Learning with Open-world Noisy Data via Class-independent Margin in Dual Representation Space

Linchao Pan[1], **Can Gao**[1, 2*], **Jie Zhou**[2, 3], **Jinbao Wang**[2, 3]

[1]College of Computer Science and Software Engineering, Shenzhen University
[2]Guangdong Provincial Key Laboratory of Intelligent Information Processing
[3]National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University
linchaopan2022@163.com, 2005gaocan@163.com, jie_jpu@163.com, wangjb@szu.edu.cn

Abstract

Learning with Noisy Labels (LNL) aims to improve the model generalization when facing data with noisy labels, and existing methods generally assume that noisy labels come from known classes, called closed-set noise. However, in real-world scenarios, noisy labels from similar unknown classes, i.e., open-set noise, may occur during the training and inference stage. Such open-world noisy labels may significantly impact the performance of LNL methods. In this study, we propose a novel dual-space joint learning method to robustly handle the open-world noise. To mitigate model overfitting on closed-set and open-set noises, a dual representation space is constructed by two networks. One is a projection network that learns shared representations in the prototype space, while the other is a One-Vs-All (OVA) network that makes predictions using unique semantic representations in the class-independent space. Then, bi-level contrastive learning and consistency regularization are introduced in two spaces to enhance the detection capability for data with unknown classes. To benefit from the memorization effects across different types of samples, class-independent margin criteria are designed for sample identification, which selects clean samples, weights closed-set noise, and filters open-set noise effectively. Extensive experiments demonstrate that our method outperforms the state-of-the-art methods and achieves an average accuracy improvement of 4.55% and an AUROC improvement of 6.17% on CIFAR80N.

Code — https://github.com/iCAN-SZU/LOND-DRS

Introduction

Deep Neural Networks (DNNs) have achieved great success in many fields, while their effectiveness heavily relies on a large amount of data with accurate and complete labels. Nevertheless, these high-quality labels are generally annotated by domain experts at the expense of high cost, thus some less-costly techniques are used to collect large-scale datasets, such as web crawling and crowdsourcing (Song et al. 2023). As a result, these non-expert manners inevitably introduce annotation errors and generate datasets with noisy labels. It has been shown that the over-parameterized DNNs can easily memorize the noisy labels (Zhang et al. 2017), thereby degrading model generalization and performance. Therefore, it is crucial to develop effective methods for DNNs to learn from data with noisy labels.

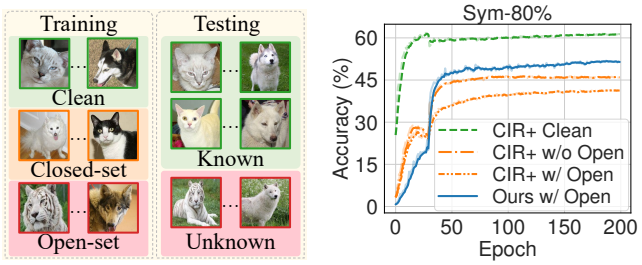

Figure 1: Illustration of the LOND setup and the effect of open-set noise. **Left:** In addition to closed-set noise, open-set noise is also present in the training and testing stage. **Right:** Open-set noise (w/ Open) significantly degrades performance on CIFAR80N with 80% symmetric noise.

Learning with Noisy Labels (LNL) has emerged as an effective learning paradigm for data with noisy labels (Song et al. 2023). Recently, a variety of methods have been proposed to train robust models directly on all data (Han et al. 2018a; Zhang and Sabuncu 2018; Yi et al. 2022), while others try to identify label noise and reduce its negative effects with carefully designed strategies (Ortego et al. 2021; Karim et al. 2022; Li et al. 2022, 2024; Yi et al. 2023). Most previous works generally focus on closed-set noise, which assumes that noisy labels still belong to known classes.

In more realistic scenarios, open-set noise from similar unknown classes may also occur in the training and inference stage (Wu et al. 2021), which poses significant challenges to LNL methods. This problem setup, called *Learning with Open-world Noisy Data (LOND)*, is illustrated in the left subplot of Figure 1. For instance, in a cat-dog classification problem, images of cats and dogs may be mislabeled as each other, resulting in closed-set noise. Due to similar features, open-set noise also occurs in the training and inference stage, where the similar unknown classes of lion and wolf may be labeled as cat and dog. These open-set noises significantly degrade the performance of the existing LNL model (Yi et al. 2023), as shown in the orange lines of the right subplot. Therefore, addressing the LOND problem re-

*Corresponding author.

quires the model to learn a classifier for known classes, and a detector for open-set noise, which is also known as an out-of-distribution (OOD) detector.

To address the LOND problem, we perform joint learning in dual representation space, i.e., the prototype space and the class-independent space, which is effective in dealing with closed-set and open-set noises (as shown in Figure 1). Specifically, to mitigate the model overfitting on the mixed noises, two networks are trained simultaneously, where the projection network uses learnable class prototypes to learn representations with shared semantics, and the One-Vs-All (OVA) network converts a multi-classification task into a multi-binary classification task by constructing a binary classifier for each class. Thus, the OVA network learns representations with unique semantics in the class-independent space, which reduces the inter-class competition in softmax and decouples the activations between clean and noisy labels. To enhance the detection of open-set samples, we introduce bi-level contrastive learning and consistency regularization in the two spaces. Subsequently, class-independent margin criteria are used to identify clean and noisy samples. For clean samples, the neighbor margin criterion aggregates OVA outputs from neighbor samples in the prototype space to ensure precise identification. On the other hand, open-set noise may have representations similar to closed-set samples, but its label can be considered as the negative class of all known classes. To measure the degree of open-set noise, the negative margin criterion is developed based on the closeness of negative probabilities in the class-independent space. For the remaining closed-set noise, we design a sample weight mechanism using the neighbor margin to measure the contribution to model learning. Our contributions are summarized as follows.

(1) We propose a novel joint learning framework with dual representation space using class-independent margin. Our method enhances performance on both classification for known classes and detection for open-set noise.

