Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

LoD: Loss-difference OOD Detection by Intentionally Label-Noisifying Unlabeled Wild Data

Chuanxing Geng[1,2,3], **Qifei Li**[1], **Xinrui Wang**[1], **Dong Liang**[1,3], **Songcan Chen**[1,3] and **Pong C. Yuen**[2*]

[1]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics
[2]Department of Computer Science, Hong Kong Baptist University
[3]MIIT Key Laboratory of Pattern Analysis and Machine Intelligence
{gengchuanxing, liqifei, wangxinrui, liangdong, s.chen}@nuaa.edu.cn, pcyuen@comp.hkbu.edu.hk

Abstract

Using unlabeled wild data containing both in-distribution (ID) and out-of-distribution (OOD) data to improve the safety and reliability of models has recently received increasing attention. Existing methods either design customized losses for labeled ID and unlabeled wild data then perform joint optimization, or first filter out OOD data from the latter then learn an OOD detector. While achieving varying degrees of success, two potential issues remain: (i) Labeled ID data typically dominates the learning of models, inevitably making models tend to fit OOD data as IDs; (ii) The selection of thresholds for identifying OOD data in unlabeled wild data usually faces dilemma due to the unavailability of pure OOD samples. To address these issues, we propose a novel loss-difference OOD detection framework (LoD) by *intentionally label-noisifying* unlabeled wild data. Such operations not only enable labeled ID data and OOD data in unlabeled wild data to jointly dominate the models' learning but also ensure the distinguishability of the losses between ID and OOD samples in unlabeled wild data, allowing the classic clustering technique (e.g., K-means) to filter these OOD samples without requiring thresholds any longer. We also provide theoretical foundation for LoD's viability, and extensive experiments verify its superiority.

1 Introduction

The safety and reliability of traditional machine learning models often face challenges when deployed in real-world environments due to unexpected occurrence of out-of-distribution (OOD) data [Nguyen *et al.*, 2015]. To meet this challenge, the OOD detection problem has been studied [Hendrycks and Gimpel, 2016; Yang *et al.*, 2024], which requires the models not only predict the true class of in-distribution (ID) data but also effectively reject the OOD data. To date, numerous OOD detection methods have been developed [Liu *et al.*, 2020b; Abati *et al.*, 2019; Wang *et al.*, 2022; Hendrycks *et al.*, 2018; Katz-Samuels *et al.*, 2022], and among them, the methods leveraging unlabeled wild data containing ID and OOD samples to improve the performance of OOD detection has recently received increasing attention [Katz-Samuels *et al.*, 2022]. This mainly attributed to the fact that such data can be freely collected during the deployment of any machine learning model in its operational environment, while also allowing for the capture of the true test-time OOD distribution.

Despite the promise, harnessing the power of unlabeled wild data is non-trivial due to the heterogeneous mixture of ID and OOD samples. Existing methods either adopt a joint optimization strategy [Katz-Samuels *et al.*, 2022] or a two-step strategy (i.e., filtering and learning) [Du *et al.*, 2024]. The former aims to design customized losses for labeled ID and unlabeled wild data to jointly optimize the models in a semi-supervised learning manner. The latter first filters out OOD samples from the unlabeled wild data using customized OOD score (usually based on labeled ID data), then uses them along with labeled ID data to learn an OOD detector. While achieving varying degrees of success, two potential issues of these methods remain:

- ✓ **The model-bias issue.** Labeled ID data typically dominates the model learning in both two strategies, especially for the two-step strategy, thus inevitably making the model tend to fit OOD data as IDs.
- ✓ **Threshold selection dilemma.** The selection of thresholds for determining OOD samples in unlabeled wild data usually faces challenges due to the unavailability of pure OOD samples.

To address these issues, this work proposes a novel loss-difference OOD detection framework (abbreviated as LoD) by *intentionally label-noisifying* unlabeled wild data. LoD adopts the filtering and learning strategy and its key lies in the loss-difference filtering module with *intentional label-noises*. In this module, the whole unlabeled wild data is intentionally labeled as a single $K+1$-th class (assuming that ID data contain K classes), and then trained together with the labeled ID data through the *fully-supervised* manner of $K+1$ classification. We would like to emphasize that such operations ingeniously transform the OOD filtering problem in unlabeled wild data into a label-noise learning problem, allowing us to solve the aforementioned issues by leveraging the inher-

*Corresponding author

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

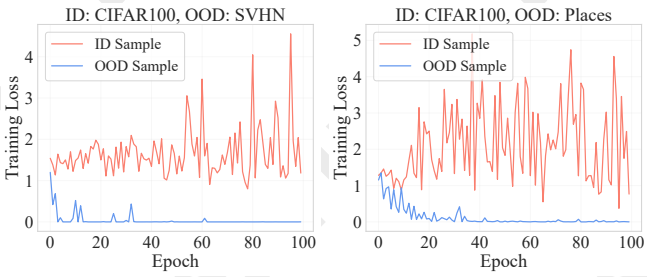

Figure 1: The cross-entropy loss changes of ID (*label-noise*) and OOD (*label-clean*) samples in unlabeled wild data when they are intentionally labeled as $K+1$-th class. These two types of samples typically exhibit different loss curves due to the differences in how learning progresses for each.

ent properties in label-noise learning. In this way, the OOD samples in unlabeled wild data is intentionally transformed into *label-clean* samples, while the ID counterparts become *label-noise* ones. The former naturally and seamlessly enables OOD samples in unlabeled wild data to jointly dominate the model learning with the labeled ID data, effectively addressing the model-bias issue. Meanwhile, the latter provides the key clues for differentiating ID and OOD samples in unlabeled wild data due to the significant differences in the loss curves between ID (*label-noise*) and OOD (*label-clean*) samples during training.

As shown in Figure 1, as the OOD samples in the unlabeled wild data are correctly labeled (*label-clean*), the model fits them well as learning progresses, leading to a gradual decrease and convergence of the loss curve. In contrast, the corresponding ID part, due to being incorrectly labeled (*label-noise*) and conflicting with the originally labeled ID data, exhibits not only higher loss values but also larger fluctuations during training. Such significant and natural differences allow us to employ classic clustering models, like K-means, to filter these OOD samples without requiring thresholds any longer. In particular, we also provide theoretical foundation to support the viability of such a module. Overall, our contributions can be highlighted as follows:

- Two potential issues (i.e., the model-bias issue and threshold selection dilemma) in this OOD research line are identified, providing some new insights for the subsequent modeling of OOD detection.
- The OOD filtering problem in unlabeled wild data is elegantly reformulated as a label-noise learning problem, leading to a novel LoD OOD detection framework, which not only effectively addresses the model-bias issue but also circumvents the threshold selection dilemma.
- Theoretical foundation is provided to support the viability of LoD. Meanwhile, extensive experiments are also conducted to demonstrate its superiority.

