International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02139-y

Visual Out-of-Distribution Detection in Open-Set Noisy Environments

Rundong He[1] · Zhongyi Han[2] · Xiushan Nie[3] · Yilong Yin[1] · Xiaojun Chang[2,4]

Received: 14 October 2023 / Accepted: 31 May 2024

Abstract
The presence of noisy examples in the training set inevitably hampers the performance of out-of-distribution (OOD) detection. In this paper, we investigate a previously overlooked problem called OOD detection under asymmetric open-set noise, which is frequently encountered and significantly reduces the identifiability of OOD examples. We analyze the generating process of asymmetric open-set noise and observe the influential role of the confounding variable, entangling many open-set noisy examples with partial in-distribution (ID) examples referred to as hard-ID examples due to spurious-related characteristics. To address the issue of the confounding variable, we propose a novel method called Adversarial Confounder REmoving (ACRE) that utilizes progressive optimization with adversarial learning to curate three collections of potential examples (easy-ID, hard-ID, and open-set noisy) while simultaneously developing invariant representations and reducing spurious-related representations. Specifically, by obtaining easy-ID examples with minimal confounding effect, we learn invariant representations from ID examples that aid in identifying hard-ID and open-set noisy examples based on their similarity to the easy-ID set. By triplet adversarial learning, we achieve the joint minimization and maximization of distribution discrepancies across the three collections, enabling the dual elimination of the confounding variable. We also leverage potential open-set noisy examples to optimize a K+1-class classifier, further removing the confounding variable and inducing a tailored K+1-Guided scoring function. Theoretical analysis establishes the feasibility of ACRE, and extensive experiments demonstrate its effectiveness and generalization. Code is available at https://github.com/Anonymous-re-ssl/ACRE0.

Keywords Out-of-distribution detection · Asymmetric open-set noise · Open-world visual recognition · Adversarial confounder removing

1 Introduction

Ensuring the reliability of machine learning models during real-world deployments is crucial, and out-of-distribution (OOD) detection plays a vital role in achieving this goal (Fang et al., 2022; Nguyen et al., 2015; Yang et al., 2021b; Hell et al., 2021). Pioneering studies have demonstrated impressive performance when the test data belongs to classes that are not seen during training (Yang et al., 2021b; Hendrycks and Gimpel, 2017; Liang et al., 2017; Liu et al., 2020a; Sun et al., 2021a). In practice, this achievement is contingent upon the essential prerequisite that the labeled training in-distribution (ID) data are devoid of any data noise.

The presence of noisy examples in the training set inevitably hampers the effectiveness of OOD detection (Wu et al., 2021). While training-time OOD detection methods make adjustments to the model training process, the inclusion of noisy examples in the training set can disrupt classification boundaries (Hendrycks et al., 2018; Chen et al., 2021; He et

Communicated by ZHUN ZHONG.

✉ Zhongyi Han
hanzhongyicn@gmail.com

Rundong He
rundong_he@mail.sdu.edu.cn

Xiushan Nie
niexsh@hotmail.com

Yilong Yin
ylyin@sdu.edu.cn

Xiaojun Chang
cxj273@gmail.com

1 School of Software, Shandong University, Jinan, China

2 Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates

3 School of Computer Science and Technology, Shandong Jianzhu University, Jinan, China

4 Australian Artificial Intelligence Institute, University of Technology Sydney, Ultimo, Australia

Published online: 16 June 2024

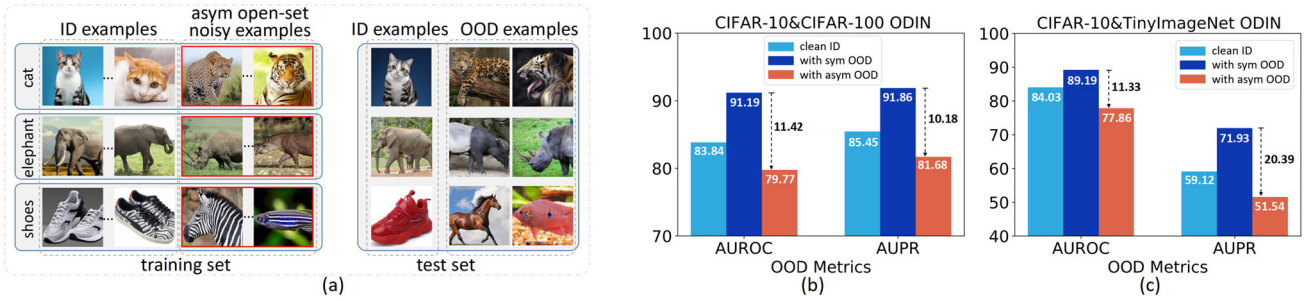

Fig. 1 Motivation illustration. **a** the problem of OOD detection under asymmetric open-set noise. **b**, **c** asymmetric open-set noise in the training set poses a greater detriment than the symmetric. AUROC and AUPR are metrics for OOD detection. The larger the AUROC and AUPR, the better the performance of the OOD detection

al., 2022a). Conversely, test-time OOD detection methods heavily rely on trained models and design an OOD scoring function for identifying unseen classes without access to the training data (Hendrycks and Gimpel, 2017; Liang et al., 2017; Liu et al., 2020a). Accordingly, their performance is irreversibly compromised when the noisy datasets interfere with the prediction uncertainties of trained models, particularly if the noisy dataset contains open-set noisy examples that closely resemble the upcoming OOD examples.

To address the challenge of OOD detection in the presence of noisy training examples, a few training-time OOD detection methods have been proposed (Wu et al., 2021; Yu and Aizawa, 2020; Wei et al., 2021). These methods primarily focus on addressing closed-set noise (e.g., label noise) and symmetric open-set noise. Symmetric open-set noise occurs when examples from unobserved classes are distributed across ID classes in a random manner without considering the characteristics of the examples. We argue that this problem can be effectively resolved by combining existing techniques in noisy label learning and OOD detection (Wu et al., 2021; Wei et al., 2021). Pioneering studies have also shown that symmetric open-set noise can even provide benefits for OOD detection (Hendrycks et al., 2018; Ming et al., 2022). Our experiments show that existing methods are limited in handling OOD detection in scenarios with previously overlooked asymmetric open-set noise. Asymmetric open-set noise frequently occurs when OOD classes are distributed towards the ID classes that share spurious-related characteristics with them, as depicted in Fig. 1a. Compared with symmetric open-set noise, the presence of asymmetric open-set noise significantly decreases the OOD detection performance by reducing the identifiability of OOD examples, as depicted in Fig. 1b and c.

We delve into the data-generating process of asymmetric open-set noise using structural causal models (Pearl, 2009). We observe that the confounding variable plays a significant role in entangling many open-set noisy examples with a subset of ID examples, referred to as "hard-ID examples", due to spurious-related characteristics. For instance, consider an ID dataset that includes the shoe class while the open-set classes encompass zebra-related classes in Fig. 1a. Due to the presence of zebra-like stripes on a subset of shoe examples, the data collection process (e.g., from web sources) often leads to misclassifying examples from the zebra class as belonging to the shoe class (Han et al., 2022a). In this scenario, the shoe examples with zebra stripes represent hard-ID examples, while the other shoes constitute easy-ID examples. The zebra stripes act as a potentially confounding variable, reducing the identifiability of OOD examples, especially when the OOD data comprise zebra-related classes.

In this paper, we introduce Adversarial Confounder REmoving (ACRE) as a solution to eliminating the confounding variable. ACRE employs progressive optimization with adversarial learning to curate three distinct collections of potential examples: easy-ID, hard-ID, and open-set noisy. Inspired by domain-invariant representation learning (Nguyen et al., 2021; Han et al., 2022b; Wang et al., 2022a; Jang et al., 2022), this approach allows us to simultaneously develop invariant representations from ID examples while reducing spurious-related representations from open-set noisy examples. To begin, we can efficiently identify the easy-ID examples with minimal effect from confounding variable by small loss criteria (Han et al., 2018; Jiang et al., 2018). From these examples, we learn invariant representations that help differentiate between hard-ID and open-set noisy examples based on their similarity to the easy-ID set. Simultaneously, we employ triplet adversarial learning to facilitate the minimization and maximization of distribution discrepancies across the three collections, successfully achieving the dual elimination of the confounding variable. In addition, we leverage potential open-set noisy examples to optimize a $K+1$-class classifier. This process aids in removing the confounding variable and induces a tailored $K+1$-Guided scoring function. We give a theoretical analysis to verify the feasibility of triplet adversarial learning in confounder removal.

In the field of computer vision, the concept of Open-World Visual Recognition has emerged as a critical area of research. Its aim is to develop resilient systems capable of handling real-world scenarios, where out-of-distribution (OOD) data is common. Our research focuses on Visual Out-of-Distribution Detection in Open-Set Noisy Environments, addressing the urgent need for algorithms that can effectively identify OOD instances within complex real-world visual data. By introducing innovative methodologies and insights to the field of Open-World Visual Recognition, our work aims to advance the frontier of visual recognition systems towards greater adaptability and robustness, to cope with dynamic and uncertain environments.

Our contributions can be summarized as follows:

- We introduce and investigate the problem of OOD detection under asymmetric open-set noise, which accommodates a variety of real-world applications but is unexplored.
- We propose a novel method called ACRE that employs an adversarial learning approach to remove the confounding variable between open-set noisy examples and hard-ID examples, resulting in improved OOD detection performance.
- Theoretical analysis and empirical results demonstrate the feasibility and effectiveness of ACRE on real-world datasets, and ACRE can pave a solid baseline for future studies.

2 Related Work

2.1 Out-of-Distribution Detection

The ability to distinguish between in-distribution (ID) and out-of-distribution (OOD) data is a fundamental concern for deploying machine learning models in real-world applications. OOD detection methods can be broadly classified into two main categories: classification-based methods (Hendrycks and Gimpel, 2017; Liang et al., 2017; Lee et al., 2018; Liu et al., 2020a; Gomes et al., 2022; Sun et al., 2022a; Ming and Li, 2023; Yang et al., 2023), density-based methods (Ren et al., 2019; Xiao et al., 2020; Morningstar et al., 2021; Zhou and Levine, 2021; Jiang et al., 2021; Zhang et al., 2021). Classification-based methods for detecting out-of-distribution (OOD) data involve modeling the conditional distribution of the in-distribution (ID) data, and then designing a scoring function to measure the uncertainty of test data. Density-based methods model the ID distribution using probabilistic models and consider test data in low-density regions as OOD data. Density-based OOD detection methods can be difficult to train and optimize, often yielding inferior performance compared to classification-based methods (Yang et al., 2021b). Therefore, in this paper, we focus on classification-based methods. Within this category, there are two main branches of research: testing-time methods (Hendrycks and Gimpel, 2017; Liang et al., 2017; Lee et al., 2018; Liu et al., 2020a; Wang et al., 2022b; Zhu et al., 2022; Song et al., 2022a; He et al., 2024b) and training-time methods (Ming et al., 2023; Ming and Li, 2023; Yang et al., 2023; Du et al., 2023; He et al., 2024ba). Test-time methods are easy to use without modifying the training procedure and objective (Yang et al., 2021b).

Unlike testing-time methods, training-time methods aim to mitigate overconfident predictions for OOD data during the training period. According to whether the OOD-supervised signals are used in the training process, training-time methods can be categorized into OOD-free and OOD-needed methods. The representatives of OOD-free methods are Wei et al. (2022), Lin et al. (2021). Wei et al. (2022) decoupled the influence of logits' norm from the training procedure by incorporating LogitNorm into the cross-entropy loss. Lin et al. (2021) exploited intermediate classifier outputs for dynamic and efficient OOD inference. The OOD-needed methods aim to calibrate the model by OOD-supervised signals, which are from auxiliary OOD datasets (Hendrycks et al., 2018; Liu et al., 2020a; Chen et al., 2021; Ming et al., 2022; Wang et al., 2023), unlabeled data He et al. (2022b), Yu and Aizawa (2019), Yang et al. (2021a), Zhou et al. (2021), Katz-Samuels et al. (2022), He et al. (2024ba), or synthetic virtual OOD data (Du et al., 2022; Tang et al., 2021; Tack et al., 2020; He et al., 2022a; Du et al., 2023).