(2) To learn robust sample representation, bi-level contrastive learning and consistency regularization are introduced in dual representation space. Moreover, a margin-based sample identification mechanism is developed to effectively distinguish and use data with mixed noisy labels.

(3) We evaluate our method on multiple synthetic and real-world datasets, which demonstrates it achieves state-of-the-art (SOTA) performance.

Related Work

Learning with Noisy Labels

LNL methods aim to improve the model generalization from data with noisy labels and can be categorized into two main types based on how they use training data. (Huang et al. 2019). The first type directly uses all data to train robust models, including robust model architectures (Han et al. 2018a), loss functions (Zhang and Sabuncu 2018), and regularization methods (Yi et al. 2022). Since these methods treat the entire training data uniformly, their performance is limited at a high noise rate. The second type aims to identify noisy labels and reduce their negative effects using different criteria. The criteria commonly are based on the model output, such as cross-entropy loss (Li, Socher, and Hoi 2020), Jensen–Shannon divergence (Karim et al. 2022), regroup median loss (Li et al. 2024), logit margin (Zhang et al. 2024). In addition to the outputs of a softmax classifier, representation neighbor information (Li et al. 2022; Ortego et al. 2021) and the outputs of OVA network (Yi et al. 2023) are introduced for further performance improvement.

The above methods focus on addressing closed-set noise in the training set. Recent methods have studied the training set containing closed-set and open-set noises. For instance, to identify open-set noise, ILON (Wang et al. 2018) and Rog (Lee et al. 2019) employ neighbor density and class-wise distribution estimation, respectively. SNCF (Albert et al. 2022) performs spectral clustering to find representations of open-set noise. Jo-src (Yao et al. 2021) and EvidentialMix (Sachdeva et al. 2021) exploit the consistency and uncertainty in model outputs to detect the mixed noises. However, these methods neglect the open-set noise during the inference. To the best of our knowledge, only NGC (Wu et al. 2021) addresses the LOND problem using neighbor graphs. It obtains a softmax classifier and a prototype-based detector. However, the softmax classifier reduces the discriminability between clean and noisy labels due to the intersection of different class activations (Yi et al. 2023). Moreover, the detector shows weak adaptability to the LOND problem, since its class prototype is simply updated by the average of class representations. On the other hand, our method learns a good classifier and detector from joint learning in the prototype and class-independent spaces.

Out-of-distribution Detection

OOD detection in classification tasks aims to detect samples from unknown classes and can be divided into two categories: training-based and post-hoc methods (Yang et al. 2024). The former requires retraining with extra or synthesized outlier data (Hendrycks, Mazeika, and Dietterich 2019; Du et al. 2022) to enhance detection ability. In contrast, post-hoc methods can be directly performed on existing classifiers without retraining them, which maintains good classification performance (Yang et al. 2022). These includes methods based on softmax (Guo et al. 2017; Hendrycks and Gimpel 2017), logits (Hendrycks et al. 2022; Liu et al. 2020; Sun, Guo, and Li 2021; Song, Sebe, and Wang 2022), and distance (Bendale and Boult 2016; Lee et al. 2018; Sun et al. 2022). Their training datasets are usually assumed to have clean labels, but this is hard to meet in real-world scenarios. A recent study (Humblot-Renaux, Escalera, and Moeslund 2024) has analyzed the effect of label noise on post-hoc methods, but it still focuses on closed-set noise. Instead, our method learns to effectively detect samples with unknown classes from datasets containing closed-set and open-set noises.

Method

Problem Statement

Formally, consider an image classification task. The training set is denoted as $D_{train} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i is an

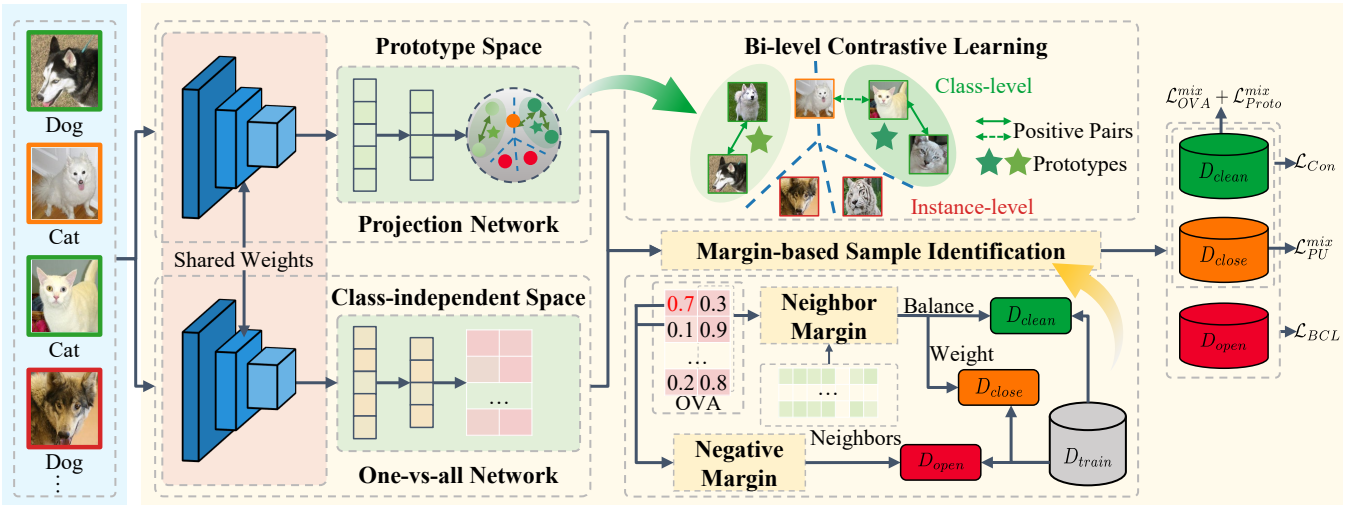

Figure 2: The overall framework of our proposed method. It uses projection and OVA networks to jointly learn in dual representation space, where bi-level contrastive learning and consistency regularization are introduced to enhance the detection of open-set noise. Then, class-independent margin criteria are used for sample identification. It uses the neighbor margin to select class-balanced clean samples D_{clean}, weighted closed-set noise D_{close}, and the negative margin to filter open-set noise D_{open}. Different losses are applied to these sample sets to obtain a classifier for known classes and a detector for open-set noise.