2 Related Works

2.1 Out-of-Distribution Detection

To improve the safety and reliability of models in detecting OOD data, various OOD methods have been developed [Zhu *et al.*, 2023; Zheng *et al.*, 2023; Wang *et al.*, 2023b; Yang *et al.*, 2024; Li *et al.*, 2024b; Behpour *et al.*, 2024; Fang *et al.*, 2024; Sharifi *et al.*, 2025], including adopting the classification confidence or entropy, modeling the ID density, leveraging auxiliary OOD data, and more. Among these, methods using auxiliary OOD data have demonstrated encouraging OOD detection performance over the counterpart without auxiliary data [Lee *et al.*, 2017; Bevandić *et al.*, 2018; Malinin and Gales, 2018; Liu *et al.*, 2020b; Chen *et al.*, 2021; Wei *et al.*, 2022; Du *et al.*, 2022; Wang *et al.*, 2023a; Sharifi *et al.*, 2025]. Despite the promise, there are two primary limitations: First, such data may not match the true distribution of OOD data in the wild; Second, collecting such data can be labor-intensive and inflexible. To address these limitations, recent works [Katz-Samuels *et al.*, 2022; Du *et al.*, 2024] proposed to leverage the unlabeled "in-the-wild" data due to they are freely collected during the deployment of any machine learning model in its operational environment, while also allowing for the capture of the true test-time OOD distribution.

Our work falls into this research line, and as mentioned earlier, though the methods in this research line have achieved varying degrees of success, they still face two potential weaknesses, i.e., the model-bias issue and the threshold selection dilemma. These motivate us to seek new methods to address these issues.

2.2 Training Neural Networks with Label Noises

In many applications [Guan *et al.*, 2018], due to the cost or difficulty of manual labeling, datasets are often annotated through online queries [Yuan *et al.*, 2024a] or crowdsourcing [Li *et al.*, 2024a]. Such annotations inevitably contain numerous mistakes, i.e., label-noises. When trained on the data mixed clean labels and noise labels, deep neural networks have been observed to first fit label-clean data during an early learning phase, and then start memorizing the label-noise data after sufficient epochs of training [Liu *et al.*, 2020a]. This phenomenon is independent of the optimizations used during training or the architectures of neural networks employed [Arpit *et al.*, 2017]. In particular, during the early learning phase, label-clean and label-noise data will have different loss curves due to the difference in how learning progresses for each type. This has been exploited in many label-noise learning works [Forouzesh *et al.*, 2022; Li *et al.*, 2023; Yuan *et al.*, 2024b; Lin *et al.*, 2024; Lienen and Hüllermeier, 2024; Yue and Jha, 2024]. For more information, please refer to the recent review work [Song *et al.*, 2022].

In this paper, we propose a novel loss-difference OOD detection framework by *intentionally label-noisifying* unlabeled wild data, which interestingly transforms the OOD filtering problem in unlabeled wild data into a label-noise learning problem. This enables us to leverage the aforementioned inherent phenomenon of label-noise learning to effectively filter OOD data from the unlabeled wild data.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

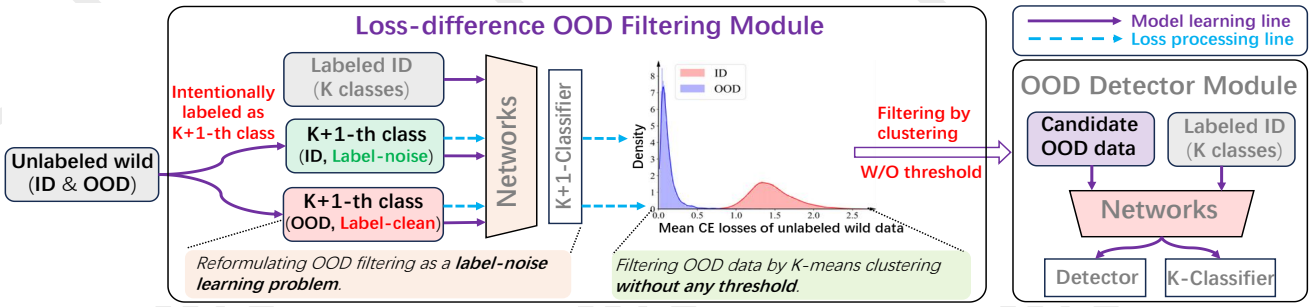

Figure 2: Overview of the loss-difference OOD detection framework by intentionally label-noisifying unlabeled wild data.

3 Methodology

3.1 Problem Formulation

Labeled ID Data Let $\mathcal{X}$ denote the input space and $\mathcal{Y} = \{1, ..., K\}$ represent the label space. Let $\mathcal{D}_{\text{in}}^{\text{train}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ denote the labeled training set drawn independently and identically from $\mathbb{P}_{\mathcal{X}\mathcal{Y}}$. $\mathbb{P}_{\text{in}}$ is the marginal distribution of $\mathbb{P}_{\mathcal{X}\mathcal{Y}}$ on $\mathcal{X}$, which is also referred to as the ID distribution.

Unlabeled Wild Data The main challenge in OOD detection is the lack of labeled OOD data. In particular, the sample space for potential OOD data can be prohibitively large, making it expensive to collect labeled OOD data. To model the realistic environment, recent works [Katz-Samuels *et al.*, 2022; Du *et al.*, 2024] incorporated unlabeled wild data $\mathcal{D}_{\text{wild}} = \{\widetilde{\boldsymbol{x}}_1, ..., \widetilde{\boldsymbol{x}}_m\}$ into OOD detection. Unlabeled wild data consists of potentially both ID and OOD data, and can be freely collected upon deploying an existing model in its natural habitats. Following [Katz-Samuels *et al.*, 2022], the Huber contamination model is employed to characterize the marginal distribution of the unlabeled wild data:

$$\mathbb{P}_{\text{wild}} := (1 - \pi)\mathbb{P}_{\text{in}} + \pi\mathbb{P}_{\text{out}}, \tag{1}$$

where $\pi \in (0, 1]$, and $\mathbb{P}_{\text{out}}$ is the OOD distribution defined over $\mathcal{X}$.