Nevertheless, the OOD detection methods commonly used in representative works assume an impeccable learning environment in which the labeled ID data is noise-free. In real-world applications, however, this assumption is often unattainable, which can severely compromise the robustness of these methods. There are only three pioneering works (Yu and Aizawa, 2020; Wei et al., 2021; Wu et al., 2021) that leverage label cleaning, geometric structure or injects open-set auxiliary data to enhance OOD detection. Although these methods have addressed the problem of OOD detection under noisy environments, the open-set noise they consider is symmetric. According to Hendrycks et al. (2018), Ming et al. (2022), Wang et al. (2023), we can know that symmetric open-set noise is helpful for OOD detection. However, in many real-world scenarios, open-set noise is not symmetric (random) but rather asymmetric (dependent). According to Fig. 1 in main body, we can know that asymmetric open-set noise is harmful to OOD detection severely. Our paper aims to address the issue of **OOD detection under asymmetric**

open-set noise, which is a highly challenging and valuable problem that has received relatively little attention.

2.2 Learning from Noisy Labels

Previous works on learning from noisy labels can be classified into three categories: label-based, sample-based, and loss-based methods. Early methods focused on correcting corrupted labels by estimating the noise transition matrix Patrini et al. (2017), Goldberger and Ben-Reuven (2017), but this approach is challenging due to the difficulty in accurately estimating the matrix. Sample-based methods (Han et al., 2018; Wei et al., 2020; Yao et al., 2023) aim to select representative samples for training, while loss-based methods (Reed et al., 2014; Zhang and Sabuncu, 2018) focus on using robust loss functions to improve model performance. However, these methods are only designed for close-set noise in the training set. More recent works (Wang et al., 2018; Sun et al., 2020; Yu and Aizawa, 2020; Sachdeva et al., 2021; Li et al., 2020; Yao et al., 2021; Li et al., 2021; Xia et al., 2022; Sun et al., 2022b; Wei et al., 2021; Wan et al., 2024) propose to handle both in-distribution (IND) and out-of-distribution (OOD) noise in training datasets. However, these approaches are not directly applicable for detecting OOD data at test time. Combining them with existing OOD detection methods may not yield satisfactory performance (Wu et al., 2021).

3 Methodology

In this section, we first introduce the learning set-up. Then, we point out the key insight by the structural causal model. According to the insight, we present adversarial confounder removing.

3.1 Introduction of Different Noise Type

We introduce a classification of different noise types, as shown in Table 1. Then, we introduce the different noise types in detailed as follows:

- **Symmetric Close-Set Noise:** The noise belongs to one of the known categories, but the label is incorrect. Moreover, the distribution of noise is uniform, meaning that samples from every category have the same probability of being incorrectly labeled as another category. This noise model does not favor any specific mislabeling pattern, and is therefore considered "symmetric."
- **Asymmetric Close-Set Noise:** The noise belongs to one of the known categories, but the label is incorrect. Moreover, the distribution of noise is uneven, with samples from certain categories more likely to be mislabeled as specific other categories. This type of noise usually occurs between categories that are similar or easily confused with one another, for instance, mislabeling a known wolf category as a known dog category is more common than mislabeling it as a known cat category.
- **Symmetric Open-Set Noise:** The noisy samples do not belong to any known category in the training set, and these samples are evenly distributed across known categories.
- **Asymmetric Open-Set Noise:** The noisy samples do not belong to any known category in the training set, but these samples have an uneven probability of being misclassified into specific known categories. For example, in an animal classification task, there may be samples from new animal categories not included in the training set, and these samples are more likely to be classified into specific categories that resemble them in appearance or ecological characteristics.

Table 1 Classification of different noise types

	Symmetric	Asymmetric
Close-Set	Sym Close-Set	Asym Close-Set
Open-Set	Sym Open-Set	Asym Open-Set

3.2 Problem Set-Up

We consider a noisy training set $\mathcal{D}_{in}^{train} = \{(\mathbf{x}_k, \tilde{\mathbf{y}}_k)\}_{k=1}^{n}$ where $\mathbf{x}_k \in \mathcal{X}$, $\tilde{\mathbf{y}}_k \in \mathcal{Y}$, $\mathcal{X}$ denotes the input space, $\mathcal{Y}$ denotes the ID label space, $\mathcal{Y} = \{1, 2, \ldots, K\}$, K denotes the number of ID classes, and n denotes the number of examples in $\mathcal{D}_{in}^{train}$. We assume the examples from ID classes are clean. In our setting of OOD detection under asymmetric open-set noise, $\mathcal{D}_{in}^{train}$ contains two types of example: (1) **ID example** $\mathbf{x}_i$ whose assigned label $\mathbf{y}_i$ is the same as the ground-truth label $\mathbf{y}_k^*$ and $\mathbf{y}_k^* \in \mathcal{Y}$; (2) **Asymmetric open-set noise example** $\mathbf{x}_o$ whose assigned label $\tilde{\mathbf{y}}_k$ does not equal to the ground-truth label $\mathbf{y}_k^*$, $\tilde{\mathbf{y}}_k \in \mathcal{Y}$ but $\mathbf{y}_k^* \notin \mathcal{Y}$, and $\tilde{\mathbf{y}}_k$ is assigned based on similarity to ID classes.

Let $\mathcal{D}^{test}$ denote the test set, which consists of ID test set $\mathcal{D}_{in}^{test}$ and OOD test set $\mathcal{D}_{out}^{test}$. The example $\mathbf{x}_{it}$ in $\mathcal{D}_{in}^{test}$ is from the ID classes. The example $\mathbf{x}_{ot}$ in $\mathcal{D}_{out}^{test}$ is from the unknown classes. The goal of OOD detection is to define a decision function $\mathcal{F}$ such that for a given test input $\mathbf{x} \in \mathcal{D}^{test}$,

$$\mathcal{F}(\mathbf{x}) = \begin{cases} 0 & \text{if } \mathbf{x} \in \mathcal{D}_{out}^{test}, \\ 1 & \text{if } \mathbf{x} \in \mathcal{D}_{in}^{test}, \end{cases} \tag{1}$$

where $\mathcal{F}(\mathbf{x}) = 1$ means that $\mathbf{x}$ is ID data and $\mathcal{F}(\mathbf{x}) = 0$ means that $\mathbf{x}$ is OOD data.

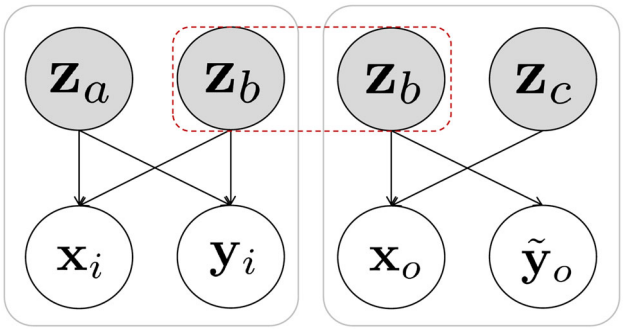

Fig. 2 The training data generating process: The gray shade of nodes indicates that the variables are unobservable

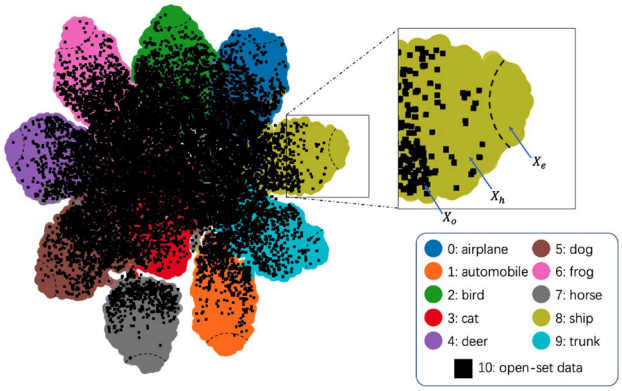

Fig. 3 The t-SNE visualization

3.3 Problem Analysis

We reveal a generating process of the training set by Fig. 2, where the shaded variables are unobservable, and the unshaded variables are observable. The left graph in Fig. 2 and Eq. (2) reveal the generating process of ID data.

$$\mathbf{z}_a \sim p_{\mathbf{z}_a}, \mathbf{z}_b \sim p_{\mathbf{z}_b}, \mathbf{x}_i = g\left(\mathbf{z}_a, \mathbf{z}_b\right) . \tag{2}$$

In the generating process of ID data, we assume that ID data $\mathbf{x}_i \in \mathcal{X}$ is generated by latent variable $\mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^m$ through a function $g : \mathcal{Z} \to \mathcal{X}$. We partition latent variable $\mathbf{z}$ into two variables $\mathbf{z} = [\mathbf{z}_a, \mathbf{z}_b]$. $\mathbf{z}_a$ denotes the private variable which only $\mathbf{x}_i$ owns. $\mathbf{z}_b$ denotes the confounding variable which both $\mathbf{x}_i$ and $\mathbf{x}_o$ own simultaneously. Further, we assume that $\mathbf{y}_i$ is generated by the private variable $\mathbf{z}_a$ and the confounding variable $\mathbf{z}_b$.

The right graph in Fig. 2 and Eq. (3) reveal the generating process of asymmetric open-set noise.

$$\mathbf{z}_b \sim p_{\mathbf{z}_b}, \mathbf{z}_c \sim p_{\mathbf{z}_c}, \mathbf{x}_o = g\left(\mathbf{z}_b, \mathbf{z}_c\right) . \tag{3}$$

In the generating process of asymmetric open-set noise, we assume that open-set data $\mathbf{x}_o \in \mathcal{X}$ is generated by latent variables $\mathbf{z}_b$ and $\mathbf{z}_c$. $\mathbf{z}_c$ denotes the private variable which only $\mathbf{x}_o$ owns. $\tilde{\mathbf{y}}_o$ denotes the noisy label of open-set data, and $\tilde{\mathbf{y}}_o$ is generated due to the biased influence $\mathbf{z}_b \to \mathbf{y}$.

During the inference stage, when the tested OOD example $\mathbf{x}_{ot}$ in D_{out}^{test} contains $\mathbf{z}_b$, $\mathbf{x}_{ot}$ is likely to be identified as ID example, resulting in poor performance of OOD detection. The existence of $\mathbf{z}_b$ is the essential reason for the performance decline of OOD detection. A pivotal insight is to **remove confounding variable** $\mathbf{z}_b$ to increase the separability of ID and OOD data, thus improving the performance of OOD detection.

We visualized the t-SNE graph of CIFAR-10 ID data contaminated with asymmetric open-set noise from CIFAR-100. In Fig. 3, the points labeled from '0' to '9' represent the features of ID examples, while the points labeled as '10' correspond to the features of asymmetric open-set noise examples. We can see that the ID data and open-set noise are difficult to separate due to their spurious-related features, acting as a confounding variable. This reduces the separability of ID and OOD data, leading to poor OOD detection performance. Based on Fig. 3, we also find that certain ID examples (at the edge) can be well separated from open-set examples, with low influence from confounding variable $\mathbf{z}_b$. These are easy-ID examples ($\mathbf{x}_e$), almost generated by $\mathbf{z}_a$. The remaining ID examples ($\mathbf{x}_h$) are hard-ID examples, generated by both $\mathbf{z}_a$ and $\mathbf{z}_b$. To address the confounding between $\mathbf{x}_h$ and $\mathbf{x}_o$, we propose adversarial confounder removing, which uses adversarial learning on $\mathbf{x}_e$, $\mathbf{x}_h$, and $\mathbf{x}_o$ to remove $\mathbf{z}_b$.

3.4 Adversarial Confounder Removing

Adversarial Confounder REmoving (ACRE) includes three components: 1) a triplet estimation module to obtain $\mathbf{x}_e$, $\mathbf{x}_h$, and $\mathbf{x}_o$; 2) a triplet adversarial learning module that uses adversarial learning on $\mathbf{x}_e$, $\mathbf{x}_h$, and $\mathbf{x}_o$ to remove $\mathbf{z}_b$; and 3) a K+1-Guided scoring function to detect OOD data. The network consists of three subnetworks: 1) feature extractor G; 2) two-head classifier (including a K-class classifier E and a K+1-class classifier C); and 3) discriminator D.

Our method includes a pre-training phase and a training phase. The pre-training phase primarily focuses on obtaining the initial feature extractor G, a K-classifier head E, and a K+1 classifier head C. The training phase mainly includes three components: triplet estimation, discriminative learning, and adversarial learning. The triplet estimation component provides uncertainty estimates and continuously updated triplets for the next two modules. The discriminative learning and adversarial learning components utilize the estimated triplets for adversarial training to remove confounding factors, thereby improving the identifiability of OOD data. Detailed optimization workflow can be seen in Fig. 4 and Algorithm 1.