input image and $y_i \in \mathcal{C} = \{1, \ldots, C\}$ is the given label. Since the labels of D_{train} may not be correct, the true label of a sample x_i is denoted as y_i^*. A clean sample has a correct label, i.e., $y_i = y_i^*$. Given $y_i \neq y_i^*$, the noisy sample falls into one of two categories: $y_i^* \in \mathcal{C}$ (closed-set noise) or $y_i^* \notin \mathcal{C}$ (open-set noise). The goal is to train a robust model that performs well on the test set $D_{test} = \{(x_i, y_i^*)\}_{i=1}^{M}$. If $y_i^* \notin \mathcal{C}$ in the test set, the sample x_i still belongs to the label space of the open-set noise in D_{train} (Wu et al. 2021).

Overview

To address the LOND problem, we propose a joint learning method in dual representation space, as illustrated in Figure 2. After input images are processed by a feature extractor with shared weights, the projection and networks are jointly learned in the prototype and class-independent spaces. To improve the detection ability of the model for open-set noise, we introduce bi-level contrastive learning on all data in the projection network and consistency regularization on closed-set data in the OVA network. Then, two class-independent margin criteria are developed for effective sample identification. Class-balanced clean samples are selected by measuring the consistency of the neighbor label, called the neighbor margin criterion. The negative margin is developed using the probabilities of negative classes in OVA outputs to filter the open-set noise. For the remaining closed-set noise, sample weights are generated by the neighbor margin. Finally, these sample sets are applied with different losses to obtain a classifier for known classes and a detector for open-set noise.

Joint Learning in Dual Representation Space

Existing LNL methods generally handle noisy labels using predictions from a softmax classifier (Yi et al. 2023). However, the softmax has a competitive mechanism and generates dependent confidence scores among similar classes, making it sensitive to noisy labels. On the other hand, current class prototypes are updated directly by the average representations (Wu et al. 2021), which limits their adaptability to open-world scenarios. To address these problems, our method performs joint learning in dual representation space, i.e., the prototype and class-independent spaces, to effectively handle both closed-set and open-set noises. Specifically, the projection network constructs the prototype space based on learnable class prototypes by learning representations with shared semantics, thus improving its adaptability to open-set noise. The OVA network constructs the class-independent space by converting a multi-classification task into a multi-binary classification task, where each class is learned independently by a sub-classifier. This decouples the activations of different classes and learns representations with unique semantics, thus improving the discriminability between clean and noisy samples.

In the prototype space, the projection network learns class prototypes by gradient descent. Let the prototypes be the normalized vector set $P = \{P_c\}_{c=1}^{C}$. Denote the feature extractor as G, the projection head as H, and the sample representation as the embedding $z_i = H(G(x_i))$. The prototype loss for a sample x_i is defined as

$$\mathcal{L}_{Proto}(x_i, y_i) = \underbrace{-P_{y_i} \cdot z_i/\tau}_{\text{tightness}} + \underbrace{\log \sum_{c=1}^{C} \exp(P_c \cdot z_i/\tau)}_{\text{contrastive}}, \tag{1}$$

where τ denotes the temperature parameter and is simply set to 0.1 in all experiments. $\mathcal{L}_{Proto}$ considers intra-class tightness and inter-class separation to learn good representation.

In the class-independent space, let the OVA classifier be F_{OVA}. The output vector for the c-th class is $p^c_{OVA}(x_i) = F^c_{OVA}(G(x_i)) = [p^c_{OVA}(z=0|x_i), p^c_{OVA}(z=1|x_i)]$, and $p^c_{OVA}(z=0|x_i) + p^c_{OVA}(z=1|x_i) = 1$. Here, $z=0$ and $z=1$ indicate the sample does not belong to and belongs to the class, respectively. Then, the OVA loss for a sample x_i is defined as

$$\begin{aligned}\mathcal{L}_{OVA}(x_i, y_i) = &- \log p^{y_i}_{OVA}(z=1|x_i) \\ &- \sum_{j=1, j\neq y_i}^{C} \log p^{y_i}_{OVA}(z=0|x_i).\end{aligned} \tag{2}$$

To further improve robustness to mixed noises, loss mixup is adopted, which provides better performance of OOD detection compared to label mixup (Pinto et al. 2022). The convex combination of x_a and x_b is defined as $x^{mix}_i = \lambda x_a + (1-\lambda)x_b$, where $\lambda \in [0,1] \sim Beta(\alpha, \alpha)$. The loss mixup is defined as

$$\mathcal{L}^{mix}_i = \lambda \mathcal{L}_i(x^{mix}_i, y_a) + (1-\lambda)\mathcal{L}_i(x^{mix}_i, y_b). \tag{3}$$

Open-set Robust Representation Learning

Robust representation learning for noisy labels is generally achieved by contrastive learning (Yi et al. 2022) and consistency regularization (Li, Socher, and Hoi 2020). However, existing methods often ignore handling open-set noise. To alleviate this problem, we introduce robust representation learning for open-set noise, including bi-level contrastive learning in the prototype space and consistency regularization in the class-independent space. This enhances the ability of the model to detect open-set noise.