Learning Goal The learning framework aims to build the OOD detector g_θ and the multi-class classifier f_θ by leveraging data from both $\mathcal{D}_{\text{in}}^{\text{train}}$ and $\mathcal{D}_{\text{wild}}$. Following [Du *et al.*, 2024], we here are interested in the following measurements for model evaluation:

$$\downarrow \text{FPR}(g_\theta) := \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\text{out}}}(\mathbb{1}\{g_\theta(\boldsymbol{x}) = \text{in}\}),$$
$$\uparrow \text{TPR}(g_\theta) := \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\text{in}}}(\mathbb{1}\{g_\theta(\boldsymbol{x}) = \text{in}\})$$

3.2 Loss-Difference OOD Detection Framework

To effectively address the two aforementioned potential issues, i.e., the model-bias issue and the threshold-selection dilemma, we innovatively propose a novel loss-difference OOD detection framework (abbreviated as LoD) by *intentionally label-noisifying* unlabeled wild data. As shown in Figure 2, LoD follows the two-step strategy and contains two main modules, i.e., loss-difference OOD filtering module and OOD detector learning module. Next, we will elaborate on the specific details of each module.

Loss-difference OOD Filtering Module

In this part, a loss-difference filtering mechanism with *intentional label-noises* is developed, which ingeniously reformulates the OOD filtering problem in unlabeled wild data as a label-noise learning problem with *controllable label-noise ratio*. This allows us to leverage the inherent properties of label-noise learning demonstrated in Section 2.2 to effectively filter OOD data from the unlabeled wild data.

In specific, we first intentionally label the whole unlabeled wild data as a single $K+1$-th class (assuming that ID data contains K classes) and then train them together with labeled ID data in a *fully-supervised manner of* $K+1$ *classification*, as follows:

$$\mathcal{L} = \frac{1}{|\mathcal{B}_{\text{in}}^{\text{train}}|} \sum_{(\boldsymbol{x}_i, y_i)\sim\mathcal{B}_{\text{in}}^{\text{train}}} \ell(\hat{y}_i, y_i) + \frac{1}{|\mathcal{B}_{\text{wild}}|} \sum_{(\boldsymbol{x}_i, y_i)\sim\mathcal{B}_{\text{wild}}} \ell(\hat{y}_i, y_{K+1}), \quad \hat{y}_i = f(\boldsymbol{x}_i, \boldsymbol{\theta}), \tag{2}$$

where $f(\cdot, \boldsymbol{\theta}) \in \mathcal{F}$ denotes the $K+1$ classifier, $l(\cdot, \cdot)$ represents the vanilla cross-entropy (CE) loss. Each training batch consists of two parts: $\mathcal{B}_{\text{in}}^{\text{train}}$ and $\mathcal{B}_{\text{wild}}$, respectively sampled from labeled ID data and unlabeled wild data. Note that the ratio of $|\mathcal{B}_{\text{in}}^{\text{train}}| : |\mathcal{B}_{\text{wild}}| \geq 1$ is controllable. In fact, we indirectly control the label-noise ratio of the learning task by controlling this ratio (for more details, please refer to Section 4).

By labeling the entire unlabeled wild data as a single $K+1$-th class, the ID samples in $\mathcal{D}_{\text{wild}}$ are intentionally converted to *label-noise* samples while the OOD samples in $\mathcal{D}_{\text{wild}}$ become *label-clean* ones. According to the inherent phenomenon of early learning stage in label-noise learning, the loss curves of these two types of labeled samples will exhibit significant difference during the early learning stage, as shown in Figure 1. This discrepancy provides us a critical clue for effectively distinguishing between them. Therefore, after training the $K+1$ classifier, we conduct clustering operations on the loss values of unlabeled wild data gained during training so as to filter the OOD samples from $\mathcal{D}_{\text{wild}}$, *which does not need the filtering thresholds any more*. Particularly, clustering in our case has a well-defined number of clusters – Two – corresponding to the ID and OOD clusters represented by their distinct loss behaviors throughout the training process, e.g., higher loss-values for ID data while lower counterparts for OOD data.

Considering the efficiency issue, we here utilize the mean of loss-values during training as the new features for each

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

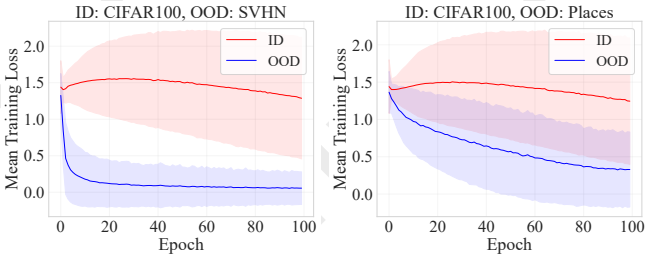

Figure 3: The mean cross-entropy loss curves respectively for all ID (*label-noise*) and OOD samples (*label-clean*) in unlabeled wild data when they are intentionally labeled as $K+1$-th class.

sample in $\mathcal{D}_{\text{wild}}$. Then the classic K-means clustering technique is employed to achieve the OOD samples filtering from unlabeled wild data. Let μ_1, μ_2 ($\mu_1 > \mu_2$) respectively denote the ID and OOD cluster centers, while d_1, d_2 respectively denote the distances between the corresponding sample and the two cluster centers. We filter the OOD samples from the unlabeled wild data by the following rule:

$$\hat{y} = \begin{cases} \text{ID data}, & \text{if } d_1 < d_2, \\ \text{OOD data}, & \text{otherwise.} \end{cases} \tag{3}$$

Remark. At first glance, labeling the entire set of unlabeled wild data as a single $K+1$-th class seems potentially to undermine the model learning. Intriguingly, however, once we switch to consider filtering the OOD samples from the label-noise perspective, such operations, just on the contrary, bring at least the following three-fold advantages:

- **First**, OOD samples in unlabeled wild data are correctly labeled (*label-clean*), naturally and seamlessly enabling them to jointly dominate the model learning with labeled ID data, thus effectively circumventing the model-bias issue.
- **Second**, as mentioned earlier, the ID samples in $\mathcal{D}_{\text{wild}}$ being erroneously labeled (*label-noise*) as $K+1$-th class contradict the label-correct ones in labeled ID data $\mathcal{D}_{\text{in}}^{\text{train}}$, thereby resulting in their loss curves exhibiting both higher values and greater fluctuations compared to those of OOD samples, such as their mean-loss curves shown in Figure 3. This discrepancy provides us a fairly clear signal to distinguish ID and OOD samples in the unlabeled wild data.
- **Third**, our LoD is data-centric in nature, wherein we just relabel the unlabeled wild data as the intentionally $K+1$-th class without any modifications to the network architectures we employed. This endows our LoD stronger applicability (for related experiments, please refer to Appendix E).

To solidly demonstrate the viability of this OOD filtering module, we also provide the theoretical analyses to support our first two claims, which will be detailed in Section 4.

