a causal approach, arXiv preprint, arXiv:2102.12353, 2021.

[25] S. Shankar, V. Piratla, S. Chakrabarti, S. Chaudhuri, P. Jyothi, S. Sarawagi, Generalizing across domains via cross-gradient training, arXiv preprint, arXiv:1804.10745, 2018.

[26] K. Zhou, Y. Yang, T. Hospedales, T. Xiang, Learning to generate novel domains for domain generalization, in: European Conference on Computer Vision, 2020, pp. 561–578.

[27] F. Lv, J. Liang, S. Li, B. Zang, C.H. Liu, Z. Wang, D. Liu, Causality inspired representation learning for domain generalization, in: Conference on Computer Vision and Pattern Recognition, 2022, pp. 8046–8056.

[28] C. Zhang, K. Zhang, Y. Li, A Causal View on Robustness of Neural Networks, Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 289–301.

[29] S. Shalev-Shwartz, S. Ben-David, Understanding Machine Learning - from Theory to Algorithms, Cambridge University Press, 2014.

[30] F. Qiao, X. Peng, Topology-aware robust optimization for out-of-distribution generalization, in: International Conference on Learning Representations, 2023.

[31] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, arXiv preprint, arXiv:1312.6199, 2013.

[32] J. Yang, P. Wang, D. Zou, Z. Zhou, K. Ding, W. Peng, H. Wang, G. Chen, B. Li, Y. Sun, X. Du, K. Zhou, W. Zhang, D. Hendrycks, Y. Li, Z. Liu, Openood: benchmarking generalized out-of-distribution detection, in: Advances in Neural Information Processing Systems, 2022.

[33] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1106–1114.

[34] L. Deng, The MNIST database of handwritten digit images for machine learning research [best of the web], IEEE Signal Process. Mag. 29 (6) (2012) 141–142.

[35] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A.Y. Ng, Reading digits in natural images with unsupervised feature learning, in: NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011, 2011.

[36] G. Kylberg, The Kylberg Texture Dataset v. 1.0, External report (Blue series) 35, Centre for Image Analysis, Swedish University of Agricultural Sciences and Uppsala University, Uppsala, Sweden, 2011.

[37] B. Zhou, À. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: a 10 million image database for scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2018) 1452–1464.

[38] Y. Bulatov, Notmnist Dataset, Google (Books/OCR), Tech. Rep. 2, 2011.

[39] H. Xiao, K. Rasul, R. Vollgraf, Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, CoRR, arXiv:1708.07747, 2017.

[40] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[41] Y. LeCun, B.E. Boser, J.S. Denker, D. Henderson, R.E. Howard, W.E. Hubbard, L.D. Jackel, Handwritten Digit Recognition with a Back-Propagation Network, Advances in Neural Information Processing Systems, 1989, pp. 396–404.

[42] K. Sricharan, A. Srivastava, Building robust classifiers through generation of confident out of distribution examples, in: Third Workshop on Bayesian Deep Learning, Conference on Neural Information Processing Systems 2018, 2018.

[43] K. Lee, H. Lee, K. Lee, J. Shin, Training confidence-calibrated classifiers for detecting out-of-distribution samples, in: 6th International Conference on Learning Representations, 2018, OpenReview.net.

[44] M. Wang, Y. Shao, H. Lin, W. Hu, B. Liu, CMG: a class-mixed generation approach to out-of-distribution detection, in: Machine Learning and Knowledge Discovery in Databases, in: Lecture Notes in Computer Science, vol. 13716, Springer, 2022, pp. 502–518.

[45] T. Tan, T. Zhao, A data-driven fuzzy system for the automatic determination of fuzzy set type based on fuzziness, Inf. Sci. (2023).

[46] H. Yu, Q. Zhang, T. Liu, J. Lu, Y. Wen, G. Zhang, Meta-add: a meta-learning based pre-trained model for concept drift active detection, Inf. Sci. (2022).

[47] B. Zou, K. Yang, X. Kui, J. Liu, S. Liao, W. Zhao, Anomaly detection for streaming data based on grid-clustering and Gaussian distribution, Inf. Sci. (2022).