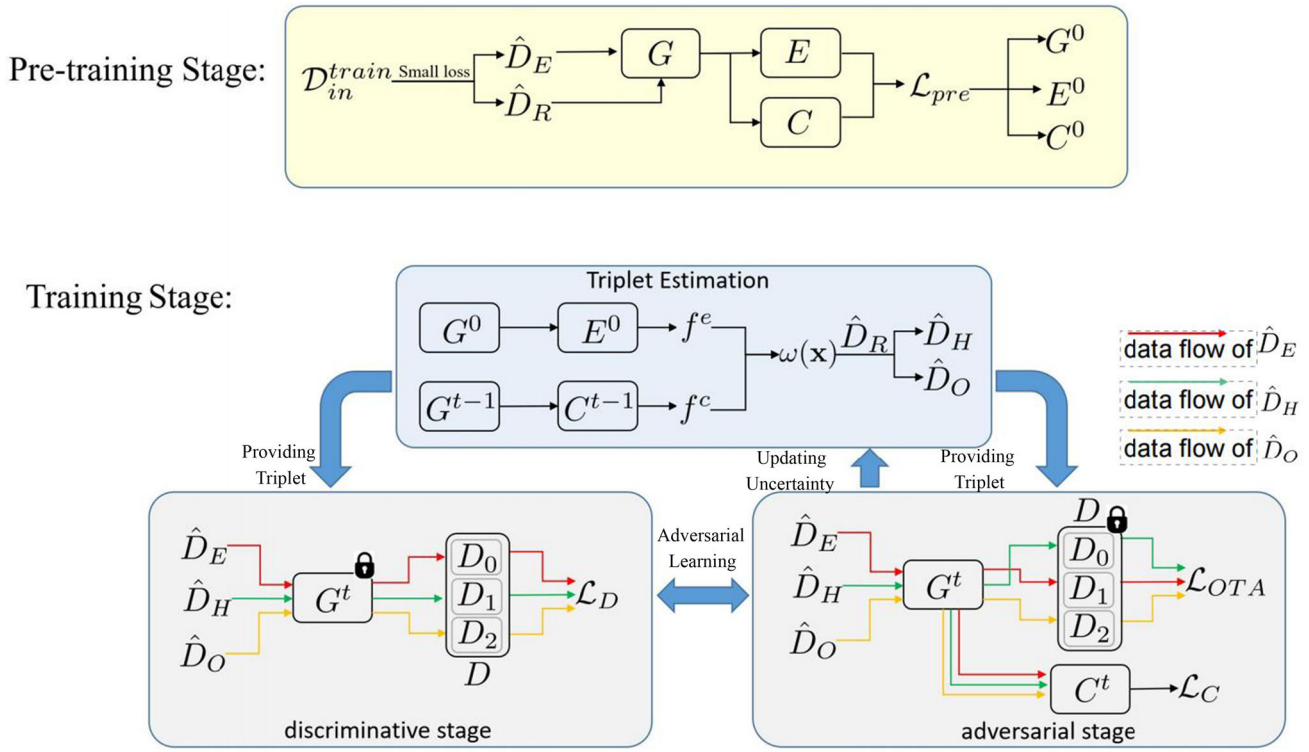

Fig. 4 The overview of ACRE for OOD detection under asymmetric open-set noise

Algorithm 1 Adversarial Confounder REmoving (ACRE).

Require: Noisy training set $\mathcal{D}_{in}^{train}$, harmonization factor τ, loss coefficients $\lambda_1, \lambda_2, \lambda_3$, training epochs n_{iter}
1: **collect** easy-ID examples $\hat{D}_E$ by Eq. (4)
2: **obtain** pre-trained G, E, and C by minimizing $\mathcal{L}_{pre}$ in Eq. (5)
3: **estimate** ID probability $\omega(\mathbf{x})$ in Eq. (6)
4: **collect** hard-ID examples $\hat{D}_H$ and open-set noisy examples $\hat{D}_O$ by $\omega(\mathbf{x})$
5: **for** $i = 1$ **to** n_{iter} **do**
6: **for** $i \in \{1, \ldots, n_{iter}\}$ **do**
7: **update** D by Eq. (9)
8: **update** G and C by Eq. (10) and Eq. (7)
9: **update** ID probability $\omega(\mathbf{x})$ in Eq. (6)
10: **recollect** hard-ID examples $\hat{D}_H$ and open-set noisy examples $\hat{D}_O$ by $\omega(\mathbf{x})$
11: **end for**
12: **compute** K+1-Guided score $S(\mathbf{x})$ by Eq. (13)
13: **distinguish** ID and OOD data by Eq. (14)

3.4.1 Triplet Estimation

In this part, we first select easy-ID examples based on small loss criteria. Then we propose $\omega(\mathbf{x})$ estimation to progressively identify potential hard-ID examples and open-set noisy examples, where $\omega(\mathbf{x})$ denotes the probability that the example $\mathbf{x}$ belongs to the ID classes.

Easy-ID examples selection. We identify the easy-ID examples by the small loss criteria (Han et al., 2018; Jiang et al., 2018):

$$\hat{D}_E = \{(\mathbf{x}_k, \tilde{\mathbf{y}}_k) | (\mathbf{x}_k, \tilde{\mathbf{y}}_k) \in \mathcal{D}_{in}^{train}, \bar{\ell}(\mathbf{x}_k) < \zeta\},$$
$$b\bar{\ell}(\mathbf{x}_k) = -\frac{1}{T_1}\sum_{t=1}^{T_1} \log\left(\frac{e^{E_{\tilde{\mathbf{y}}_k}(G(\mathbf{x}_k))}}{\sum_{j=1}^{K} e^{E_j(G(\mathbf{x}_k))}}\right), \quad (4)$$

where T_1 denotes the number of epoch to select easy-ID data, ζ is the pre-defined threshold, and $E_j(\cdot)$ denotes the j-th logit from classifier E. After obtaining the easy-ID data set $\hat{D}_E$, we use $\hat{D}_R$ to denote the remaining set, which contains both hard-ID examples and open-set noisy examples.

Progressive estimation of $\omega(\mathbf{x})$. The estimation of $\omega(\mathbf{x})$ contains two stages: the pre-training and updating stage. The updating of $\omega(\mathbf{x})$ is progressive and is performed simultaneously with Triplet Adversarial Learning (see Sect. 3.4.2). Details optimization workflow can be seen in Fig. 4.

During the pre-training stage, we utilize all examples in $\mathcal{D}_{in}^{train}$ to pre-train the feature extractor G and two-head classifier. In the optimization process, we utilize $\mathcal{D}_{in}^{train}$ to optimize E and utilize $\hat{D}_E$ to optimize C. The pre-training optimization objective $\mathcal{L}_{pre}$ is defined by

$$\mathcal{L}_{pre} = -\frac{1}{n}\sum_{(\mathbf{x}_k, \tilde{\mathbf{y}}_k) \in \mathcal{D}_{in}^{train}} \log\left(\frac{e^{E_{\tilde{\mathbf{y}}_k}(G(\mathbf{x}_k))}}{\sum_{j=1}^{K} e^{E_j(G(\mathbf{x}_k))}}\right)$$

$$-\frac{1}{n_e}\sum_{(\mathbf{x},y)\in\hat{D}_E}\log\left(\frac{e^{C_y(G(\mathbf{x}))}}{\sum_{j=1}^{K+1}e^{C_j(G(\mathbf{x}))}}\right), \tag{5}$$

where n_e denotes the number of examples in $\hat{D}_E$. After the pre-trained stage by minimizing $\mathcal{L}_{pre}$, we obtain pre-trained G, E, and C. Then, based the pre-trained G, E, and C, we estimate $\omega(\mathbf{x})$ by

$$\omega(\mathbf{x}) = (1-\tau)\cdot f^e(E(G(\mathbf{x})))+\tau\cdot(1-f^c(C(G(\mathbf{x})))), \tag{6}$$

where f^e denotes maximum softmax probability (MSP) (Hendrycks and Gimpel, 2017), f^c denotes the K+1-th softmax probability, and τ is a harmonization factor. Our experiments verify that using two different classifiers can improve the estimation of $\omega(\mathbf{x})$ because, at the initial stage, E trained with all examples outperforms C, E can primarily guide C to constantly self-growth and self-renewal trained with easy-ID examples.

During the updating stage, we begin by preliminarily splitting $\hat{D}_R$ into hard-ID set $\hat{D}_H$ and open-set noisy set $\hat{D}_O$ by $\omega\hat{D}_R$ and $(1-\omega)\hat{D}_R$, respectively. Let n_h and n_o denote the number of examples in $\hat{D}_H$ and $\hat{D}_O$. Then, to improve the estimation of $\omega(\mathbf{x})$, we rectify the label of $\hat{D}_O$ to $K+1$ and fine-tune the two-head classifier with $\hat{D}_H$ and $\hat{D}_O$ by $\mathcal{L}_C$, which is defined by

$$\begin{aligned}\mathcal{L}_C = &-\frac{\lambda_2}{n_h+n_e}\sum_{(\mathbf{x},y)\in\hat{D}_H\cup\hat{D}_E}\log\left(\frac{e^{C_y(G(\mathbf{x}))}}{\sum_{j=1}^{K+1}e^{C_j(G(\mathbf{x}))}}\right)\\ &-\frac{\lambda_3}{n_o}\sum_{\mathbf{x}\in\hat{D}_O}\log\left(\frac{e^{C_{K+1}(G(\mathbf{x}))}}{\sum_{j=1}^{K+1}e^{C_j(G(\mathbf{x}))}}\right),\end{aligned} \tag{7}$$

where λ_2 and λ_3 denote coefficients. The first item in Eq. (7) optimizes the first K outputs of C, and the second item in Eq. (7) optimizes the K+1-th output of C. Then, we update $\omega(\mathbf{x})$ by optimizing the second item in Eq. (6) with fine-tuned G and C. Besides updating $\omega(\mathbf{x})$, another advantage of the K+1-class classifier is that it can be utilized to design an OOD scoring function (see Sect. 3.4.3).

3.4.2 Triplet Adversarial Learning

The objective of this section is to remove the confounding variable through adversarial learning over the estimated triplet, comprising both the discriminative and adversarial stages.

During the discriminative stage, we introduce a triplet discriminator D as follows,

$$D(G(\mathbf{x})) = [D_0(G(\mathbf{x})), D_1(G(\mathbf{x})), D_2(G(\mathbf{x}))], \tag{8}$$

where $D_0(\cdot)$, $D_1(\cdot)$, and $D_2(\cdot)$ represent the probability that example $\mathbf{x}$ belongs to easy-ID subset, hard-ID subset, and open-set subset, respectively. Given these three dimensions outputted by D, we propose the triplet discrimination loss $\mathcal{L}_D$, which is defined by

$$\begin{aligned}\mathcal{L}_D(x) = &\frac{1}{n_e}\sum_{\mathbf{x}\in\hat{D}_E}[-\log D_0(G(\mathbf{x}))]\\ &+\frac{1}{n_h}\sum_{\mathbf{x}\in\hat{D}_H}[-\log D_1(G(\mathbf{x}))]\\ &+\frac{1}{n_o}\sum_{\mathbf{x}\in\hat{D}_O}[-\log D_2(G(\mathbf{x}))].\end{aligned} \tag{9}$$

By fixing G and optimizing D with $\mathcal{L}_D$, we can obtain an optimal discriminator D^*, which can identify which of the three triplets the example comes from.