OOD Detector Learning Module

After obtaining the candidate OOD samples $\mathcal{D}_{\text{out}}$ from the unlabeled wild data, we training an OOD detector g_θ using them together with labeled ID data $\mathcal{D}_{\text{in}}^{\text{train}}$. Similar to [Du *et al.*, 2024], we adopt the following optimization objective:

$$\mathcal{L}(g_\theta) = \mathbb{E}_{\boldsymbol{x}\in\mathcal{D}_{\text{in}}^{\text{train}}} \mathbb{1}\{g_\theta(\boldsymbol{x}) \le 0\} + \mathbb{E}_{\widetilde{\boldsymbol{x}}\in\mathcal{D}_{\text{out}}} \mathbb{1}\{g_\theta(\widetilde{\boldsymbol{x}}) > 0\}, \tag{4}$$

where the binary sigmoid loss is employed as the smooth approximation of the $0/1$ loss to make it tractable. In addition, a K-class classifier f_θ is also trained using CE loss on labeled ID data along with g_θ to ensure the ID accuracy. Algorithm 1 denotes the entire workflow of our LoD.

Algorithm 1 LoD OOD Detection Framework

Input: In-distribution data $\mathcal{D}_{in}^{train}$, unlabeled wild data $\mathcal{D}_{\text{wild}}$, Max Epoch T, Batch size $|\mathcal{B}|$.
Output: OOD detector g_θ and classifier f_θ.

 # Loss-difference OOD detection module
Initializing: Model parameters $\boldsymbol{\theta}$, $\mathcal{D}_{\text{wild}}$ labeled as $K+1$-th class, loss record matrix $\mathcal{V} = \{\} \in \mathbb{R}^{|\mathcal{D}_{\text{wild}}|\times T}$.
for *epoch* $= 1$ to T **do**
  Batch $\mathcal{B} = \mathcal{B}_{\text{in}}^{\text{train}} \cup \mathcal{B}_{\text{wild}}$, where $\mathcal{B}_{\text{in}}^{\text{train}}$ samples from $\mathcal{D}_{\text{in}}^{\text{train}}$ and $\mathcal{B}_{\text{wild}}$ samples from $\mathcal{D}_{\text{wild}}$.
  Update $K+1$ classifier $f(\cdot, \boldsymbol{\theta})$ based on Eq.(2).
  Record losses of unlabeled wild data. $\mathcal{V} \leftarrow \mathcal{V} \cup \{l_i \mid i \in (1, |\mathcal{B}_{\text{wild}}|)\}$
end for
Calculate the mean-loss set of wild data $\{u_i\} = \texttt{mean}(\mathcal{V})$, where $\{u_i\} \in \mathbb{R}^{|\mathcal{D}_{\text{wild}}|\times 1}$.
Cluster and detect candidate OOD samples set $\mathcal{D}_{\text{out}}$ based on Eq.(3).
 # OOD detector learning module
for *epoch* $= 1$ to T **do**
  Batch $\mathcal{B} = \mathcal{B}_{\text{in}}^{\text{train}} \cup \mathcal{B}_{\text{out}}$, where $\mathcal{B}_{\text{in}}^{\text{train}}$ samples from $\mathcal{D}_{\text{in}}^{\text{train}}$ and $\mathcal{B}_{\text{out}}$ samples from $\mathcal{D}_{\text{out}}$.
  Update f_θ and g_θ based on Eq.(4).
end for

4 Theoretical Analysis

4.1 Mitigation of The Model Bias

For the first claim in Subsection 3.2, we here provide a theoretical analysis at the gradient level to demonstrate that in our LoD framework, labeled ID data and OOD data in $\mathcal{D}_{\text{wild}}$ can jointly dominate the model learning. Let N_1 denote the number of samples in $\mathcal{B}_{\text{in}}$, while N_2 and N_3 denote the number of samples respectively from IDs and OODs in $\mathcal{B}_{\text{wild}}$. Then Eq.(2) can be rewritten in the following form:

$$\mathcal{L} = \underbrace{\frac{1}{N_1}\sum_{(\boldsymbol{x}_i,y_i)\sim\mathcal{B}_{\text{in}}} \ell(\hat{y}_i, y_i) + \frac{1}{N_2}\sum_{(\boldsymbol{x}_i,y_i)\sim\mathcal{B}_{\text{wild}}} \ell(\hat{y}_i, y_{K+1})}_{\text{ID data}} + \underbrace{\frac{1}{N_3}\sum_{(\boldsymbol{x}_i,y_i)\sim\mathcal{B}_{\text{wild}}} \ell(\hat{y}_i, y_{K+1})}_{\text{OOD data}}. \tag{5}$$

Let $\nabla\mathcal{L}_{N_k} = \frac{1}{N_k}\sum_{i=1}^{N_k} \nabla l(\hat{y}_i, y_i), k = 1,2,3$, denote the gradient of the corresponding part with respect to the model parameters $\boldsymbol{\theta}$. For the OOD samples in $\mathcal{B}_{\text{wild}}$, evidently,

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

they are correctly labeled (label-clean), the model parameters therefore will be updated in the correct gradient direction.

As for ID samples, they consist of two parts: one part sampled from $\mathcal{D}_{\text{in}}^{\text{train}}$ (*label-clean*), and the other part sampled from $\mathcal{D}_{\text{wild}}$ (*label-noise*). Then the update of the model parameters $\boldsymbol{\theta}$ is as follows:

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^{t} - \eta(\nabla\mathcal{L}_{N_1} + \nabla\mathcal{L}_{N_2}), \tag{6}$$

where t denotes the number of steps for model update, and η is the learning rate. According to Eq.(6), the update of $\boldsymbol{\theta}$ is determined by $(\nabla\mathcal{L}_{N_1} + \nabla\mathcal{L}_{N_2})$. Since $|\mathcal{B}_{\text{in}}^{\text{train}}| > |\mathcal{B}_{\text{wild}}|$ and $|\mathcal{B}_{\text{wild}}| \geq N_2$, we have

$$|\mathcal{B}_{\text{in}}^{\text{train}}| > |\mathcal{B}_{\text{wild}}| \geq N_2.$$

This indicates that correctly labeled ID samples dominate the updating of model parameters, especially when $|\mathcal{B}_{\text{in}}^{\text{train}}| \gg N_2$. In summary, we have the labeled ID data $\mathcal{D}_{\text{in}}^{\text{train}}$ and the OOD data in $\mathcal{D}_{\text{wild}}$ that can jointly dominate the model learning, thus effectively addressing the model-bias issue.

4.2 Discriminability between ID and OOD CE Mean-Losses

As mentioned earlier, the key to our LoD lies in ingeniously transforming the OOD filtering problem into a label-noise learning problem with controllable label-noise ratio, which allows us to leverage the established theoretical foundation of label-noise learning [Liu *et al.*, 2020a; Yue and Jha, 2024] to ensure the feasibility of our LoD. The work [Liu *et al.*, 2020a] has shown that the phenomenon in early learning stage, when training with noisy labels, is intrinsic to high-dimensional classification tasks, even in the simplest setting, far from being a peculiar feature of deep neural networks. Therefore, for the second claim in Subsection 3.2, a theoretical analysis of loss gap between ID (label-noise) and OOD (label-clean) data in $\mathcal{D}_{\text{wild}}$ is provided here using a similar setting in [Liu *et al.*, 2020a].