In the prototype space, contrastive learning is performed at the class and instance levels for the identified closed-set and open-set samples, respectively. On two-view data with strong and weak augmentations, the loss of the bi-level contrastive learning for a single sample x_i is defined as follows:

$$\mathcal{L}_{BCL}(z_i, y_i) = \frac{1}{1+|P(i)|}\mathcal{L}_{BCL,i}, \tag{4}$$

where $P(i)$ denotes the index set of views of other samples with shared labels ($y_i = y_j$) in a minibatch data B, and $|\cdot|$ indicates the cardinality of the set. $\mathcal{L}_{BCL,i}$ is defined as

$$\begin{aligned}\mathcal{L}_{BCL,i} = &- \log \frac{exp(z_i \cdot z^*_i/\tau)}{\sum^{2|B|}_{r=1, r\neq i} exp(z_i \cdot z_r/\tau)} \\ &- \sum_{j\in P(i)} \log \frac{w(x_i)\cdot w(x_j)\cdot exp(z_i \cdot z_j/\tau)}{\sum^{2|B|}_{r=1, r\neq i} exp(z_i \cdot z_r/\tau)},\end{aligned} \tag{5}$$

where z^*_i denotes another view of x_i, and $w(\cdot)$ is the sample weight derived from the neighbor margin.

In the class-independent space, the loss of consistency regularization with respect to the output of the OVA network is defined as

$$\begin{aligned}\mathcal{L}_{Con}(x_i, y_i) = \sum_{c=1}^{C} \sum_{j\in(0,1)} &||p^c_{OVA}(z=j|t_s(x_i)) \\ &- p^c_{OVA}(z=j|t_w(x_i))||^2_2,\end{aligned} \tag{6}$$

where t_s and t_w denote strong and weak augmentations.

Margin-based Sample Identification

Sample identification in LNL aims to distinguish clean and noisy samples based on estimated quality. In general, existing methods design estimation criteria based on the outputs of a softmax classifier, but they are hard to reduce inter-class competition in softmax (Yi et al. 2023). Although the logits of a softmax classifier have been used to design margin criteria (Zhang et al. 2024), they neglect open-set noise. To handle the mixed noises, we design the class-independent margin criteria in dual representation space, including the neighbor margin and the negative margin.

A clean sample has higher consistency between its given label and the neighbor label that is aggregated from the neighbor samples (Ortego et al. 2021). To obtain the neighbor label, previous methods aggregate the given labels or softmax probabilities of neighbors, which may not be effective at a high noise rate. Instead, we aggregate the OVA probability outputs in the class-independent space based on the representation neighbors. The probability of class c of the neighbor label of sample x_i is defined as follows:

$$q^c_{Neigh}(x_i) = \sum_{j=1}^{k} w^{Neigh}_{ij} p^c_{OVA}(x_i), \tag{7}$$

where k indicates the number of nearest neighbors, and w^{Neigh}_{ij} denotes the normalized weight based on the distance between sample x_i and neighbor sample x_j, which is defined as $w^{Neigh}_{ij} = exp(z_i \cdot z_j/\tau) / \sum^k_{j=1} exp(z_i \cdot z_j/\tau)$.

To measure the consistency between the given label and the neighbor label, the neighbor margin is defined as

$$\begin{aligned}M_{Neigh}(x_i) = &\, q^{y_i}_{Neigh}(z=1|x_i) \\ &- \frac{1}{K}\sum_{j=1, j\neq y_i}^{K} q^j_{Neigh}(z=1|x_i),\end{aligned} \tag{8}$$

where $\sum^K_{j=1, j\neq y_i} q^j_{Neigh}(z=1|x_i)$ is the sum of the top-k probabilities. The larger the value of M^i_{Neigh} (x is omitted), the more likely it is to be the clean sample. To select clean samples, a naive method is to use a fixed threshold on the margin. However, since the margin is larger for easy classes and smaller for hard classes, this method can result in a class-imbalanced set of clean samples. To ensure class balance, the clean sample set for each class c is defined as

$$D^c_{clean} = \{(x_i, y_i) : M_{Neigh}(x_i) \leq \gamma_c\}, \tag{9}$$

where γ_c is a class-wise dynamic threshold determined by the α_{ID}-quantile of the class consistency degree that is the number of neighbor labels equal to given labels.

Open-set noise may have similar features to closed-set samples with the same given label. Their representations with shared semantics make open-set noise difficult to identify. In fact, an open-set sample can be considered as not belonging to any known classes. In other words, it belongs to the negative class of all known classes. The OVA network estimates the probability that a sample does not belong to each class, called the negative probability. If the negative probability of a given label is close to that of other labels,

the sample is identified as an open-set sample, belonging to the negative class of all known classes. Thus, the negative margin is defined as

$$M_{Neg}(x_i) = |p^{y_i}_{OVA}(z=0|x_i) - \max_{j \neq y_i} p^{j}_{OVA}(z=0|x_i)|. \tag{10}$$

The smaller the value of M_{Neg}, the more likely it is an open-set noise sample. These samples are generally treated as a novel class and can be selected using a single threshold. Thus, the set of open-set noise is

$$\begin{aligned} D_{open} = \{(x_i, y_i) : & M_{Neg}(x_i) \leq \gamma_{Neg}, \\ & (x_i, y_i) \notin D_{clean}\}, \end{aligned} \tag{11}$$

where γ_{Neg} is determined by the first α_{OOD} percentage of M_{Neg} sorted in ascending order.

To effectively use the rest closed-set noise, the sample weights are defined as

$$w(x_i) = \begin{cases} 1, & x_i \in D_{clean} \\ 0, & x_i \in D_{open} \\ (M^i_{Neigh}+1)/(M^{\max}_{Neigh}+1), & x_i \in D_{close} \end{cases}, \tag{12}$$

where $M^{\max}_{Neigh}$ is the maximum margin and $D_{close} = D_{train} - D_{clean} - D_{open}$. In addition to bi-level contrastive learning, these sample weights are also used in the pseudo-label learning of class prototypes. After mixing closed-set samples as in (Li, Socher, and Hoi 2020), the pseudo-label loss for prototypes of a closed-set sample x_i is defined as

$$\begin{aligned} \mathcal{L}_{PU}(x_i, y_i) &= ||p_i - Sharpen(\bar{y}_i, w(x_i), T)||^2_2, \\ Sharpen(\bar{y}_i, w(x_i), T) &= \left\{ \frac{\bar{y}^{w(x_i)/T}_{i,c}}{\sum_{j=1}^{C} \bar{y}^{w(x_i)/T}_{i,j}} \right\}_{c=1}^{C}, \end{aligned} \tag{13}$$

where p_i is the softmax probability of the prototype logits $\{P_c \cdot z_i/\tau\}_{c=1}^{C}$, $\bar{y}_i$ is the average probability of the sample with strong and weak augmentations, and T is the sharpening parameter that is set to 0.5 in this study.