During the adversarial stage, our approach tackles the removal of the confounding variable by learning invariant representations across easy-ID and hard-ID examples and minimizing the spurious-related representations between hard-ID examples and open-set noisy examples. However, designing an effective adversarial loss remains a critical challenge, as prior adversarial methods typically employ two-dimensional discriminators (Gui et al., 2023; Ganin et al., 2016). In contrast, our method utilizes a three-dimensional discriminator. To address this issue, we propose the OOD-aware triplet adversarial loss $\mathcal{L}_{OTA}$:

$$\begin{aligned}\mathcal{L}_{OTA}(x) = &\frac{1}{n_e}\sum_{\mathbf{x}\in\hat{D}_E}[-\log D_1(G(\mathbf{x}))]\\ &+\frac{1}{n_h}\sum_{\mathbf{x}\in\hat{D}_H}[-\log D_0(G(\mathbf{x}))]\\ &+\frac{1}{n_o}\sum_{\mathbf{x}\in\hat{D}_O}[-\log D_2(G(\mathbf{x}))].\end{aligned} \tag{10}$$

Adversarial optimization. Based on Eq. (9) and Eq. (10), we define the bi-level optimization by

$$\begin{aligned}&\min_G \frac{1}{n_e}\sum_{\mathbf{x}\in\hat{D}_E}[-\log D_1^*(G(\mathbf{x}))]\\ &+\frac{1}{n_h}\sum_{\mathbf{x}\in\hat{D}_H}[-\log D_0^*(G(\mathbf{x}))]\\ &+\frac{1}{n_o}\sum_{\mathbf{x}\in\hat{D}_O}[-\log D_2^*(G(\mathbf{x}))],\\ &s.t. D^* = \arg\min_D \frac{1}{n_e}\sum_{\mathbf{x}\in\hat{D}_E}[-\log D_0(G(\mathbf{x}))]\end{aligned}$$

$$+\frac{1}{n_h}\sum_{\mathbf{x}\in\hat{D}_H}\left[-\log D_1(G(\mathbf{x}))\right]$$
$$+\frac{1}{n_o}\sum_{\mathbf{x}\in\hat{D}_O}\left[-\log D_2(G(\mathbf{x}))\right], \tag{11}$$

The first two items in Eq. (11) achieve the minimization of distribution discrepancies across easy-ID and hard-ID data to learn invariant representations to remove $\mathbf{z}_b$. The last two items achieve the maximization of distribution discrepancies across hard-ID data and open-set noisy data, enabling the dual elimination of spurious-related representations to remove $\mathbf{z}_b$. Triplet adversarial learning increases the distribution discrepancies across hard-ID data and open-set noisy data, which contributes to better estimation of $\omega(\mathbf{x})$ and curates three collections of potential examples more reliably.

Although confounding variable $\mathbf{z}_b$ can be removed by optimizing Eq. (11), the classifier still outputs an overconfident prediction to OOD data. The reason is that, during the training stage, ID example $(\mathbf{x}_i, \mathbf{y}_i)$ and asymmetric open-set example $(\mathbf{x}_o, \tilde{\mathbf{y}}_o)$ are used for optimizing classifier. Classifier is constantly learning spurious influence from $\mathbf{x}_o$ to $\tilde{\mathbf{y}}_o$. Although $\mathbf{x}_o$ is only generated by $\mathbf{z}_c$ after removing $\mathbf{z}_b$, classifier also learns another spurious influence: $\mathbf{z}_c \rightarrow \tilde{\mathbf{y}}_o$. During the inference stage, when the tested OOD example $\mathbf{x}_{ot}$ in D_{out}^{test} contains $\mathbf{z}_c$, $\mathbf{x}_{ot}$ is likely to be identified as ID example, resulting in poor performance of OOD detection. To remove the spurious influence $\mathbf{z}_c \rightarrow \tilde{\mathbf{y}}_o$, we propose to rectify the label of $\mathbf{x}_o$ to $K+1$, and optimate model by connecting Eq. (11) with Eq. (7), which helps learn a correct influence $\mathbf{z}_c \rightarrow K+1$, improve the classifier's separability of ID and OOD data, and enhance the estimation of ID probability $\omega(\mathbf{x})$.

Therefore, considering Eqs. (11) and (7), we define the final optimization formally by

$$\min_{G,C} \lambda_1 \mathcal{L}_{OTA} + \mathcal{L}_C, \quad s.t. \quad D^* = \arg\min_D \mathcal{L}_D\,, \tag{12}$$

where λ_1 are coefficient.

3.4.3 OOD Detection in Testing Stage

During the test stage, we define the K+1-Guided scoring function $S(\mathbf{x})$ from classifier C by

$$S(\mathbf{x}) = \frac{e^{C_{K+1}(G(\mathbf{x}))}}{\sum_{j=1}^{K+1} e^{C_j(G(\mathbf{x}))}}\,, \tag{13}$$

where C_j denotes the j-th logit from C. For OOD detection, one can exercise the thresholding mechanism to distinguish between ID and OOD data by

$$G_\gamma(\mathbf{x}) = \begin{cases} \text{OOD} & S(\mathbf{x}) \geq \gamma\,, \\ \text{ID} & S(\mathbf{x}) < \gamma\,. \end{cases} \tag{14}$$

The threshold γ is chosen so that a high fraction of ID data (e.g., 95%) is correctly classified (Sun et al., 2021b). All the used notations can be seen in Table 2.

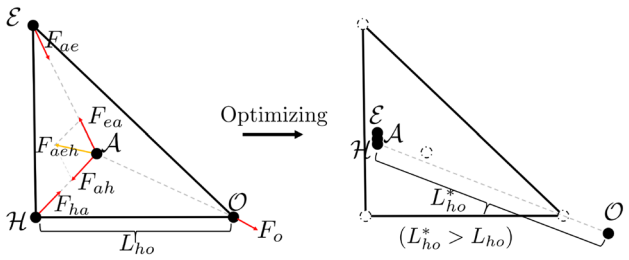

Fig. 5 The force analysis of optimizing $\mathcal{L}_{OTA}$

4 Theoretical Analysis

According to inner-level optimization in Eq. (11), we minimize $\mathcal{L}_D$ loss to find optimal D^* with a fixed G. The output of D^* is $D^*(z) = \left[\frac{P_E(z)}{3P_{avg}(z)}, \frac{P_H(z)}{3P_{avg}(z)}, \frac{P_O(z)}{3P_{avg}(z)}\right]$, where $P_{avg}(z) = (P_E(z) + P_H(z) + P_O(z))/3$, z denotes the output from G. $P_E(z)$, $P_H(z)$, and $P_O(z)$ are the feature distributions of easy-ID data, hard-ID data, and open-set data, respectively. We optimize G given D^* by minimizing $\mathcal{L}_{OTA}$ loss. Then we can obtain the following Theorem.

Theorem 1 *(Proof in Appendix) $\mathcal{L}_{OTA}$ loss can be expressed as,*

$$\begin{aligned}\mathcal{L}_{OTA} = &KL\left(P_H \| P_{avg}\right) + 3KL\left(P_{avg} \| P_H\right) \\ &+ KL\left(P_E \| P_{avg}\right) + 3KL\left(P_{avg} \| P_E\right) \\ &- KL\left(P_O \| P_{avg}\right) + O_{EH} + 5\log 3\,,\end{aligned} \tag{15}$$

where

$$O_{EH} = \int_z \left(P_O(z)\log\frac{P_H(z)}{3P_{avg}(z)} + P_O(z)\log\frac{P_E(z)}{3P_{avg}(z)}\right)dz\,. \tag{16}$$

By minimizing $\mathcal{L}_{OTA}$ loss, the confounding variable can be removed (Fig. 5).

Remark We analyze $KL\left(P_H \| P_{avg}\right) + 3KL\left(P_{avg} \| P_H\right) + KL\left(P_E \| P_{avg}\right) + 3KL\left(P_{avg} \| P_E\right) - KL\left(P_O \| P_{avg}\right)$ by the analysis of forces in the field of physics. Since the KL divergence is asymmetric, it can be viewed as a force approximately. We use F_{ea}, F_{ha}, F_{ah}, F_{ae} to denote $KL\left(P_E \| P_{avg}\right)$,

Table 2 The summary of all the used notations

$\mathcal{X}, \mathcal{Y}, \mathcal{Z}$	Input space, ID label space, latent space
K	The number of ID classes
$\mathcal{D}_{in}^{train}, \mathcal{D}^{test}, \mathcal{D}_{in}^{test}, \mathcal{D}_{out}^{test}$	Noisy training set, test set, ID test set, OOD test set
n	The number of examples in $\mathcal{D}_{in}^{train}$
$(x_k, \tilde{y}_k)$	The example with index of k in $\mathcal{D}_{in}^{train}$
y^*	Ground-truth label
$\tilde{y}_o$	Noisy label of open-set example
x_{it}, x_{ot}	The example in $\mathcal{D}_{in}^{test}$, The example in $\mathcal{D}_{out}^{test}$
$\mathcal{F}$	Decision function for OOD detection
$\mathbf{x}_i, \mathbf{x}_o, \mathbf{x}_e, \mathbf{x}_h$	ID example, OOD example, easy-ID example, hard-ID example
$\mathbf{z}_a$	The private variable which only $\mathbf{x}_i$ owns
$\mathbf{z}_c$	The private variable which only $\mathbf{x}_o$ owns
$\mathbf{z}_b$	The confounding variable which both $\mathbf{x}_i$ and $\mathbf{x}_o$ own simultaneously
E, C, D, G	K-class classifier, K+1-class classifier, discriminator, feature extractor
f^e	The maximum softmax probability of E
f^c	The K+1-th softmax probability of C
τ	The harmonization factor to control two items about certainty estimation
$\hat{D}_E, \hat{D}_H, \hat{D}_O$	Easy-ID data set, hard-ID data set, open-set noisy set
$\hat{D}_R$	The remaining set which contains both hard-ID and open-set noisy examples
ω	Certainty estimation score
ζ	The pre-defined threshold for obtaining $\hat{D}_E$
$\lambda_1, \lambda_2, \lambda_3$	The coefficients about $\mathcal{L}_C$ and $\mathcal{L}_{OTA}$
$S(x)$	The K+1-Guided scoring function

$KL\left(P_H \| P_{avg}\right)$, $KL\left(P_{avg} \| P_H\right)$, $KL\left(P_{avg} \| P_E\right)$, respectively. $\mathcal{E}, \mathcal{H}, \mathcal{O}, \mathcal{A}$ denote P_E, P_H, P_O, P_A, located at the three vertices and the center of the triangle, respectively. F_{aeh} denotes the resultant force, and its direction represents the direction $\mathcal{A}$ moves. F_{ha} and F_{ea} will keep $\mathcal{E}$ and $\mathcal{H}$ moving closer to $\mathcal{A}$. By optimizing $\mathcal{L}_{OTA}$, $KL\left(P_E \| P_{avg}\right)$, $KL\left(P_H \| P_{avg}\right)$, $KL\left(P_{avg} \| P_H\right)$, $\left(P_{avg} \| P_E\right)$ will keep decreasing until $P_E \approx P_H \approx P_{avg}$. We use F_a denote $-KL\left(P_O \| P_{avg}\right)$. Minimizing $\mathcal{L}_{OTA}$ increases $KL\left(P_O \| P_{avg}\right)$, resulting in $\mathcal{O}$ constantly moving away from $\mathcal{A}$. D_{AO} denotes the distance of $\mathcal{A}$ and $\mathcal{O}$ in the optimal G. Moreover, minimizing $\mathcal{L}_{OTA}$ will decrease O_{EH} and the output of the open-set data on D_0 and D_1. $P_E \approx P_H \approx P_{avg}$ dictates that we learn invariant representations across P_E and P_H, and $\mathcal{O}$ constantly moving away from $\mathcal{A}$ dictates that we minimize the spurious-related representations, successfully removing confounding variable.

5 Experiments

To validate the effectiveness of ACRE, we conduct a comprehensive performance evaluation, comparing it against state-of-the-art methods.

5.1 Setup

Following Wu et al. (2021), we choose CIFAR-10 and CIFAR-100 as the ID benchmark datasets, and choose CIFAR-100, TinyImageNet (Deng et al., 2009), and Places365 (Zhou et al., 2017) as the OOD benchmark datasets. Taking the ID dataset CIFAR-10 and OOD dataset CIFAR-100 as an example (abbreviated as CIFAR-10&CIFAR-100), the generation process of the noisy training set $\mathcal{D}_{in}^{train}$ can be described in three steps. Firstly, we train a supervised model on CIFAR-10 using cross-entropy loss. Secondly, a certain percentage of open-set examples from CIFAR-100 are randomly selected, and their pseudo-labels are predicted using the trained supervised model. Lastly, the open-set examples with pseudo-labels are integrated into the training set. This procedure approximates the actual noise generation process. Table 3 shows the dataset configurations for open-set noisy environments.

We compare our proposed **ACRE** with state-of-the-art methods: **MSP** (Hendrycks and Gimpel, 2017), **ODIN** (Liang et al., 2017), **Mahalanobis** (Lee et al., 2018), **Energy** (Liu et al., 2020a), **GradNorm** (Huang et al., 2021), **RankFeat** (Song et al., 2022b), **LogitNorm** (Wei et al., 2022), **NGC** (Wu et al., 2021), and **ODNL** (Wei et al., 2021).