Considering a two class dataset that consists of n independent samples $(\boldsymbol{x}_i, y_i)$ drawn from a mixture of two Gaussians in $\mathbb{R}^d$ as follows.

$$\boldsymbol{x} \sim \mathcal{N}(+\boldsymbol{v}, \sigma^2\boldsymbol{I}_{d\times d}), \quad \text{if } y = +1$$
$$\boldsymbol{x} \sim \mathcal{N}(-\boldsymbol{v}, \sigma^2\boldsymbol{I}_{d\times d}), \quad \text{if } y = -1,$$

where $\boldsymbol{v}$ is an arbitrary unit vector in $\mathbb{R}^d$ and σ^2 is a small constant. Denote y as the true hidden label and $\widetilde{y}$ as the observed label. Assume that for any sample $\boldsymbol{x}_i$,

$$\widetilde{y} = \begin{cases} y_i, & \text{with probability } 1-\Delta, \\ -y_i, & \text{with probability } \Delta, \end{cases} \tag{7}$$

where $\Delta \in (0, 1/2)$ is the label-noise ratio. Let us consider a linear classifier $f(\cdot, \boldsymbol{\theta})$ trained by gradient descent on CE loss:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^{2\times d}} \mathcal{L}_{CE}(\boldsymbol{\theta}) := -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{2} y_i \log(f(\boldsymbol{x_i}, \boldsymbol{\theta})). \tag{8}$$

In order to correctly classify the true classes well (and not overfit to the noisy labels), the rows of $\boldsymbol{\theta}$ should be correlated with the vector $\boldsymbol{v}$. Let $\nabla\mathcal{L}_{CE}(\boldsymbol{\theta})$ denote the gradient of Eq.(8). According to [Liu *et al.*, 2020a], we have the following lemma.

Lemma 1 (Early-learning succeeds [Liu *et al.*, 2020a])**.** *Denote by $\{\boldsymbol{\theta}_t\}$ the iterates of gradient descent with step size η. For any $\Delta \in (0, 1/2)$, there exists a constant δ_Δ, depending only on Δ, such that if $\delta \leq \delta_\Delta$, then with high probability $1 - o(1)$, there exists a $T = \Omega(1/\eta)$ such that: for all $t < T$, we have $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\| \leq 1$ and*

$$-\nabla\mathcal{L}_{CE}(\boldsymbol{\theta}_t)^T\boldsymbol{v}/\|\nabla\mathcal{L}_{CE}(\boldsymbol{\theta}_t)\| \geq 1/6.$$

Lemma 1 indicates that under the condition of label-noise ratio Δ, the model parameters $\boldsymbol{\theta}$ update along the proper gradient direction during the early learning stage. This means, during this period, the loss curves of ID (*label-noise*) and OOD (*label-clean*) samples in $\mathcal{D}_{\text{wild}}$ will have significantly different characteristics, with larger loss values and greater fluctuations for ID samples versus smaller loss values and smaller fluctuations for OOD ones. To theoretically analyze this, we have the following proposition.

Proposition 1. *Let l_i denote the loss value of each sample in $\mathcal{D}_{wild}$, which is bounded by R. $\overline{l_{in}} = \frac{1}{|\mathcal{D}_{in}^{wild}|}\sum_{i\in\mathcal{D}_{in}^{wild}} l_i$ and $\overline{l_{out}} = \frac{1}{|\mathcal{D}_{out}^{wild}|}\sum_{i\in\mathcal{D}_{out}^{wild}} l_i$ respectively denote the mean losses of ID and OOD sets from unlabeled wild data $\mathcal{D}_{wild}$, and $n = |\mathcal{D}_{in}^{wild}| + |\mathcal{D}_{out}^{wild}|$. Under the Lemma 1, with high probability, we have*

$$\overline{l_{in}} - \overline{l_{out}} \geq 1 - 2e^{-\boldsymbol{\theta}^T\boldsymbol{v} + \frac{1}{2}\|\boldsymbol{\theta}\|^2\delta^2} - \mathcal{O}(\frac{R}{\sqrt{n}}).$$

Proposition 1 demonstrates that the cross-entropy mean losses of ID and OOD samples in $\mathcal{D}_{\text{wild}}$ are distinguishable, just as the two curves shown in Figure 3. The proof is provided in Appendix A of supplementary materials (https://github.com/ChuanxingGeng/LoD).

5 Experiments

5.1 Implementation Details

Our LoD (https://github.com/ChuanxingGeng/LoD) framework contains two main modules, i.e., loss-difference OOD filtering module and OOD detector learning module. For these two modules, we follow [Du *et al.*, 2024; Katz-Samuels *et al.*, 2022] and employ Wide ResNet [Zagoruyko, 2016] with 40 layers and widen factor of 2 as the backbone. Moreover, for the loss-difference OOD filtering module, we use stochastic gradient descent with a momentum of 0.9 as the optimizer, and set the initial learning rate to 0.01. We train for 100 epochs using cosine learning rate decay, a batch size of 128 in which $|\mathcal{B}_{\text{in}}^{\text{train}}| : |\mathcal{B}_{\text{wild}}| = 3 : 1$, and a dropout rate of 0.3. For the OOD detector learning module, similar to [Du *et al.*, 2024], we load a pre-trained ID classifier and add an additional linear layer which utilize the penultimate-layer features of ID classifier for binary classification. The initial learning rate is set to 0.001, and the remaining training configurations are consistent with those of the former module. All experiments are conducted on a single NVIDIA RTX 3090 GPU.