Training and Inference

On the three identified sample subsets, the total loss of our method is

$$\begin{aligned} \mathcal{L} = & \sum_{D_{clean}} \mathcal{L}^{mix}_{OVA} + \sum_{D_{clean}} \mathcal{L}^{mix}_{Proto} + \sum_{D_{close}} \mathcal{L}^{mix}_{PU} \\ & + \lambda_{Con} \sum_{D_{clean} \cup D_{close}} \mathcal{L}_{Con} + \lambda_{BCL} \sum_{D_{train}} \mathcal{L}_{BCL}. \end{aligned} \tag{14}$$

During inference, the OVA output is used to evaluate the classification performance for known classes. For the OOD detection performance, the OVA output and the prototypes are combined to calculate the following OOD score:

$$s(x_i) = p^{y_{proto}}_{OVA}(z=0|x_i), \tag{15}$$

where y_{proto} is the label predicted by the class prototypes.

Experiments

In this section, our method is compared with other SOTA methods on multiple datasets. Specifically, the effectiveness of our method is evaluated on closed-world, open-world, and real-world noisy data. The ablation study is also conducted to verify the importance of different modules of our method.

Experimental Setup

Datasets. The effectiveness of our method is evaluated on the CIFAR80N, CIFAR100N, Web-Aircraft, Web-Car, and Web-Bird datasets. Specifically, the CIFAR100 dataset (Krizhevsky 2009), containing 50,000 training images and 10,000 test images, is used as the base dataset. Synthetic noises are added to this dataset to generate the CIFAR80N and CIFAR100N datasets, following the settings in (Yao et al. 2021). They contain both symmetric and asymmetric types of noise with specified noise rates. In addition, the last 20 classes of CIFAR100 are added to the test set of CIFAR80N to validate the OOD detection performance of our method. To further test our method under more challenging scenarios, web datasets are used, including Web-Aircraft, Web-Car, and Web-Bird (Sun et al. 2021). These datasets are collected via image search engines, inevitably resulting in unknown noise rates and complex noise types.

Implementation Details. For experiments on the CIFAR datasets, a seven-layer CNN (Yao et al. 2021) is used as the backbone network. It is trained using SGD with a momentum of 0.9, a weight decay of 0.0005, and an initial learning rate of 0.05 adjusted by cosine annealing. Both the batch size and the projection dimension are set to 128. Set $\alpha = 1$ in the Beta distribution, $K = 3$ for symmetric noise, and $K = 1$ for asymmetric noise. The network is trained for 300 epochs, including a 50-epoch warm-up phase. For OOD detection methods, the settings are consistent with those in (Yang et al. 2022). For web datasets, ResNet50 pre-trained on ImageNet is adopted and trained using SGD consistent with the CIFAR experiments. This training uses a batch size of 64 and an initial learning rate of 0.005. The ResNet50 is trained for 120 epochs with a 10-epoch warm-up phase, where the prototype loss is added with a weight of 10 to fully use the pre-trained knowledge. Moreover, set $\alpha = 0.5$, $K = 1$, and $\alpha_{ID} = 0.5$.

Baselines. On CIFAR100N and CIFAR80N, our method is compared with recent SOTA methods, including Co-teaching (Han et al. 2018b), Co-teaching+ (Yu et al. 2019), DivideMix (Li, Socher, and Hoi 2020), NGC, NCE (Li et al. 2022), CIR+ (Yi et al. 2023), NPN-hard (Sheng et al. 2024), and SED (Sheng et al. 2025). The performance of the model trained using only cross-entropy (denoted as standard) is also shown. Most results of these methods are from SED, and † denotes the re-implemented performance using open-source code (except for NPN-hard on CIFAR100N). On the web datasets, we also compare the following SOTA methods: Jo-SRC, UNICON (Karim et al. 2022), and DISC (Li et al. 2023). For OOD detection, SED is combined with multiple post-hoc methods. These include distance-based methods like MDS (Lee et al. 2018), KNN (Sun et al. 2022), and OpenMax (Bendale and Boult 2016), and softmax-based methods like MSP (Hendrycks and Gimpel 2017) and Temp-Scaling (Guo et al. 2017), logit-based methods like MLS (Hendrycks et al. 2022), EBO (Liu et al. 2020), REACT (Sun, Guo, and Li 2021), and RankFeat (Song, Sebe, and Wang 2022). Classification performance is measured by the top-1 accuracy metric and OOD detection is verified using the AUROC metric.

Methods	Publication	CIFAR100N			CIFAR80N		
		Sym-20%	Sym-80%	Asym-40%	Sym-20%	Sym-80%	Asym-40%
Standard	-	35.50	3.84	28.43	29.37	4.20	22.25
Co-teaching	NeurIPS 2018	56.21	22.83	37.26	60.38	16.59	42.42
Co-teaching+	ICML 2019	52.87	18.55	38.78	53.97	12.29	43.01
DivideMix	ICLR 2020	57.76	28.98	43.75	57.47	21.18	37.47
NGC†	ICCV 2021	60.95	40.10	45.50	64.67	37.12	47.83
NCE	ECCV 2022	54.58	35.23	49.90	58.53	39.34	56.40
CIR+†	AAAI 2023	57.73	44.95	52.67	61.11	45.75	56.47
NPN-hard†	AAAI 2024	65.27	36.88	60.11	66.07	35.38	64.09
SED	ECCV 2024	66.50	38.15	58.29	69.10	42.57	60.87
Ours	-	**67.11**	**48.33**	**65.22**	**69.61**	**49.30**	**67.27**

Table 1: Average accuracy (%) on closed-world noisy data (CIFAR100N) and open-world noisy data (CIFAR80N) over the last ten epochs, where "Sym" and "Asym" denote the symmetric and asymmetric noise, respectively.