Table 3 Dataset configurations for open-set noisy environments

$\mathcal{D}_{in}^{train}$	$\mathcal{D}_{in}^{test}$	$\mathcal{D}_{out}^{test}$
CIFAR-10 (CIDAR-100)	CIFAR-10	CIFAR-100
CIFAR-10 (TinyImagenet)	CIFAR-10	TinyImagenet
CIFAR-10 (Places365)	CIFAR-10	Places365
CIFAR-100 (TinyImagenet)	CIFAR-100	TinyImagenet
CIFAR-100 (Places365)	CIFAR-100	Places365
TinyImagenet (CIFAR-100)	TinyImagenet	CIFAR-100
ImageNet-100 (ImageNet-100-200)	ImageNet-100	ImageNet-100-200
CIFAR-10 (CIFAR-100)	CIFAR-10	iSUN, Places365, Texture, SVHN, LSUN-C, LSUN-R
CIFAR-100 (TinyImagenet)	CIFAR-100	iSUN, Places365, Texture, SVHN, LSUN-C, LSUN-R

Similar to Liu et al. (2020b), we measure the following metrics for OOD detection: (1) the false positive rate (FPR95) of OOD examples when the true positive rate of ID examples is at 95%; (2) the area under the receiver operating characteristic curve (AUROC); and (3) the area under the precision-recall curve (AUPR).

We implement our algorithm in PyTorch using the standard ResNet-50 backbone. The model is trained for 4,000 iterations with ACRE. Mini-batch SGD is employed with a weight decay of 5e-4 and Nesterov momentum of 0.9. The learning rate is 0.002 and follows a cosine annealing schedule.

5.2 Results

Table 4 presents the results obtained on the CIFAR-10 in-distribution (ID) dataset, along with three out-of-distribution (OOD) datasets: CIFAR-100, TinyImageNet, and Places365. ACRE demonstrates superior performance compared to all test-time OOD detection methods, as shown in Table 4. For instance, when compared to RankFeat, ACRE exhibits an **8.68%** improvement in AUROC, **7.03%** improvement in AUPR, and **27.72%** improvement in FPR95 on the CIFAR-100 OOD dataset. Similarly, in comparison to Energy, ACRE achieves an **15.29%** improvement in AUROC, **9.98%** improvement in AUPR, and **48.99%** improvement in FPR95 on the TinyImageNet OOD dataset, demonstrating the effectiveness of confounding variable removing. These results also demonstrate that test-time OOD detection methods are vulnerable to asymmetric open-set noise. Moreover, the results indicate that CIFAR-100 and TinyImageNet are closer to CIFAR-10 than Places365. Nevertheless, ACRE consistently achieves the best performance across all evaluation metrics, regardless of the proximity of the OOD dataset, thereby confirming its robustness.

Table 5 presents the results obtained on the CIFAR-100 in-distribution (ID) dataset, along with two out-of-distribution (OOD) datasets: TinyImageNet and Places365. ACRE surpasses all training-time OOD detection methods, including LogitNorm, NGC, and ODNL, as indicated in Table 5. For instance, compared to LogitNorm, ACRE achieves an **18.68%** improvement in AUROC, **25.68%** improvement in AUPR, and **41.03%** improvement in FPR95 on the TinyImageNet OOD dataset. NGC and ODNL are designed to address scenarios where the ID training set contains close-set noise and symmetric open-set noise. In comparison, ACRE achieves **25.69%** and **39.27%** improvement in FPR95 on the TinyImageNet OOD dataset, respectively, outperforming NGC and ODNL. Notably, the test-time OOD detection method MSP exhibits a higher AUROC by **0.61%** and **3.32%** compared to NGC and ODNL, respectively, on the Places365 OOD dataset. These results emphasize the limitations of combining noisy label learning and OOD detection techniques in effectively handling OOD detection under asymmetric open-set noise Table 6.

5.3 More Results on Larger-Scale Datasets

To validate the effectiveness and generalizability of ACRE, we also conduct experiments on larger-scale datasets TinyImageNet and ImageNet. For TinyImageNet, we use TinyImageNet as the in-distribution (ID) data and obtain out-of-distribution (OOD) data from CIFAR-100. For ImageNet, we treat the first 100 classes (ImageNet-100) as the ID classes and the classes with index of 100 to 200 (ImageNet-100-200) as OOD classes. The specific experimental results are shown in Table 7. According to Table 7, ACRE comprehensively surpasses all the baseline methods, which demonstrates the effectiveness of our approach in addressing out-of-distribution (OOD) detection under asymmetric open-set noise. ACRE exhibits strong performance on both standard-scale datasets and larger-scale datasets, which also reflects ACRE's good generalization ability.

Table 4 The effectiveness of ACRE on the ID datasets CIFAR-10 and OOD datasets CIFAR-100, TinyImageNet, and Places365.

Method	CIFAR-100			TinyImageNet			Places365		
	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓
MSP (Hendrycks and Gimpel, 2017)	77.67	80.49	83.78	73.45	46.80	86.28	84.28	61.20	70.59
ODIN (Liang et al., 2017)	79.77	81.68	78.35	77.86	51.54	77.88	84.72	57.25	68.07
Energy (Liu et al., 2020a)	79.76	81.67	78.38	77.87	51.54	77.67	84.90	58.06	67.44
Mahalanobis (Lee et al., 2018)	72.79	73.79	77.93	85.12	59.81	57.61	83.09	60.94	77.70
GradNorm (Huang et al., 2021)	70.38	72.65	84.62	65.49	40.66	89.22	48.60	21.94	96.16
ReAct (Sun et al., 2021a)	78.47	80.67	83.69	76.87	50.66	82.17	83.38	54.21	73.61
RankFeat (Song et al., 2022b)	79.78	81.68	78.40	78.00	51.60	77.46	84.31	58.23	67.38
EED (He et al., 2024ba)	83.62	84.42	72.31	84.43	37.61	63.38	82.76	50.71	69.72
MMD (He et al., 2024ba)	81.41	83.85	75.99	82.95	46.77	73.98	85.00	67.67	75.43
LAPS (He et al., 2024b)	82.35	82.18	72.98	78.27	35.87	75.37	82.39	52.73	73.25
LogitNorm (Wei et al., 2022)	75.79	73.92	79.14	71.56	32.72	81.84	71.42	30.79	84.93
NGC (Wu et al., 2021)	84.29	84.50	70.31	83.63	42.45	65.53	86.96	65.15	55.71
ODNL (Wei et al., 2021)	75.07	72.60	80.96	71.42	29.81	80.79	83.90	57.64	67.72
ACRE (ours)	**88.46**	**88.71**	**50.68**	**93.16**	**61.52**	**28.68**	**96.11**	**83.96**	**16.88**

The ratio of asymmetric open-set noise is 50%. All values are percentages. ↑ (↓) indicates larger (smaller) values are better. Bold numbers are superior results

Table 5 OOD detection results of ACRE and comparison with competitive baselines on the ID datasets CIFAR-100

Method	TinyImageNet			Places365		
	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓
MSP (Hendrycks and Gimpel, 2017)	72.46	21.89	84.06	70.50	32.51	87.83
ODIN (Liang et al., 2017)	70.47	17.49	84.38	67.31	26.61	89.28
Energy (Liu et al., 2020a)	70.46	17.50	84.38	67.43	26.73	88.99
Mahalanobis (Lee et al., 2018)	67.63	29.72	93.12	54.91	33.22	98.30
GradNorm (Huang et al., 2021)	50.28	13.13	96.93	50.84	20.64	99.03
ReAct (Sun et al., 2021a)	66.97	14.33	85.36	61.80	21.34	90.99
RankFeat (Song et al., 2022b)	70.55	17.55	84.01	67.86	26.15	88.94
EED (He et al., 2024ba)	71.36	19.41	84.67	68.45	29.15	89.47
MMD (He et al., 2024ba)	71.12	32.16	91.21	67.50	44.98	92.18
LAPS (He et al., 2024b)	70.03	16.89	84.62	67.01	26.12	89.28
LogitNorm (Wei et al., 2022)	69.15	20.69	88.47	65.42	28.97	87.91
NGC (Wu et al., 2021)	75.56	32.24	73.13	69.89	31.13	86.80
ODNL (Wei et al., 2021)	70.41	20.30	86.71	67.18	30.95	89.31
ACRE (ours)	**87.83**	**46.37**	**47.44**	**93.17**	**77.27**	**30.53**

Bold values indicate the superior results
The ratio of asymmetric open-set noise is 20%

Table 6 The effectiveness of ACRE on the ID datasets CIFAR-10 and OOD datasets CIFAR-100, TinyImageNet, and Places365 under $K+1$-Guided score

Method	CIFAR-100			TinyImageNet			Places365		
	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓
$K+1$	79.12	75.86	73.91	77.52	22.42	67.21	79.11	37.59	70.24
$K+1$ + ACRE	88.46	88.71	50.68	93.16	61.52	28.68	96.11	83.96	16.88

Table 7 The effectiveness of ACRE on larger-scale datasets

Method	TinyImageNet&CIFAR-100			ImageNet-100&ImageNet-100-200		
	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓
MSP (Hendrycks and Gimpel, 2017)	64.86	67.98	93.53	72.98	76.43	87.15
ODIN (Liang et al., 2017)	62.07	63.78	94.32	73.39	75.54	86.81
Energy (Liu et al., 2020a)	61.98	63.71	94.53	73.38	75.51	86.80
Mahalanobis (Lee et al., 2018)	54.28	58.96	96.54	71.97	75.88	85.45
GradNorm (Huang et al., 2021)	51.82	51.71	93.67	67.75	67.92	85.26
ReAct (Sun et al., 2021a)	60.99	62.08	94.23	73.13	74.59	86.28
RankFeat (Song et al., 2022b)	61.98	63.71	94.53	73.38	75.51	86.80
EED (He et al., 2024ba)	68.10	70.67	91.35	70.92	88.10	88.23
MMD (He et al., 2024ba)	55.55	60.26	93.72	67.97	86.72	86.92
LAPS (He et al., 2024b)	67.68	69.40	90.28	68.15	85.87	87.72
LogitNorm (Wei et al., 2022)	54.26	57.87	96.40	64.00	67.77	92.05
NGC (Wu et al., 2021)	70.95	68.02	83.74	73.83	73.42	77.85
ODNL (Wei et al., 2021)	63.58	66.16	92.49	42.90	44.67	96.79
ACRE (ours)	**93.28**	**92.26**	**26.71**	**83.43**	**83.19**	**62.23**

All values are percentages. ↑ (↓) indicates larger (smaller) values are better. Bold numbers are superior results

Table 8 The verification of the effectiveness of $\mathcal{L}_{OTA}$ and K+1-Guided score

E	C	AUROC ↑	AUPR ↑	FPR95 ↓
✓	–	94.61	73.43	20.98
–	✓	95.51	80.23	18.84
✓	✓	**96.11**	**83.96**	**16.88**

Bold values indicate the superior results

5.4 Analyses

We analyze the individual strengths of three components: the adversarial loss $\mathcal{L}_{OTA}$, ID probability $\omega(\mathbf{x})$ estimation through the two-head classifier, and the K+1-Guided scoring function.

Effectiveness of $\mathcal{L}_{OTA}$. Table 9 reports the results of with/without adversarial loss $\mathcal{L}_{OTA}$. For instance, when CIFAR-10 is the ID dataset and Places365 is the OOD dataset, the removal of $\mathcal{L}_{OTA}$ leads to a **5.99%** decrease in AUPR and a **24.84%** increase in FPR95. These findings demonstrate the importance of removing the confounding variable and $\mathcal{L}_{OTA}$ effectively achieve that.

Effectiveness of $\omega(\mathbf{x})$ estimation by two-head classifier. Table 8 presents the results of ACRE with different $\omega(\mathbf{x})$ estimation approaches: only using E to estimate $\omega(\mathbf{x})$, only using C to estimate $\omega(\mathbf{x})$, using both E and C to estimate $\omega(\mathbf{x})$. Comparing the results, solely relying on E leads to a decline in OOD detection performance, with approximately **1.50%** decrease in AUROC, **10.53%** decrease in AUPR, and **4.10%** increase in FPR95. Similarly, using only C also results in a decline in OOD detection performance, with **0.60%** decrease in AUROC, **3.73%** decrease in AUPR, and **1.96%** increase in FPR95. These findings validate the effectiveness of $\omega(\mathbf{x})$ estimation by the two-head classifier (Fig. 6).