Evaluation Metrics. Similar to [Du *et al.*, 2024; Katz-Samuels *et al.*, 2022], we adopt the following evaluation metrics: (1) the false positive rate (FPR95) of OOD examples when true positive rate of ID examples is at 95%, (2)

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Methods	OOD Dataset												ACC
	SVHN		Places		LSUN-Crop		LSUN-Resize		Textures		Average		
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	
π=0.1													
OE (ICLR'19)	1.57	99.63	60.24	83.43	3.83	99.26	0.93	99.79	27.89	93.35	18.89	95.09	71.65
Energy(w/OE) (NeurIPS'20)	1.47	99.68	54.67	86.09	2.52	99.44	2.68	99.50	37.26	91.26	19.72	95.19	73.46
WOODS (ICML'22)	0.12	99.96	29.58	90.60	0.11	99.96	0.07	99.96	9.12	96.65	7.80	97.43	75.22
SAL (ICLR'24)	0.07	99.95	3.53	99.06	0.06	99.94	0.02	99.95	5.73	98.65	1.88	99.51	73.71
LoD (Ours)	**0**	**100**	**3.34**	**99.16**	**0**	**100**	**0**	**100**	**4.79**	**98.87**	**1.63**	**99.61**	73.85
π=0.5													
OE (ICLR'19)	2.86	99.05	40.21	88.75	4.13	99.05	1.25	99.38	22.86	94.63	14.26	96.17	73.38
Energy(w/OE) (NeurIPS'20)	2.71	99.34	34.82	90.05	3.27	99.18	2.54	99.23	30.16	94.76	14.70	96.51	72.76
WOODS (ICML'22)	0.17	99.80	21.87	93.73	0.48	99.61	1.24	99.54	9.95	95.97	6.74	97.73	73.91
SAL (ICLR'24)	0.02	99.98	**1.27**	99.62	0.04	99.96	0.01	99.99	5.64	99.16	1.40	99.74	73.77
LoD (Ours)	**0**	**100**	1.53	**99.66**	**0**	**100**	**0**	**100**	**3.72**	**99.19**	**1.05**	**99.77**	74.32
π=0.9													
OE (ICLR'19)	0.84	99.36	19.78	96.29	1.64	99.57	0.51	99.75	12.74	94.95	7.10	97.98	72.02
Energy(w/OE) (NeurIPS'20)	0.97	99.64	17.52	96.53	1.36	99.73	0.94	99.59	14.01	95.73	6.96	98.24	73.62
WOODS (ICML'22)	0.05	99.98	11.34	95.83	0.07	99.99	0.03	99.99	6.72	98.73	3.64	98.90	73.86
SAL (ICLR'24)	0.03	99.99	2.79	99.89	0.05	99.99	0.01	99.99	5.88	**99.53**	1.75	**99.88**	74.01
LoD (Ours)	**0**	**100**	**0.48**	**99.90**	**0**	**100**	**0**	**100**	**2.78**	99.41	**0.65**	99.86	74.34

Table 1: Evaluation results of FPR95↓ (%), AUROC↑ (%) and ACC↑ (%) on standard benchmarks. CIFAR100 is ID, and bold numbers highlight the best results.

Methods	Dataset										ACC
	CIFAR10		CIFAR+10		CIFAR+50		TinyImageNet		Average		
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	
$\pi = 0.1$											
OE (ICLR'19)	30.83	94.9	11.40	97.98	22.21	95.98	82.3	75.34	36.69	91.05	91.45
Energy(w/OE) (NeurIPS'20)	38.36	89.85	16.40	96.51	36.18	90.49	88.48	74.30	44.86	87.79	86.98
WOODS (ICML'22)	32.33	93.70	22.39	95.95	22.12	95.76	74.60	78.62	37.86	91.01	92.43
SAL (ICLR'24)	12.95	97.35	4.76	98.88	10.66	97.63	48.35	86.71	19.18	95.14	91.50
LoD (Ours)	**2.56**	**99.40**	**1.50**	**99.62**	**1.96**	**99.39**	**47.61**	**91.55**	**13.41**	**97.49**	91.44
$\pi = 0.5$											
OE (ICLR'19)	13.77	97.68	4.08	99.09	9.80	98.27	76.13	80.62	25.95	93.92	91.82
Energy(w/OE) (NeurIPS'20)	9.16	97.70	3.70	98.98	10.01	97.43	75.93	83.58	24.70	94.42	87.91
WOODS (ICML'22)	17.89	96.64	12.50	97.69	12.68	97.68	70.60	81.42	28.42	93.36	92.53
SAL (ICLR'24)	12.76	97.38	4.84	98.87	10.86	97.60	48.17	86.77	19.16	95.16	91.39
LoD (Ours)	**2.32**	**99.47**	**1.04**	**99.71**	**1.96**	**99.46**	**46.44**	**91.52**	**12.94**	**97.54**	91.33
$\pi = 0.9$											
OE (ICLR'19)	6.40	98.71	1.56	99.50	4.94	98.97	67.45	84.98	20.09	95.54	92.10
Energy(w/OE) (NeurIPS'20)	2.95	98.63	1.30	99.41	2.18	98.52	58.84	88.92	16.32	96.37	89.58
WOODS (ICML'22)	12.82	97.50	10.98	98.03	10.51	98.07	68.01	82.82	25.58	94.11	92.17
SAL (ICLR'24)	12.95	97.34	4.30	98.91	11.11	97.56	49.19	86.66	19.39	95.12	91.41
LoD (Ours)	**2.19**	**99.45**	**1.04**	**99.77**	**1.90**	**99.45**	**45.24**	**91.80**	**12.59**	**97.62**	91.50

Table 2: Evaluation results of FPR95↓ (%), AUROC↑ (%) and ACC↑ (%) on hard benchmarks, and bold numbers highlight the best results..

Area Under the Receiver Operating Characteristic curve (AUROC), and (3) ID classification Accuracy (ACC).

To comprehensively evaluate our LoD framework, we conduct extensive experiments on both standard benchmarks and hard benchmarks (newly curated in this paper) detailed in the following subsections. Moreover, limited by space, we defer additional experiments in the supplementary materials, including results on CIFAR10 (Appendix C), results on unseen OOD datasets (Appendix D), and results on different network structures (Appendix E).

5.2 Experiments on Standard Benchmarks

Datasets. For standard benchmarks, we here follow [Du *et al.*, 2024; Katz-Samuels *et al.*, 2022], and choose CIFAR100 as in-distribution (ID) datasets ($\mathbb{P}_{\text{in}}$). For the out-of-distribution (OOD) test datasets ($\mathbb{P}_{\text{out}}$), we use a diverse collection of natural image datasets including SVHN [Netzer *et al.*, 2011], Textures [Cimpoi *et al.*, 2014], Places [Zhou *et al.*, 2017], LSUN-Crop [Yu *et al.*, 2015] and LSUN-Resize [Yu *et al.*, 2015]. For the unlabeled wild data ($\mathbb{P}_{\text{wild}}$), we follow [Du *et al.*, 2024] and mix datasets by combining a subset of ID data with OOD data under different mixture proportions $\pi \in \{0.1, 0.5, 0.9\}$. Specifically, the ID dataset is split into two equal halves (25,000 images per half), with one half used to mix with an OOD dataset (e.g., SVHN) to create the unlabeled wild data ($\mathbb{P}_{\text{wild}}$).

