Enhanced Sample Selection with Confidence Tracking: Identifying Correctly Labeled yet Hard-to-Learn Samples in Noisy Data

Weiran Pan[1, 2], **Wei Wei**[1, 2*], **Feida Zhu**[3], **Yong Deng**[4]
[1] School of Computer Science and Technology, Huazhong University of Science and Technology, China
[2] Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL), China
[3] School of Computing and Information Systems, Singapore Management University, Singapore
[4] State Grid Fujian Electric Power Co.
{panwr, weiw}@hust.edu.cn, fdzhu@smu.edu.sg, hhdyyong@qq.com

Abstract

We propose a novel sample selection method for image classification in the presence of noisy labels. Existing methods typically consider small-loss samples as correctly labeled. However, some correctly labeled samples are inherently difficult for the model to learn and can exhibit high loss similar to mislabeled samples in the early stages of training. Consequently, setting a threshold on per-sample loss to select correct labels results in a trade-off between precision and recall in sample selection: a lower threshold may miss many correctly labeled hard-to-learn samples (low recall), while a higher threshold may include many mislabeled samples (low precision). To address this issue, our goal is to accurately distinguish correctly labeled yet hard-to-learn samples from mislabeled ones, thus alleviating the trade-off dilemma. We achieve this by considering the trends in model prediction confidence rather than relying solely on loss values. Empirical observations show that only for correctly labeled samples, the model's prediction confidence for the annotated labels typically increases faster than for any other classes. Based on this insight, we propose tracking the confidence gaps between the annotated labels and other classes during training and evaluating their trends using the Mann-Kendall Test. A sample is considered potentially correctly labeled if all its confidence gaps tend to increase. Our method functions as a plug-and-play component that can be seamlessly integrated into existing sample selection techniques. Experiments on several standard benchmarks and real-world datasets demonstrate that our method enhances the performance of existing methods for learning with noisy labels.

Code — https://github.com/Aliinton/ConfidenceTracking

1 Introduction

The remarkable success of deep learning methods in classification tasks can largely be attributed to high-quality datasets. However, collecting such datasets through manual labeling can be both time-consuming and expensive in many applications. Acquiring data via online queries (Li et al. 2017) or crowdsourcing (Xiao et al. 2015) can construct large-scale datasets at a lower cost but inevitably introduces noise labels. Existing research (Zhang et al. 2021) has shown that deep neural networks can easily fit noisy data, resulting in poor generalization. Therefore, developing algorithms robust to noisy labels is of great practical importance (Natarajan et al. 2013).

Sample selection methods aim to identify correct labels from noisy data, which is increasingly crucial for current deep learning models to effectively learn from noisy labels. There is a general consensus that the small-loss criterion is an effective approach, which assumes samples with small losses are more likely to have correct labels. In this context, the Co-teaching family (Han et al. 2018; Yu et al. 2019; Wei et al. 2020; Xia et al. 2022) and other state-of-the-art methods (Li, Socher, and Hoi 2020; Li et al. 2023) have been proposed. Typically, these methods select samples with losses below a threshold for training. So a higher threshold introduces more incorrect labels while a lower threshold excludes more correct labels, creating a trade-off dilemma between precision and recall in sample selection. To alleviate this issue, one possible solution is establishing another sample selection criterion that can distinguish correct labels from incorrect ones in high-loss data. Then we can combine it with the small-loss criterion to improve recall while maintaining precision in sample selection.

To achieve this, we propose considering the changing trends in model predictions to identify correct labels rather than relying solely on loss values. This is motivated by our empirical observations that although some correctly labeled samples may obtain similar loss values to mislabeled ones, their training dynamics are still distinguishable. As shown in Figure 1, some correctly labeled data can be hard to fit by the model and exhibit similar high loss to mislabeled data in the early training stage. It is difficult to distinguish them by setting a threshold on loss values. But we also observe that, only for the correctly labeled samples, the model's prediction confidence for annotated labels tends to rise more quickly than for other classes. For instance, with correctly labeled dog images, the model's prediction confidence for the "dog" class (*i.e.,* the posterior probability of the image belonging to the "dog" class predicted by the model) increases more rapidly than for any other class. Conversely, for cat images mislabeled as "dog", the model's prediction confidence for the "cat" class rises slightly faster than for the "dog" class. This observation suggests it's possible to identify correctly labeled samples by considering training