Effectiveness of the K+1-Guided scoring function. Table 9 reports the results obtained with and without the K+1-Guided scoring function. It reveals that replacing the K+1-Guided scoring function in ACRE with the MSP score leads to a considerable decrease in performance. Specifically, the average AUROC experiences a drop of 9.74%, the average AUPR declines by 16.93%, and the average FPR95 increases by 33.52%. Our findings emphasize the significant advantages of our approach, as it not only successfully eliminates the confounding variable but also enables the training of a K+1-class classifier C, which yields a novel OOD score. Remarkably, this new OOD score proves to be effective for detecting OOD instances under the presence of asymmetric open-set noise.

Effectiveness of ACRE to eliminate confounding factors. To verify that our method can eliminate confounding factors, we fixed the score as K+1-Guided score and then tested the separability between open-set data and ID (In-Distribution) data with and without the use of our ACRE. The results are shown in Table 6. According to Table 6, we can clearly see that after using ACRE, the separability between ID and OOD (Out-Of-Distribution) data significantly increases, indicating that our adversarial training to remove interfering factors is effective. Since confounding factors are key obstacles to the separability between ID and OOD, the improvement in separability after using ACRE also indicates that the confounding factors have been successfully removed by ACRE. To further prove that our method can increase the separability of ID and OOD data by removing

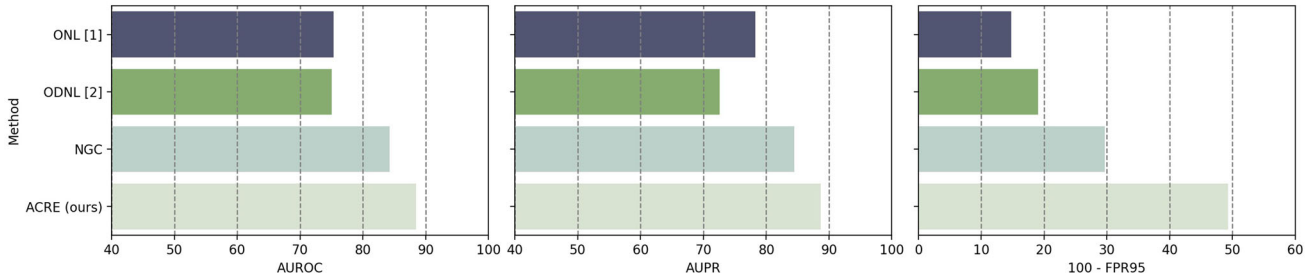

Fig. 6 A comparative analysis with methods specifically addressing open-set noise

Table 9 The verification of the effectiveness of $\mathcal{L}_{OTA}$ and K+1-Guided score

	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓
	w(w/o) $\mathcal{L}_{OTA}$			K+1-Guided score (MSP)		
CIFAR-10&CIFAR-100	**88.46** (87.74)	**88.71** (88.26)	**50.68** (53.12)	**88.46** (84.63)	**88.71** (85.75)	**50.68** (68.05)
CIFAR-10&TinyImageNet	**93.16** (90.94)	**61.52** (54.03)	**28.68** (37.68)	**93.16** (87.78)	**61.52** (52.01)	**28.68** (56.07)
CIFAR-10&Places365	**96.11** (92.01)	**83.96** (77.97)	**16.88** (41.72)	**96.11** (88.55)	**83.96** (70.91)	**16.88** (55.85)
CIFAR-100&TinyImageNet	**87.83** (86.85)	**46.37** (43.20)	**47.44** (49.65)	**87.83** (74.24)	**46.37** (25.24)	**47.44** (81.09)
CIFAR-100&Places365	**93.17** (92.77)	**77.27** (75.45)	**30.53** (32.12)	**93.17** (74.86)	**77.27** (39.31)	**30.53** (80.64)
average	**91.75** (90.06)	**71.57** (67.78)	**34.82** (42.86)	**91.75** (82.01)	**71.57** (54.64)	**34.82** (68.34)

Bold values indicate the superior results

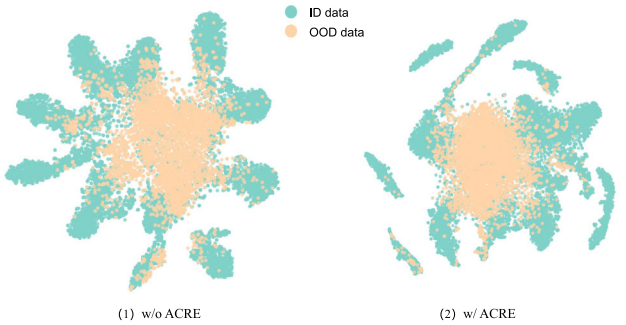

Fig. 7 Comparison of t-SNE plots without and with ACRE under CIFAR-10&Places365

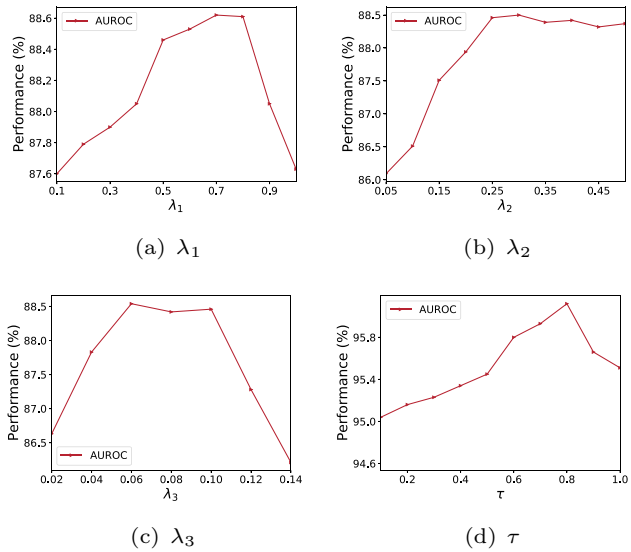

Fig. 8 An analysis of λ_1, λ_2, λ_3, and τ under different values

confounder, we add t-SNE visualization in Fig. 7. According to Fig. 7, ACRE improves the separability between ID and OOD data, while reducing the variance among OOD data.

5.5 Comparative Analysis with Methods Specifically Addressing Open-Set Noise

To further validate the effectiveness of ACRE, we compare ACRE with methods specifically addressing open-set noise. We compare with ONL (Wang et al., 2018), ODNL (Wei et al., 2021), NGC (Wu et al., 2021), and our ACRE. ONL (Wang et al., 2018) detects open-set noise and learns deep discriminative features in an iterative fashion. However, it primarily targets symmetric open-set noise. Symmetric open-set noise does not impair the performance of out-of-distribution detection. In our paper, we investigate asymmetric open-set noise, which is detrimental to out-of-distribution detection. ODNL (Wei et al., 2021) mitigates the problem of label noise by incorporating symmetric open-set noise. Unlike it, we address asymmetric open-set noise and focus on the task of out-of-distribution (OOD) detection. NGC (Wu et al., 2021) investigates the classification problem in scenarios mixed with closed-set noise and symmetric open-set noise. We integrate it with the OOD detection method as one of the baselines for analysis. The experimental results are shown in

Table 10 An analysis of the choice of ζ

Method	ζ=0.1			ζ=0.2			ζ=0.3		
	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓
ACRE (ours)	88.46	88.71	50.68	87.11	86.46	51.44	85.34	83.59	53.67

Table 11 The K+1-Guided score with (without) ACRE

	AUROC ↑	AUPR ↑	FPR95 ↓
(a)			
iSUN	**85.07** (82.10)	**87.93** (81.59)	**70.71** (73.15)
Places365	**84.15** (76.66)	**65.41** (43.39)	**70.52** (82.74)
Texture	**92.57** (87.82)	**95.29** (91.01)	**33.26** (52.29)
SVHN	**83.79** (77.61)	**76.43** (55.90)	**72.96** (86.41)
LSUN-C	**91.90** (74.02)	**93.55** (73.77)	**48.61** (90.80)
LSUN-R	**85.50** (82.46)	**88.03** (81.60)	**75.40** (76.25)
average	**87.16** (80.11)	**84.44** (71.21)	**61.91** (76.94)
(b)			
iSUN	**92.82** (71.31)	**92.02** (73.09)	**24.61** (85.19)
Places365	**74.77** (65.56)	**44.48** (30.92)	**76.98** (89.29)
Texture	**75.84** (67.21)	**83.24** (76.46)	**72.16** (85.25)
SVHN	**74.10** (69.52)	**62.92** (47.54)	92.78 (**90.96**)
LSUN-C	**78.59** (73.49)	**77.13** (69.68)	**66.12** (72.81)
LSUN-R	**95.08** (72.56)	**94.49** (73.37)	**18.16** (84.12)
average	**81.87** (69.94)	**75.71** (61.84)	**58.47** (84.60)

(a): CIFAR-10&CIFAR100; (b): CIFAR-100&TinyImageNet
Bold values indicate the superior results

Fig. 6. According to Fig. 6, our method has a clear advantage compared to baseline methods involving open-set noise. Conventional solutions for symmetric open-set noise are not suitable for direct application in asymmetric open-set noise scenarios, which further validates the high value and significance of researching OOD detection in the context of asymmetric open-set noise.

5.6 Sensitivity of Hyperparameters

Our method has four most important hyperparameters, including τ, λ_1, λ_2 and λ_3. Then, we analyze the sensitivity of them in detail. The sensitivity experiments about λ_1, λ_2 and λ_3 are based on ID dataset CIFAR-10 with OOD dataset CIFAR-100. The sensitivity experiment about τ is based on the ID dataset CIFAR-10 with the OOD dataset Places365.

The sensitivity of τ. τ is the harmonic factor in Eq. (6), which balances the outputs by E and C. Results can be seen in Fig. 8d. Figure 8d shows that choosing a proper value of τ is critical for the performance of OOD detection to a certain extent.

The sensitivity of λ_1, λ_2, **and** λ_3. λ_1 in Eq. (12) acts to balance the OOD-aware triplet adversarial loss $\mathcal{L}_{OTA}$ which is aimed to remove the confounding variable. λ_2 and λ_3 in Eq. (7) act to balance the optimization of the first K outputs of C and the $K+1$-th output of C. The impact of selecting appropriate values for λ_1, λ_2, and λ_3 are crucial for achieving optimal performance in OOD detection, as highlighted in Fig. 8a, b, and c, respectively. Specifically, when the ID dataset is CIFAR-10 and the OOD dataset is CIFAR-100, varying values of λ_1, λ_2, and λ_3 lead to significant differences in the results, with a **1.02%** difference of AUROC in λ_1, a **2.41%** difference of AUROC in λ_2, and a **2.33%** difference of AUROC in λ_3 observed. These findings underscore the importance of carefully choosing the value of λ_1, λ_2, and λ_3 when training the model by our proposed ACRE.

The sensitivity of ζ. ζ is a hype-parameter to select easy ID samples. Table 10 reflects the sensitivity of our method to ζ. According to Table 10, it is clear that our method is quite sensitive to the choice of ζ. When ζ changes from 0.1 to 0.3, the AUROC decreases by 3.12%. A larger ζ means a larger selected easy-ID data set, but it introduces more potential open-set noise. Therefore, we should choose a smaller ζ value to ensure that the selected easy-ID data set contains less open-set noise.

5.7 Generalization of ACRE

Generalization to unseen OOD datasets. In the previous experimental setup, $\mathbf{x}_o$ in $\mathcal{D}_{in}^{train}$ and $\mathbf{x}_{ot}$ in $\mathcal{D}_{out}^{test}$ are from the identical distribution, which follows the setting of Wu et al. (2021), Yu and Aizawa (2020), Zhou et al. (2021). To demonstrate the generalization capability of ACRE, we evaluate its performance on unseen OOD datasets, that is $\mathbf{x}_o$ and $\mathbf{x}_{ot}$ are not from the identical distribution. In our experiments, during the training phase, we utilize CIFAR-10&CIFAR-100 and CIFAR-100&TinyImageNet. However, during the inference phase, we evaluate the model on six unseen (new) OOD datasets, including iSUN, Places365, Texture, SVHN, LSUN-C, and LSUN-R. The results are presented in Table 11. These results highlight that removing the confounding variable effectively mitigates the detrimental effects of asymmetric open-set noise and enhances its capability to detect unseen OOD data. For instance, under CIFAR-100&TinyImageNet, ACRE achieves an average AUROC gain of 11.93%, an average AUPR gain of 13.87%, and an average FPR95 gain of 26.13%. These compelling results substantiate the generalization ability of ACRE.