Main Results. We mainly compare our LoD with 4 latest methods using unlabeled wild data including Outlier Exposure (OE) [Hendrycks *et al.*, 2018], energy-regularization learning (Energy) [Liu *et al.*, 2020b], WOODS [Katz-Samuels *et al.*, 2022], and SAL [Du *et al.*, 2024]. Table 1 presents a comprehensive comparison of different methods on standard benchmarks, highlighting the substantial advantages of our proposed LoD. Across all datasets and π values, our approach consistently delivers superior performance, achieving an FPR95 close to 0%, which is significantly lower

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

than the current SOTA baseline, SAL. Notably, on the most challenging Textures, our method outperforms SAL with substantial reductions in FPR95 by 0.94%, 1.92%, and 3.10% for $\pi = 0.1, 0.5, 0.9$, respectively. Moreover, while existing SOTA methods demonstrate strong performance in AUROC, our LoD achieves notable improvements even in this aspect. Importantly, our LoD maintains competitive in-distribution accuracy, matching or surpassing the performance of SOTA methods such as SAL and WOODS across various π values.

5.3 Experiments on Hard Benchmarks

Datasets. In the settings of standard benchmarks, the ID and OOD samples are sourced from different datasets with inherently distinct distributions, which actually indirectly reduces the difficulty of OOD detection. As shown in Table 1, many methods, including ours, have achieved exceptionally high performance. To further demonstrate the advantages of our LoD, we here curate more challenging benchmarks, called hard benchmarks. Different from standard benchmarks, the ID and OOD samples on hard benchmarks come from the same dataset with different classes.

In specific, taking CIFAR10 as an example, we first randomly select 6 classes as ID data and the remaining 4 classes as OOD data. Then, similar to the splitting protocol of standard benchmarks, the training set of 6 ID classes is divided into two halves (15,000 images per half). One half is used as labeled ID data, while the other half is mixed with the data from 4 OOD classes to create the unlabeled wild data. We here select CIFAR10, CIFAR+10, CIFAR+50, and TinyImageNet [Vaze *et al.*, 2022] to curate the hard OOD benchmarks, and more details can be found in Appendix B of supplementary materials.

Main Results. Since the four methods we compared do not conduct the experiments on these benchmarks, we reproduce the results according to the source codes provided by them. Table 2 reports the detailed results on hard benchmarks. Across all datasets and under various π values, our LoD achieves better FPR95 and AUROC performance compared to existing methods, indicating that its OOD detection has stronger generalization. Notably, compared to the SOTA baseline SAL [Du *et al.*, 2024], our method reduces FPR95 by substantial margins of 5.77%, 6.22%, and 6.80% on average when $\pi = 0.1, 0.5, 0.9$, respectively. Especially on CIFAR10, where LoD outperforms SAL more than 10% in case of FPR95. In particular, on the most challenging TinyImageNet, LoD consistently surpasses SAL by a large margin of 4.84%, 4.75%, and 5.14% in terms of AUROC when $\pi = 0.1, 0.5, 0.9$, respectively. Besides, our LoD also maintains competitive ID classification accuracy compared to the SOTA baseline, comprehensively demonstrating the effectiveness of our LoD.

5.4 Experiments on Different Ratios and Epochs

Results on Different Ratios of $|\mathcal{B}_{\text{in}}^{\text{train}}|/|\mathcal{B}_{\text{wild}}|$

According to Section 4.1, the larger the ratio $|\mathcal{B}_{\text{in}}^{\text{train}}|/|\mathcal{B}_{\text{wild}}|$, the more dominant the labeled ID data in $\mathcal{D}_{\text{in}}^{\text{train}}$ and the OOD data in $\mathcal{D}_{\text{wild}}$ are in model learning, thus leading to better model performance. To verify this, we conduct experiments in different ratios of $|\mathcal{B}_{\text{in}}^{\text{train}}|/|\mathcal{B}_{\text{wild}}|$. Figure 4 shows the results. As the ratio increases, the model performance consistently improves across all benchmarks, strongly supporting our claim. Considering computational efficiency, $|\mathcal{B}_{\text{in}}^{\text{train}}|/|\mathcal{B}_{\text{wild}}|$ is set to $3:1$ in all of our experiments.

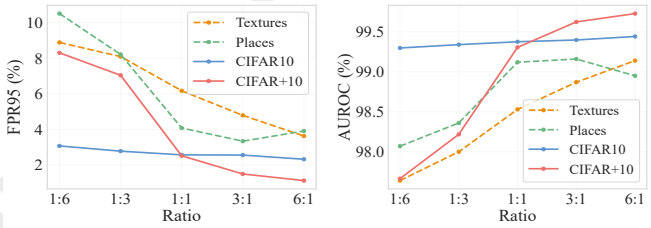

Figure 4: Experiments in different rations ($|\mathcal{B}_{\text{in}}^{\text{train}}|/|\mathcal{B}_{\text{wild}}|$) on standard benchmarks (dashed lines) and hard benchmarks (solid lines).

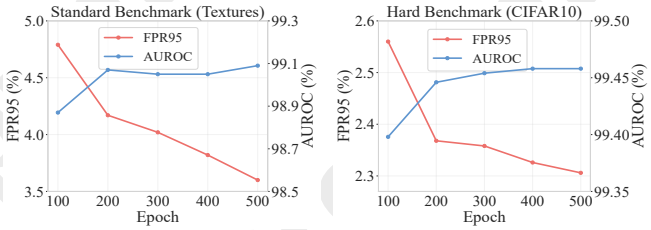

Figure 5: The impacts of training epochs on results respectively in standard and hard benchmarks.

Impact of Epoch in Early-learning Succeeds

As shown in Proposition 1, the early-learning succeeds is the key to our LoD. To clearly demonstrate the appropriate number of training epochs, we conduct the epoch experiments on standard benchmark (take Textures as an example) and hard benchmark (take CIFAR10 as an example) respectively. Figure 5 shows the results, and we can observe a steady performance improvement in our LoD from 100 to 500 training epochs. At first glance, this phenomenon seems inconsistent with the early-learning succeeds in the traditional label-noise learning field, which is usually shorter. However, please note that in our work setting, the label-noise ratio is controlled within an appropriate range by controlling the ratio of $|\mathcal{B}_{\text{in}}^{\text{train}}|/|\mathcal{B}_{\text{wild}}|$, meaning that correctly labeled samples all along dominate the network's learning. This further verifies the operability of LoD due to the long period early-learning succeeds. Considering efficiency issues, the training epochs of all experiments in this paper are set to 100 epochs.

6 Conclusion

In this paper, we innovatively propose a loss-difference OOD detection framework by *intentionally label-noisifying* unlabeled wild data, which ingeniously transforms the OOD filtering problem in unlabeled wild data into a label-noise learning problem with controllable label-noise ratio. Importantly, LoD not only effectively addresses the model-bias issue commonly associated with existing methods, but also circumvents the threshold selection dilemma inherent in these approaches.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Acknowledgments

This research was supported in part by the National Natural Science Foundation of China (62106102, 62376126, 62272229), in part by the NSFC-Hong Kong joint collaboration research fund CRS_HKU703/24, in part by the Hong Kong Scholars Program under Grant XJ2023035, in part by the Fundamental Research Funds for the Central Universities under Grant NS2024058.