CIFAR80N	MDS	KNN	OpenMax	MSP	TempScaling	MLS	EBO	REACT	RankFeat	NGC	Ours
Sym-20%	48.02	44.47	60.93	63.92	64.43	63.73	63.45	62.90	56.54	67.97	**75.97**
Sym-80%	51.06	48.95	57.17	60.23	61.44	61.83	61.72	61.01	55.22	59.17	**63.82**
Asym-40%	47.21	45.97	60.61	61.48	61.97	61.95	61.88	61.44	56.41	67.64	**73.50**
Avg.	48.76	46.46	59.57	61.88	62.61	62.50	62.35	61.78	56.06	64.93	**70.10**

Table 2: AUROC (%) comparison with SED combined with SOTA post-hoc methods and NGC, where "Avg." denotes the average performance on three cases.

Comparisons with State-of-the-art Methods

Evaluation on Closed-world and Open-world Noisy Data The results validated on the closed-world noisy data CIFAR100N are shown in Table 1. It is clearly evident that our method outperforms other methods. In the most challenging case (i.e., Sym-80%), our method improves by at least 3.38%. This indicates that our method can handle closed-set noise effectively.

On the open-world noisy data CIFAR80N, our method also outperforms other methods in both classification and OOD detection performance, as evidenced by Tables 1 and 2. For the classification performance in Table 1, our method still maintains an accuracy of 49.30% at Sym-80%. Toward more realistic asymmetric noise, our method improves the accuracy by at least 3.18%. These results show that our method achieves superior classification performance even in the presence of high noise rates and complex noise types. For the OOD detection performance in Table 2, our method clearly achieves the best AUROC scores. Compared to other methods, our method increases at least 6.17% on average. In all, our method can effectively identify open-set noise while robustly classifying known classes.

Evaluation on Real-world Noisy Data. The comparison performance on real-world datasets is shown in Table 3. The results confirm that our method outperforms other SOTA methods on average. Specifically, our method achieves an accuracy of 89.92% on Web-Aircraft, 80.95% on Web-Bird, and 89.45% on Web-Car, with an average improvement of at least 0.94% over other methods. These results demonstrate that our method effectively handles real-world data with unknown noise rates and types. This can be attributed to the effective joint learning in dual representation space and the strategy of sample identification, which significantly reduces the negative effects of the mixed noises.

Ablation Study

Effects of Different Modules. An ablation study is conducted to investigate the contributions of different modules and losses used in the proposed method. Table 4 shows the classification and OOD detection performance at Sym-80% and Asym-40% on CIFAR80N. The average performance is also listed, which balances the classification performance and the OOD detection performance.

It is evident that each module enhances the average metric. Our method employs joint learning in the prototype and class-independent spaces to mitigate mixed noise overfitting. It is worth noting that our warm-up model alone obtains 33.95% accuracy at Sym-80% and 55.66% accuracy at Asym-40%, which outperforms some SOTA methods in Table 1. In addition to warm-up, we use $\mathcal{L}_{OVA}$ and $\mathcal{L}_{Proto}$ on the clean samples as a baseline, where the samples are identified by the neighbor margin. This enhances the average performance by 8.47% compared to (1). It suggests that learning in dual representation space could effectively handle closed-set and open-set noises. Moreover, $\mathcal{L}_{PU}$ is used for pseudo-label learning in the prototype space. This additional learning further enhances the performance on asymmetric noise. For example, at Asym-40%, the model with the loss of $\mathcal{L}_{PU}$ achieves a 1.00% improvement compared to (2). This may be because the neighbor margin-weighted temperature

Methods	Publication	Backbone	Web-Aircraft	Web-Bird	Web-Car	Avg.
Standard	-	ResNet50	60.80	64.40	60.60	61.93
Co-teaching	NeurIPS 2018	ResNet50	79.54	76.68	84.95	80.39
Co-teaching+	ICML 2019	ResNet50	74.80	70.12	76.77	73.90
DivideMix	ICLR 2020	ResNet50	82.48	74.40	84.27	80.38
NGC	ICCV 2021	ResNet50	78.64	75.37	82.48	78.83
Jo-SRC	CVPR 2021	ResNet50	82.73	81.22	88.13	84.03
UNICON	CVPR 2022	ResNet50	85.18	81.20	88.15	84.84
NCE	ECCV 2022	ResNet50	84.94	80.22	86.38	83.85
DISC	CVPR 2023	ResNet50	85.27	81.08	88.31	84.89
NPN-hard	AAAI 2024	ResNet50	86.02	80.91	88.26	85.06
SED	ECCV 2024	ResNet50	86.62	**82.00**	88.88	85.83
Ours	-	ResNet50	**89.92**	80.95	**89.45**	**86.77**

Table 3: Accuracy (%) comparison with SOTA methods on Web-Aircraft, Web-Bird, and Web-Car, where the results of existing methods are mainly copied from (Sheng et al. 2025).

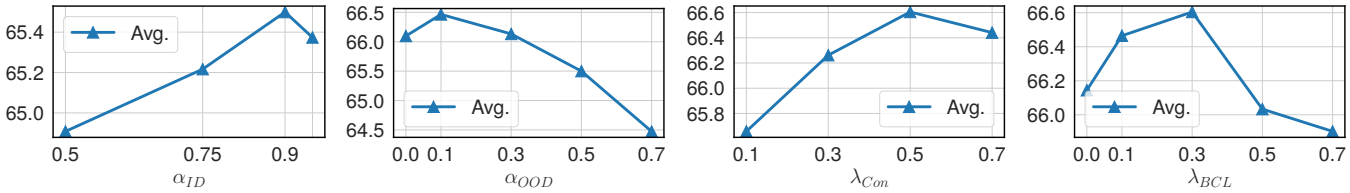

Figure 3: The sensitivity of hyper-parameters α_{ID}, α_{OOD}, λ_{Con}, and λ_{BCL}, where "Avg." denotes the average of accuracy and AUROC of three cases (Sym-20%, Sym-80%, and Asym-40%) on CIFAR80N.