Generalization to different noisy ratios. To assess the generalization capability of ACRE, we conduct experiments on CIFAR-10&CIFAR-100 with varying ratios of asym-

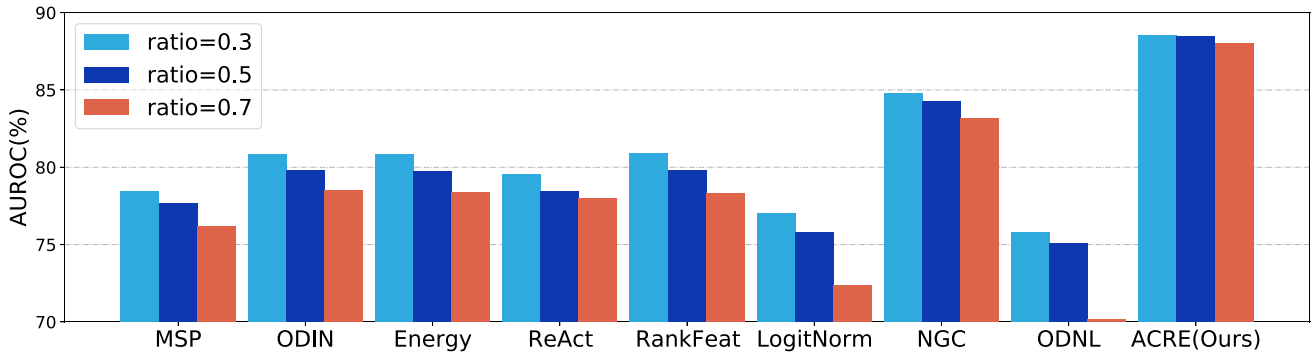

Fig. 9 AUROC(%) of ACRE and compared baselines with different noisy ratios

metric open-set noise: 30%, 50%, and 70%. The results are depicted in Fig. 9. The findings can be summarized as follows. Firstly, as the ratio of asymmetric open-set noise increases, the performance of OOD detection declines across all methods, confirming the adverse impact of asymmetric open-set noise on OOD detection. Secondly, ACRE consistently outperforms all methods by a significant margin, affirming the effectiveness of removing the confounding variable as a strategy to mitigate the negative effects of asymmetric open-set noise on OOD detection. Lastly, ACRE demonstrates superior stability compared to the baselines, exhibiting reduced susceptibility to the influence of asymmetric open-set noise. These observations validate the robustness and generalization capability of ACRE.

6 Conclusion

In this paper, we investigated a previously overlooked problem in detecting OOD examples under the presence of asymmetric open-set noise. Despite its broad applications in the real world, this problem presents significant challenges. To address this problem, we proposed Adversarial Confounder REmoving (ACRE) that introduces triplet estimation and triplet adversarial learning to remove the confounding variable between open-set noisy examples and hard-ID examples. Our method is substantiated by rigorous theoretical analysis and compelling empirical results, highlighting its feasibility and effectiveness. We believe that ACRE establishes a solid foundation for future studies.

The primary limitation of the proposed method lies in its ability to handle solely asymmetric open-set noise, where the ground-truth label exists outside the ID label space, while the assigned label tends to align with the known class that exhibits similar spurious-related characteristics. In future research, we aim to tackle the challenge of OOD detection under multiple noisy environments.

Appendix A: The Proof of Theorem 1

Proof First, we fix the feature extractor G, and minimize the distribution discrimination loss $\mathcal{L}_D$.

$$\begin{aligned}\min_D \mathcal{L}_D(x) =& \mathbb{E}_{P_E(x)}[-\log D_0(G(x))] \\ &+ \mathbb{E}_{P_H(x)}[-\log D_1(G(x))] \\ &+ \mathbb{E}_{P_O(x)}[-\log D_2(G(x))] \\ =& -\int_{x\sim P_E(x)} \log D_0(G(x))dx \\ &- \int_{x\sim P_H(x)} \log D_1(G(x))dx \\ &- \int_{x\sim P_O(x)} \log D_2(G(x))dx \\ =& -\int_{z\sim P_E(z)} \log D_0(z)dz \\ &- \int_{z\sim P_H(z)} \log D_1(z)dz \\ &- \int_{z\sim P_O(z)} \log D_2(z)dz \\ =& \int_z (-P_E(z)\log D_0(z) - P_H(z)\log D_1(z) \\ &- P_O(z)\log D_2(z))\,dz \end{aligned} \tag{A1}$$

$D_0(z)+D_1(z)+D_2(z) = 1$ for all z. Therefore, we transform the above optimization problem into an optimization problem with constraints as follows:

$$\begin{aligned}\min_D & -P_E(z)\log D_0(z) - P_H(z)\log D_1(z) \\ & - P_O(z)\log D_2(z) \\ s.t.\quad & D_0(z) + D_1(z) + D_2(z) = 1 \end{aligned} \tag{A2}$$

To solve the optimization problem with constraints, we use the Lagrange multiplier method.

$$\begin{aligned}\min_D \tilde{\mathcal{L}}_D := & -P_E(z)\log D_0(z) - P_H(z)\log D_1(z) \\ & - P_O(z)\log D_2(z) \\ & + v(D_0(z)+D_1(z)+D_2(z)-1) \end{aligned} \tag{A3}$$

where v denotes the Lagrange variable.

We compute the derivative of $\tilde{\mathcal{L}}_D$ with respect to D and v as follows:

$$\begin{aligned}&\frac{\partial \tilde{\mathcal{L}}_D}{\partial D_0(z)} = \frac{-P_E(z)}{D_0(z)} + v = 0 \quad \Leftrightarrow \quad D_0(z) = \frac{P_E(z)}{v} \\ &\frac{\partial \tilde{\mathcal{L}}_D}{\partial D_1(z)} = \frac{-P_H(z)}{D_1(z)} + v = 0 \quad \Leftrightarrow \quad D_1(z) = \frac{P_H(z)}{v} \\ &\frac{\partial \tilde{\mathcal{L}}_D}{\partial D_2(z)} = \frac{-P_O(z)}{D_2(z)} + v = 0 \quad \Leftrightarrow \quad D_2(z) = \frac{P_O(z)}{v} \\ &\frac{\partial \tilde{\mathcal{L}}_D}{\partial v} = D_0(z) + D_1(z) + D_2(z) - 1 = 0 \\ &\quad \Leftrightarrow \quad D_0(z) + D_1(z) + D_2(z) = 1 \end{aligned} \tag{A4}$$

According to the above equations, we can know

$$D_0(z) + D_1(z) + D_2(z) = \frac{P_E(z)}{v} + \frac{P_H(z)}{v} + \frac{P_O(z)}{v} = 1\,, \tag{A5}$$

where

$$v = P_E(z) + P_H(z) + P_O(z) = 3P_{avg}\,. \tag{A6}$$

Thus, we obtain optimal D^* as

$$\begin{aligned} D^*(z) &= [D_0^*(z), D_1^*(z), D_2^*(z)] \\ &= \left[\frac{P_E(z)}{3P_{avg}(z)}, \frac{P_H(z)}{3P_{avg}(z)}, \frac{P_O(z)}{3P_{avg}(z)}\right]. \end{aligned} \tag{A7}$$

Then, during optimizing G through minimizing $\mathcal{L}_{OTA}$, we fix D with D^*.

$$\begin{aligned}\min_G \mathcal{L}_{OTA}(x) &= \mathbb{E}_{P_E(x)}[-\log D_1^*(G(x))] \\ &\quad + \mathbb{E}_{P_H(x)}[-\log D_0^*(G(x))] + \mathbb{E}_{P_O(x)}[-\log D_2^*(G(x))] \\ &= \int_z \big(-P_E(z)\log D_1^*(z) - P_H(z)\log D_0^*(z) \\ &\quad -P_O(z)\log D_2^*(z)\big)\,dz \\ &= \int_z \Bigg(-P_E(z)\log\frac{P_H(z)}{3P_{avg}(z)} - P_H(z)\log\frac{P_E(z)}{3P_{avg}(z)} \\ &\quad -P_O(z)\log\frac{P_O(z)}{3P_{avg}(z)}\Bigg)\,dz \\ &= \int_z \Bigg(-P_E(z)\log\frac{P_H(z)}{3P_{avg}(z)} - P_H(z)\log\frac{P_E(z)}{3P_{avg}(z)} \\ &\quad -P_O(z)\log\frac{P_O(z)}{3P_{avg}(z)}\Bigg)\,dz \\ &= \int_z \Bigg((P_H(z)+P_O(z)-3P_{avg})\log\frac{P_H(z)}{3P_{avg}(z)} \\ &\quad +(P_E(z)+P_O(z)-3P_{avg})\log\frac{P_E(z)}{3P_{avg}(z)} \\ &\quad -P_O(z)\log\frac{P_O(z)}{3P_{avg}(z)}\Bigg)\,dz \\ &= \int_z \Bigg(P_H(z)\log\frac{P_H(z)}{3P_{avg}(z)} + P_O(z)\log\frac{P_H(z)}{3P_{avg}(z)} \\ &\quad -3P_{avg}\log\frac{P_H(z)}{3P_{avg}(z)} + P_E(z)\log\frac{P_E(z)}{3P_{avg}(z)} \\ &\quad +P_O(z)\log\frac{P_E(z)}{3P_{avg}(z)} - 3P_{avg}\log\frac{P_E(z)}{3P_{avg}(z)} \\ &\quad -P_O(z)\log\frac{P_O(z)}{3P_{avg}(z)}\Bigg)\,dz \\ &= KL\left(P_H\|3P_{avg}\right) + KL\left(3P_{avg}\|P_H\right) + KL\left(P_E\|3P_{avg}\right) \\ &\quad + KL\left(3P_{avg}\|P_E\right) - KL\left(P_O\|3P_{avg}\right) \\ &\quad + \int_z \left(P_O(z)\log\frac{P_H(z)}{3P_{avg}(z)} + P_O(z)\log\frac{P_E(z)}{3P_{avg}(z)}\right)dz \\ &= KL\left(P_H\|P_{avg}\right) + 3KL\left(P_{avg}\|P_H\right) + KL\left(P_E\|P_{avg}\right) \\ &\quad + 3KL\left(P_{avg}\|P_E\right) - KL\left(P_O\|P_{avg}\right) + 5\log 3 \\ &\quad + \int_z \left(P_O(z)\log\frac{P_H(z)}{3P_{avg}(z)} + P_O(z)\log\frac{P_E(z)}{3P_{avg}(z)}\right)dz \\ &= KL\left(P_H\|P_{avg}\right) + 3KL\left(P_{avg}\|P_H\right) + KL\left(P_E\|P_{avg}\right) \\ &\quad + 3KL\left(P_{avg}\|P_E\right) - KL\left(P_O\|P_{avg}\right) \\ &\quad + O_{EH} + 5\log 3\,, \end{aligned} \tag{A8}$$

where O_{EH} denotes $\int_z \left(P_O(z)\log\frac{P_H(z)}{3P_{avg}(z)} + P_O(z)\log\frac{P_E(z)}{3P_{avg}(z)}\right)dz$ for convenience. Then, we analyze $KL\left(P_H\|P_{avg}\right) + 3KL\left(P_{avg}\|P_H\right) + KL\left(P_E\|P_{avg}\right) + 3KL\left(P_{avg}\|P_E\right) - KL\left(P_O\|P_{avg}\right)$ by the analysis of forces in the field of physics. Since the KL dispersion is asymmetric, it can be viewed as a force approximately. As shown in Fig. 5, we use F_{ea}, F_{ha}, F_{ah}, F_{ae} to denote $KL\left(P_E\|P_{avg}\right)$, $KL\left(P_H\|P_{avg}\right)$, $KL\left(P_{avg}\|P_H\right)$, $\left(P_{avg}\|P_E\right)$, respectively. $\mathcal{E}$, $\mathcal{H}$, $\mathcal{O}$, $\mathcal{A}$ denote P_E, P_H, P_O, P_A, located at the three vertices and the center of the triangle, respectively. F_{aeh} denotes the resultant force, and its direction represents the direction $\mathcal{A}$ moves. F_{ha} and F_{ea} will keep $\mathcal{E}$ and $\mathcal{H}$ moving closer to $\mathcal{A}$. By optimizing $\mathcal{L}_{OTA}$, $KL\left(P_E\|P_{avg}\right)$, $KL\left(P_H\|P_{avg}\right)$, $KL\left(P_{avg}\|P_H\right)$, $\left(P_{avg}\|P_E\right)$ will keep decreasing until $P_E \approx P_H \approx P_{avg}$. We use F_a denote $-KL\left(P_O\|P_{avg}\right)$. Minimizing $\mathcal{L}_{OTA}$ increases $KL\left(P_O\|P_{avg}\right)$, resulting in $\mathcal{O}$ constantly moving away from $\mathcal{A}$. D_{AO} denotes the distance of $\mathcal{A}$ and $\mathcal{O}$ in the optimal G. Moreover, minimizing $\mathcal{L}_{OTA}$ will decrease O_{EH} and the output of the open-set data on D_0 and D_1, contribut-

ing to enhancing the separability of ID and OOD distribution as well. □

Author Contributions Conceptualization: H-RD, H-ZY; Methodology: H-RD, H-ZY; Theoretical analysis: H-RD; Writing-original draft preparation: H-RD, H-ZY; Writing-review and editing: H-RD, H-ZY, N-XS, Y-YL, C-XJ; Funding acquisition: Y-YL.