References

[Abati *et al.*, 2019] Davide Abati, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Latent space autoregression for novelty detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 481–490, 2019.

[Arpit *et al.*, 2017] Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017.

[Behpour *et al.*, 2024] Sima Behpour, Thang Long Doan, Xin Li, Wenbin He, Liang Gou, and Liu Ren. Gradorth: a simple yet efficient out-of-distribution detection with orthogonal projection of gradients. *Advances in Neural Information Processing Systems*, 36, 2024.

[Bevandić *et al.*, 2018] Petra Bevandić, Ivan Krešo, Marin Oršić, and Siniša Šegvić. Discriminative out-of-distribution detection for semantic segmentation. *arXiv preprint arXiv:1808.07703*, 2018.

[Chen *et al.*, 2021] Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Atom: Robustifying out-of-distribution detection using outlier mining. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21*, pages 430–445. Springer, 2021.

[Cimpoi *et al.*, 2014] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

[Du *et al.*, 2022] Xuefeng Du, Xin Wang, Gabriel Gozum, and Yixuan Li. Unknown-aware object detection: Learning what you don't know from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13678–13688, 2022.

[Du *et al.*, 2024] Xuefeng Du, Zhen Fang, Ilias Diakonikolas, and Yixuan Li. How does unlabeled data provably help out-of-distribution detection? *arXiv preprint arXiv:2402.03502*, 2024.

[Fang *et al.*, 2024] Zhen Fang, Yixuan Li, Feng Liu, Bo Han, and Jie Lu. On the learnability of out-of-distribution detection. *Journal of Machine Learning Research*, 25, 2024.

[Forouzesh *et al.*, 2022] Mahsa Forouzesh, Hanie Sedghi, and Patrick Thiran. Leveraging unlabeled data to track memorization. *arXiv preprint arXiv:2212.04461*, 2022.

[Guan *et al.*, 2018] Melody Guan, Varun Gulshan, Andrew Dai, and Geoffrey Hinton. Who said what: Modeling individual labelers improves classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[Hendrycks and Gimpel, 2016] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[Hendrycks *et al.*, 2018] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

[Katz-Samuels *et al.*, 2022] Julian Katz-Samuels, Julia B Nakhleh, Robert Nowak, and Yixuan Li. Training ood detectors in their natural habitats. In *International Conference on Machine Learning*, pages 10848–10865. PMLR, 2022.

[Lee *et al.*, 2017] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.

[Li *et al.*, 2023] Yifan Li, Hu Han, Shiguang Shan, and Xilin Chen. Disc: Learning from noisy labels via dynamic instance-specific selection and correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24070–24079, 2023.

[Li *et al.*, 2024a] Huiru Li, Liangxiao Jiang, and Chaoqun Li. Certainty weighted voting-based noise correction for crowdsourcing. *Pattern Recognition*, 150:110325, 2024.

[Li *et al.*, 2024b] Tianqi Li, Guansong Pang, Xiao Bai, Wenjun Miao, and Jin Zheng. Learning transferable negative prompts for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17584–17594, 2024.

[Lienen and Hüllermeier, 2024] Julian Lienen and Eyke Hüllermeier. Mitigating label noise through data ambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13799–13807, 2024.

[Lin *et al.*, 2024] Yexiong Lin, Yu Yao, and Tongliang Liu. Learning the latent causal structure for modeling label noise. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[Liu *et al.*, 2020a] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33:20331–20342, 2020.

[Liu *et al.*, 2020b] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

[Malinin and Gales, 2018] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018.

[Netzer *et al.*, 2011] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 4. Granada, 2011.

[Nguyen *et al.*, 2015] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.

[Sharifi *et al.*, 2025] Sina Sharifi, Taha Entesari, Bardia Safaei, Vishal M Patel, and Mahyar Fazlyab. Gradient-regularized out-of-distribution detection. In *European Conference on Computer Vision*, pages 459–478. Springer, 2025.

[Song *et al.*, 2022] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems*, 34(11):8135–8153, 2022.

[Vaze *et al.*, 2022] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. *the International Conference on Learning Representations*, abs/2110.06207, 2022.

[Wang *et al.*, 2022] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4921–4930, 2022.

[Wang *et al.*, 2023a] Qizhou Wang, Zhen Fang, Yonggang Zhang, Feng Liu, Yixuan Li, and Bo Han. Learning to augment distributions for out-of-distribution detection. *Advances in neural information processing systems*, 36:73274–73286, 2023.

[Wang *et al.*, 2023b] Qizhou Wang, Junjie Ye, Feng Liu, Quanyu Dai, Marcus Kalander, Tongliang Liu, Jianye Hao, and Bo Han. Out-of-distribution detection with implicit outlier transformation. *In International Conference on Learning Representations*, 2023.

[Wei *et al.*, 2022] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *International conference on machine learning*, pages 23631–23644. PMLR, 2022.

[Yang *et al.*, 2024] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024.

[Yu *et al.*, 2015] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[Yuan *et al.*, 2024a] Shunjie Yuan, Xinghua Li, Yinbin Miao, Haiyan Zhang, Ximeng Liu, and Robert H Deng. Combating noisy labels by alleviating the memorization of dnns to noisy labels. *IEEE Transactions on Multimedia*, 2024.

[Yuan *et al.*, 2024b] Suqin Yuan, Lei Feng, and Tongliang Liu. Early stopping against label noise without validation data. In *The Twelfth International Conference on Learning Representations*, 2024.

[Yue and Jha, 2024] Chang Yue and Niraj K Jha. Ctrl: Clustering training losses for label error detection. *IEEE Transactions on Artificial Intelligence*, 2024.

[Zagoruyko, 2016] Sergey Zagoruyko. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[Zheng *et al.*, 2023] Haotian Zheng, Qizhou Wang, Zhen Fang, Xiaobo Xia, Feng Liu, Tongliang Liu, and Bo Han. Out-of-distribution detection learning with unreliable out-of-distribution sources. *Advances in Neural Information Processing Systems*, 36:72110–72123, 2023.

[Zhou *et al.*, 2017] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

[Zhu *et al.*, 2023] Jianing Zhu, Yu Geng, Jiangchao Yao, Tongliang Liu, Gang Niu, Masashi Sugiyama, and Bo Han. Diversified outlier exposure for out-of-distribution detection via informative extrapolation. *Advances in Neural Information Processing Systems*, 36:22702–22734, 2023.