Methods/Noise	Sym-80%			Asym-40%		
	ACC	AUROC	Avg.	ACC	AUROC	Avg.
(1).Warmup	33.95	56.11	45.03	55.66	67.35	61.50
(2).Baseline	46.95	62.68	54.82	66.15	71.16	68.65
(3).(2)+$\mathcal{L}_{PU}$	46.95	62.51	54.73	67.00	72.30	69.65
(4).(3)+$\mathcal{L}_{Con}$	47.65	63.03	55.34	66.40	72.92	69.66
(5).(4)+$\mathcal{L}_{BCL}$	**49.30**	**63.82**	**56.56**	**67.27**	**73.50**	**70.39**

Table 4: Ablation study of our method on CIFAR80N with 80% symmetric and 40% asymmetric noise rates.

smooths the noise prediction and reduces noise overfitting. In addition, we apply $\mathcal{L}_{Con}$ on the closed-set samples in the class-independent space. Adding $\mathcal{L}_{Con}$ improves the average AUROC by 0.57% on both types of noises. This may be attributed to that $\mathcal{L}_{Con}$ requires the model to predict consistently on closed-set samples with different perturbations. On all training data, we further apply bi-level contrastive learning. Our proposed model improves the accuracy by 1.65% at Sym-80%. The reason for this may be that $\mathcal{L}_{BCL}$ can construct high-quality sample pairs from the instance and class levels for better representation learning.

Sensitivity Analysis of Hyper-parameters. Our method introduces the parameter of K in the neighbor margin, α_{ID} and α_{OOD} in the sample identification, and λ_{Con} and λ_{BCL} in the total loss function. Detailed analysis of K is provided in the supplementary material, while the results for the other four hyper-parameters are shown in Figure 3. The performance is measured by the average metric of three cases (Sym-20%, Asym-40%, Sym-80%) on CIFAR80N. The performance is enhanced when increasing the parameter of α_{ID} since more clear samples are selected. But after reaching the value of 0.9, the performance is degraded dramatically. Conversely, the performance is decreased with a larger value of α_{OOD} because more potentially useful samples of known classes are filtered out. For the trade-off parameters of losses, a moderate value of λ_{Con} is preferred for balancing its weight with $\mathcal{L}_{OVA}$ to keep the classification performance. Similarly, to achieve this balance, a smaller λ_{BCL} should be taken. The method achieves the best balance between classification and OOD detection performance when $\alpha_{ID} = 0.9$, $\alpha_{OOD} = 0.1$, $\lambda_{Con} = 0.5$, and $\lambda_{BCL} = 0.3$.

Conclusion

In this study, we propose the joint learning method with dual representation space to address the LOND problem. Specifically, projection and OVA networks are simultaneously trained to reduce the model overfitting on closed-set and open-set noises. To enhance the detection of open-set noise, our method introduces bi-level contrastive learning and consistency regularization. Moreover, class-independent margin criteria are designed to identify clean samples, closed-set noise, and open-set noise, and jointly minimize different losses on these sample subsets to benefit from the memorization effects. Our method outperforms other SOTA methods on the open-world noisy data.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 62476171, 62476172, 62076164, and 62206122), the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2024A1515011367), the Guangdong Provincial Key Laboratory (Grant No. 2023B1212060076), and the Shenzhen Institute of Artificial Intelligence and Robotics for Society.

References

Albert, P.; Arazo, E.; O'Connor, N. E.; and McGuinness, K. 2022. Embedding Contrastive Unsupervised Features to Cluster In- And Out-of-Distribution Noise in Corrupted Image Datasets. In *Computer Vision – ECCV 2022*, 402–419.

Bendale, A.; and Boult, T. E. 2016. Towards Open Set Deep Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1563–1572.

Du, X.; Wang, Z.; Cai, M.; and Li, Y. 2022. VOS: Learning What You Don't Know by Virtual Outlier Synthesis. In *International Conference on Learning Representations*.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, 1321–1330.

Han, B.; Yao, J.; Niu, G.; Zhou, M.; Tsang, I.; Zhang, Y.; and Sugiyama, M. 2018a. Masking: A New Perspective of Noisy Supervision. In *Advances in Neural Information Processing Systems*, 5836–5846.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018b. Co-Teaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels. In *Advances in Neural Information Processing Systems*, 8527–8537.

Hendrycks, D.; Basart, S.; Mazeika, M.; Zou, A.; Kwon, J.; Mostajabi, M.; Steinhardt, J.; and Song, D. 2022. Scaling Out-of-Distribution Detection for Real-World Settings. In *Proceedings of the 39th International Conference on Machine Learning*, 8759–8773.

Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *International Conference on Learning Representations*.

Hendrycks, D.; Mazeika, M.; and Dietterich, T. 2019. Deep Anomaly Detection with Outlier Exposure. In *International Conference on Learning Representations*.

Huang, J.; Qu, L.; Jia, R.; and Zhao, B. 2019. O2U-Net: A Simple Noisy Label Detection Approach for Deep Neural Networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3326–3334.

Humblot-Renaux, G.; Escalera, S.; and Moeslund, T. B. 2024. A Noisy Elephant in the Room: Is Your Out-of-Distribution Detector Robust to Label Noise? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22626–22636.

Karim, N.; Rizve, M. N.; Rahnavard, N.; Mian, A.; and Shah, M. 2022. UNICON: Combating Label Noise Through Uniform Selection and Contrastive Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9676–9686.