Funding This work is supported by the National Natural Science Foundation of China (62176139, 62176141), the Major Basic Research Project of the Natural Science Foundation of Shandong Province (ZR2021ZD15), the Shandong Provincial Natural Science Foundation for Distinguished Young Scholars (ZR2021JQ26), and the Taishan Scholar Project of Shandong Province (tsqn202103088).

Data Availability Not applicable.

Code Availability Not applicable.

Declarations

Conflict of interest The author declares that he has no confict of interest.

Ethics Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

References

Chen, J., Li, Y., Wu, X., et al. (2021). Atom: Robustifying out-of-distribution detection using outlier mining. In *ECML*, pp. 430–445.

Deng, J., Dong, W., Socher, R., et al. (2009). Imagenet: A large-scale hierarchical image database. In *CVPR, IEEE*, pp. 248–255.

Du, X., Wang, Z., Cai, M., et al. (2022). Vos: Learning what you don't know by virtual outlier synthesis. In *ICLR*.

Du, X., Sun, Y., Zhu, X., et al. (2023). Dream the impossible: Outlier imagination with diffusion models. In *Advances in Neural Information Processing Systems*

Fang, Z., Li, Y., Lu, J., et al. (2022). Is out-of-distribution detection learnable? In *NeurIPS*.

Ganin, Y., Ustinova, E., Ajakan, H., et al. (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research, 17*, 59:1-59:35.

Goldberger, J., & Ben-Reuven, E. (2017). Training deep neural-networks using a noise adaptation layer. In *5th international conference on learning representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Gomes, E. D. C., Alberge, F., Duhamel, P., et al. (2022). Igeood: An information geometry approach to out-of-distribution detection. In *ICLR*.

Gui, J., Sun, Z., Wen, Y., et al. (2023). A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Transactions on Knowledge and Data Engineering, 35*(4), 3313–3332.

Han, B., Yao, Q., Yu, X., et al. (2018). Co-teaching: Robust training of deep neural networks with extremely noisy labels. NeurIPS 31.

Han, Z., Gui, X. J., Sun, H., et al. (2022a). Towards accurate and robust domain adaptation under multiple noisy environments. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Han, Z., Sun, H., & Yin, Y. (2022). Learning transferable parameters for unsupervised domain adaptation. *IEEE Transactions on Image Processing, 31*, 6424–6439.

He, R., Han, Z., Lu, X., et al. (2022a). Ronf: Reliable outlier synthesis under noisy feature space for out-of-distribution detection. In *ACM MM*, pp. 4242–4251.

He, R., Han, Z., Lu, X., et al. (2022b). Safe-student for safe deep semi-supervised learning with unseen-class unlabeled data. In *CVPR*, pp. 14585–14594.

He, R., Han, Z., Lu, X., et al. (2024). SAFER-STUDENT for safe deep semi-supervised learning with unseen-class unlabeled data. *IEEE Transactions on Knowledge and Data Engineering, 36*(1), 318–334. https://doi.org/10.1109/TKDE.2023.3279139

He, R., Yuan, Y., Han, Z., et al. (2024b). Exploring channel-aware typical features for out-of-distribution detection. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 12402–12410.

Hell, F., Hinz, G., Liu, F., et al. (2021). Monitoring perception reliability in autonomous driving: Distributional shift detection for estimating the impact of input data on prediction accuracy. In *Computer science in cars symposium*, pp 1–9.

Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*.

Hendrycks, D., Mazeika, M., & Dietterich, T. (2018). Deep anomaly detection with outlier exposure. In *ICLR*.

Huang, R., Geng, A., & Li, Y. (2021). On the importance of gradients for detecting distributional shifts in the wild. *NeurIPS, 34*, 677–689.

Jang, J., Na, B., Shin, D., et al. (2022). Unknown-aware domain adversarial learning for open-set domain adaptation. In *NeurIPS*.

Jiang, D., Sun, S., & Yu, Y. (2021). Revisiting flow generative models for out-of-distribution detection. In *ICLR*.

Jiang, L., Zhou, Z., Leung, T., et al. (2018). Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*, pp. 2304–2313.

Katz-Samuels, J., Nakhleh, J. B., Nowak, R., et al. (2022). Training ood detectors in their natural habitats. In *International conference on machine learning*, PMLR, pp. 10848–10865.

Lee, K., Lee, K., Lee, H., et al. (2018). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. NeurIPS 31.

Li, J., Xiong, C., & Hoi, S. C. (2020). Mopro: Webly supervised learning with momentum prototypes. arXiv preprint arXiv:2009.07995.

Li, J., Xiong, C., & Hoi, S. C. (2021). Learning from noisy data with robust representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9485–9494.

Liang, S., Li, Y., & Srikant, R. (2017). Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*.

Lin, Z., Roy, S. D., & Li, Y. (2021). Mood: Multi-level out-of-distribution detection. In *CVPR*, pp. 15313–15323.

Liu, W., Wang, X., Owens, J., et al. (2020). Energy-based out-of-distribution detection. *NeurIPS, 33*, 21464–21475.

Liu, W., Wang, X., Owens, J. D., et al. (2020b). Energy-based out-of-distribution detection. In *NeurIPS*.

Ming, Y., & Li, Y. (2023). How does fine-tuning impact out-of-distribution detection for vision-language models? International Journal of Computer Vision.

Ming. Y., Fan. Y., & Li. Y. (2022). Poem: Out-of-distribution detection with posterior sampling. In *ICML*, pp 15650–15665.

Ming, Y., Sun, Y., Dia, O., et al. (2023). How to exploit hyperspherical embeddings for out-of-distribution detection? In *Proceedings of the international conference on learning representations*.

Morningstar, W., Ham, C., Gallagher, A., et al. (2021). Density of states estimation for out of distribution detection. In *AISTATS*, pp 3232–3240.

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, pp. 427–436.

Nguyen, A. T., Tran, T., Gal, Y., et al. (2021). Domain invariant representation learning with domain density transformations. *NeurIPS, 34*, 5264–5275.

Patrini, G., Rozza, A., Krishna Menon, A., et al. (2017). Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1944–1952.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Reed, S., Lee, H., Anguelov, D., et al. (2014). Training deep neural networks on noisy labels with bootstrapping. arXiv preprint arXiv:1412.6596.

Ren, J., Liu, P. J., Fertig, E., et al. (2019). Likelihood ratios for out-of-distribution detection. NeurIPS 32.

Sachdeva, R., Cordeiro, F. R., Belagiannis, V., et al. (2021). Evidentialmix: Learning with combined open-set and closed-set noisy labels. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3607–3615.

Song, Y., Sebe, N., & Wang, W. (2022a). Rankfeat: Rank-1 feature removal for out-of-distribution detection. arXiv preprint arXiv:2209.08590.

Song, Y., Sebe, N., & Wang, W. (2022b). Rankfeat: Rank-1 feature removal for out-of-distribution detection. In *NeurIPS*.

Sun, Y., Guo, C., & Li, Y. (2021a). React: Out-of-distribution detection with rectified activations. In *NeurIPS*, pp. 144–157.

Sun, Y., Guo, C., & Li, Y. (2021b). React: Out-of-distribution detection with rectified activations. In *NeurIPS*.

Sun, Y., Ming, Y., Zhu, X., et al. (2022a) Out-of-distribution detection with deep nearest neighbors. In *ICML*.

Sun, Z., Hua, X. S., Yao, Y., et al. (2020). Crssc: salvage reusable samples from noisy data for robust learning. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 92–101.

Sun, Z., Shen, F., Huang, D., et al. (2022b). Pnp: Robust learning from noisy labels by probabilistic noise prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5311–5320.

Tack, J., Mo, S., Jeong, J., et al. (2020). Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in Neural Information Processing Systems, 33*, 11839–11852.

Tang, K., Miao, D., Peng, W., et al. (2021). Codes: Chamfer out-of-distribution examples against overconfidence issue. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1153–1162.

Wan, W., Wang, X., Xie, M. K., et al. (2024). Unlocking the power of open set: A new perspective for open-set noisy label learning. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 15438–15446.

Wang, F., Han, Z., Gong, Y., et al. (2022a). Exploring domain-invariant parameters for source free domain adaptation. In *CVPR*, pp. 7151–7160.

Wang, H., Li, Z., Feng, L., et al. (2022b). Vim: Out-of-distribution with virtual-logit matching. In *IEEE/CVF conference on computer vision and pattern recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, pp. 4911–4920.

Wang, Q., Fang, Z., Zhang, Y., et al. (2023). Learning to augment distributions for out-of-distribution detection. In *Advances in Neural Information Processing Systems*.

Wang, Y., Liu, W., Ma, X., et al. (2018). Iterative learning with open-set noisy labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8688–8696.

Wei, H., Feng, L., Chen, X., et al. (2020). Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13726–13735.

Wei, H., Tao, L., Xie, R., et al. (2021). Open-set label noise can improve robustness against inherent label noise. *Advances in Neural Information Processing Systems, 34*, 7978–7992.

Wei, H., Xie, R., Cheng, H., et al. (2022). Mitigating neural network overconfidence with logit normalization. In *ICML*.

Wu, Z. F., Wei, T., Jiang, J., et al. (2021). Ngc: A unified framework for learning with open-world noisy data. In *ICCV*, pp. 62–71.

Xia, X., Han, B., Wang, N., et al. (2022). Extended t: Learning with mixed closed-set and open-set noisy labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xiao, Z., Yan, Q., & Amit, Y. (2020). Likelihood regret: An out-of-distribution detection score for variational auto-encoder. *Advances in Neural Information Processing Systems, 33*, 20685–20696.

Yang, J., Wang, H., Feng, L., et al. (2021a). Semantically coherent out-of-distribution detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8301–8309.

Yang, J., Zhou, K., Li, Y., et al. (2021b). Generalized out-of-distribution detection: A survey. arXiv preprint arXiv:2110.11334.

Yang, J., Zhou, K., & Liu, Z. (2023). Full-spectrum out-of-distribution detection. *International Journal of Computer Vision, 131*(10), 2607–2622. https://doi.org/10.1007/S11263-023-01811-Z

Yao, Y., Sun, Z., Zhang, C., et al. (2021). Jo-src: A contrastive approach for combating noisy labels. In *CVPR*, pp. 5192–5201.

Yao, Y., Gong, M., Du, Y., et al. (2023). Which is better for learning with noisy labels: the semi-supervised method or modeling label noise? In *International conference on machine learning*, PMLR, pp. 39660–39673.

Yu, Q., & Aizawa, K. (2019). Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9518–9526.

Yu, Q., & Aizawa, K. (2020). Unknown class label cleaning for learning with open-set noisy labels. In *ICIP*, pp. 1731–1735.

Zhang, L., Goldstein, M., & Ranganath, R. (2021). Understanding failures in out-of-distribution detection with deep generative models. In *ICML*, pp. 12427–12436.

Zhang, Z., & Sabuncu, M. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems* 31.

Zhou, A., & Levine, S. (2021). Amortized conditional normalized maximum likelihood: Reliable out of distribution uncertainty estimation. In *ICML*, pp. 12803–12812.

Zhou, B., Lapedriza, A., Khosla, A., et al. (2017). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(6), 1452–1464.

Zhou, Z., Guo, L. Z., Cheng, Z., et al. (2021). Step: Out-of-distribution detection in the presence of limited in-distribution labeled data. *Advances in Neural Information Processing Systems, 34*, 29168–29180.

Zhu, Y., Chen, Y., Xie, C., et al. (2022). Boosting out-of-distribution detection with typical features. In *NeurIPS*.