Krizhevsky, A. 2009. Learning multiple layers of features from tiny images. Technical report, University of Toronto.

Lee, K.; Lee, K.; Lee, H.; and Shin, J. 2018. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *Advances in Neural Information Processing Systems*, 7167–7177.

Lee, K.; Yun, S.; Lee, K.; Lee, H.; Li, B.; and Shin, J. 2019. Robust Inference via Generative Classifiers for Handling Noisy Labels. In *Proceedings of the 36th International Conference on Machine Learning*, 3763–3772.

Li, F.; Li, K.; Tian, J.; and Zhou, J. 2024. Regroup Median Loss for Combating Label Noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 13474–13482.

Li, J.; Li, G.; Liu, F.; and Yu, Y. 2022. Neighborhood Collective Estimation for Noisy Label Identification and Correction. In *Computer Vision – ECCV 2022*, 128–145.

Li, J.; Socher, R.; and Hoi, S. C. H. 2020. DivideMix: Learning with Noisy Labels as Semi-Supervised Learning. In *International Conference on Learning Representations*.

Li, Y.; Han, H.; Shan, S.; and Chen, X. 2023. 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 (CVPR)*, 24070–24079.

Liu, W.; Wang, X.; Owens, J.; and Li, Y. 2020. Energy-Based Out-of-distribution Detection. In *Advances in Neural Information Processing Systems*, 21464–21475.

Ortego, D.; Arazo, E.; Albert, P.; O'Connor, N. E.; and McGuinness, K. 2021. Multi-Objective Interpolation Training for Robustness To Label Noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6606–6615.

Pinto, F.; Yang, H.; Lim, S. N.; Torr, P.; and Dokania, P. 2022. Using Mixup as a Regularizer Can Surprisingly Improve Accuracy & Out-of-Distribution Robustness. In *Advances in Neural Information Processing Systems*, 14608–14622.

Sachdeva, R.; Cordeiro, F. R.; Belagiannis, V.; Reid, I.; and Carneiro, G. 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 (WACV)*, 3607–3615.

Sheng, M.; Sun, Z.; Cai, Z.; Chen, T.; Zhou, Y.; and Yao, Y. 2024. Adaptive integration of partial label learning and negative learning for enhanced noisy label learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 4820–4828.

Sheng, M.; Sun, Z.; Chen, T.; Pang, S.; Wang, Y.; and Yao, Y. 2025. Foster Adaptivity and Balance in Learning with Noisy Labels. In *Computer Vision – ECCV 2024*, 217–235.

Song, H.; Kim, M.; Park, D.; Shin, Y.; and Lee, J.-G. 2023. Learning From Noisy Labels With Deep Neural Networks: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11): 8135–8153.

Song, Y.; Sebe, N.; and Wang, W. 2022. RankFeat: Rank-1 Feature Removal for Out-of-distribution Detection. In *Advances in Neural Information Processing Systems*, 17885–17898.

Sun, Y.; Guo, C.; and Li, Y. 2021. ReAct: Out-of-distribution Detection With Rectified Activations. In *Advances in Neural Information Processing Systems*, 144–157.

Sun, Y.; Ming, Y.; Zhu, X.; and Li, Y. 2022. Out-of-Distribution Detection with Deep Nearest Neighbors. In *Proceedings of the 39th International Conference on Machine Learning*, 20827–20840.

Sun, Z.; Yao, Y.; Wei, X.-S.; Zhang, Y.; Shen, F.; Wu, J.; Zhang, J.; and Shen, H. T. 2021. Webly Supervised Fine-Grained Recognition: Benchmark Datasets and an Approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10602–10611.

Wang, Y.; Liu, W.; Ma, X.; Bailey, J.; Zha, H.; Song, L.; and Xia, S.-T. 2018. Iterative Learning with Open-Set Noisy Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8688–8696.

Wu, Z.-F.; Wei, T.; Jiang, J.; Mao, C.; Tang, M.; and Li, Y.-F. 2021. NGC: A Unified Framework for Learning With Open-World Noisy Data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 62–71.

Yang, J.; Wang, P.; Zou, D.; Zhou, Z.; Ding, K.; Peng, W.; Wang, H.; Chen, G.; Li, B.; Sun, Y.; Du, X.; Zhou, K.; Zhang, W.; Hendrycks, D.; Li, Y.; and Liu, Z. 2022. OpenOOD: Benchmarking Generalized Out-of-Distribution Detection. In *Advances in Neural Information Processing Systems*, 32598–32611.

Yang, J.; Zhou, K.; Li, Y.; and Liu, Z. 2024. Generalized Out-of-Distribution Detection: A Survey. *International Journal of Computer Vision*, 132(12): 5635–5662.

Yao, Y.; Sun, Z.; Zhang, C.; Shen, F.; Wu, Q.; Zhang, J.; and Tang, Z. 2021. Jo-Src: A Contrastive Approach for Combating Noisy Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5192–5201.

Yi, L.; Liu, S.; She, Q.; McLeod, A. I.; and Wang, B. 2022. On Learning Contrastive Representations for Learning With Noisy Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16682–16691.

Yi, R.; Guan, D.; Huang, Y.; and Lu, S. 2023. Class-Independent Regularization for Learning with Noisy Labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3276–3284.

Yu, X.; Han, B.; Yao, J.; Niu, G.; Tsang, I.; and Sugiyama, M. 2019. How Does Disagreement Help Generalization against Label Corruption? In *Proceedings of the 36th International Conference on Machine Learning*, 7164–7173.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2017. Understanding Deep Learning Requires Rethinking Generalization. In *International Conference on Learning Representations*.

Zhang, S.; Li, Y.; Wang, Z.; Li, J.; and Liu, C. 2024. Learning with Noisy Labels Using Hyperspherical Margin Weighting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 16848–16856.

Zhang, Z.; and Sabuncu, M. 2018. Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. In *Advances in Neural Information Processing Systems*, 8778–8788.