*Corresponding author.
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

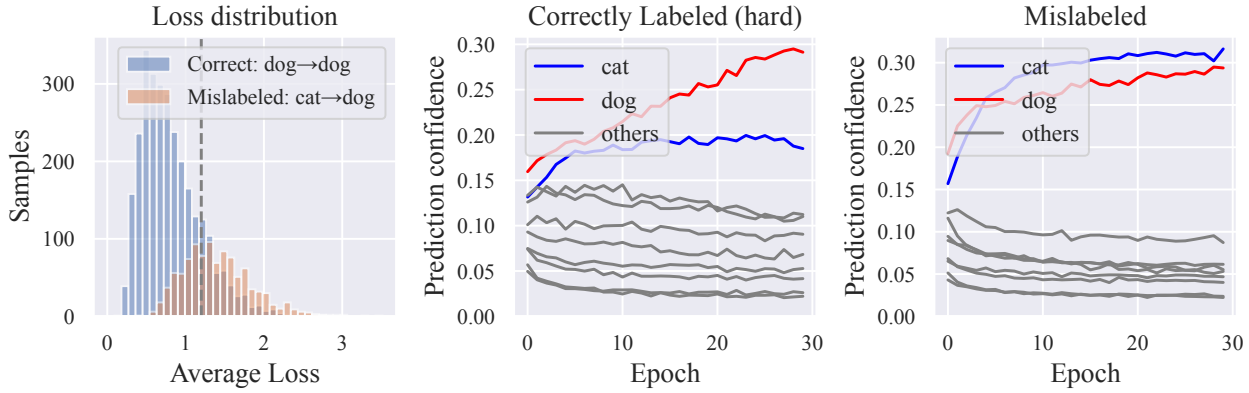

Figure 1: Illustration of Confidence Tracking. We train a PreActResNet-18 model using cross-entropy loss and an SGD optimizer on CIFAR-10N-Worst (CIFAR-10 dataset with human-annotated real-world noisy labels, its noise rate is 40.21%). The left graph presents the average per-sample loss distribution in the first 30 epochs. We regard samples with an average loss greater than 1.2 (indicated by the dotted vertical line) as hard-to-learn ones and show the model prediction confidence trajectories on hard-to-learn dogs' images (middle graph) and mislabeled cats' images (right graph).

dynamics even when they are indistinguishable from mislabeled samples in terms of loss values.

Motivated by this finding, we propose a novel sample selection criterion that selects correct labels by monitoring how model predictions change during training, dubbed Confidence Tracking (CT). Specifically, we track confidence gaps in model predictions between annotated labels and other classes during training. If all confidence gaps of a sample tend to increase, we regard it as a potentially correctly labeled sample. *Unlike the small loss criterion which mainly considers the loss values, our method focuses on the changing trend in model predictions, allowing us to distinguish correctly labeled yet hard-to-learn samples from mislabeled ones within high-loss data.* In practice, our method functions as a plug-and-play component that can be combined with popular sample selection methods (Arazo et al. 2019; Kim et al. 2021; Pleiss et al. 2020; Li et al. 2023) to enhance their performance.

Our method is related to AUM (Pleiss et al. 2020) which also considers confidence gaps but selects samples with relatively large average logit margins as correctly labeled. Similar to the small-loss criterion, setting a threshold on average logit margins to select correct labels still faces the trade-off dilemma between precision and recall. Our experiments in Section 4 demonstrate that CT accurately selects correct labels from samples rejected by AUM or the small-loss criterion, improving recall while maintaining precision in sample selection, and bringing performance gains to various benchmarks. To sum up, our key contributions are as follows:

- We analyze why correctly labeled and mislabeled samples exhibit different training dynamics from the perspective of coherent gradients (Chatterjee 2020) and provide supporting evidence.
- We propose a novel sample selection method based on monitoring changes in model predictions during training, termed Confidence Tracking (CT), which is a plug-and-play component that can integrate with existing sample selection methods and accurately distinguish correctly labeled yet hard-to-learn samples from mislabeled ones.
- We experimentally show that our method improves the performance of the existing sample selection methods on various benchmarks and real-world datasets.

2 Related Work

We review representative noise-robust methods and sample selection strategies in learning with noisy labels (LNL), excluding studies that assume access to clean label subsets (Xiao et al. 2015; Hendrycks et al. 2018; Qu, Mo, and Niu 2021; Tu et al. 2023).

Noise-robust methods. These methods address noisy labels through robust loss functions, loss correction, label correction, and regularization. *Robust loss functions* like MAE (Ghosh, Kumar, and Sastry 2017), GCE (Zhang and Sabuncu 2018), SCE(Wang et al. 2019), $\mathcal{L}_{\mathrm{DMI}}$(Xu et al. 2019), NCE(Ma et al. 2020), GJS(Englesson and Azizpour 2021), f-divergence(Wei and Liu 2021) and Peer loss(Liu and Guo 2020) are designed to mitigate noise. *Loss correction* methods estimate noise transition matrices, achieving success in class-dependent matrices estimating (Patrini et al. 2017; Yao et al. 2020), though accurately estimating the more general instance-dependent noise transition matrix remains challenging without additional assumptions (Xia et al. 2020; Berthon et al. 2021; Yao et al. 2021; Jiang et al. 2022; Cheng et al. 2022; Yang et al. 2022; Li et al. 2024). *Label correction* replaces noisy labels with model outputs (Tanaka et al. 2018; Yi and Wu 2019). Bootstraping (Reed et al. 2014) and M-correction (Arazo et al. 2019) use a convex combination of noisy labels and model predictions for training. *Regularization* methods constrain model capacity to prevent memorization. For instance, ELR (Liu et al. 2020) uses temporal regularization, NCR (Iscen et al. 2022) enforces similarity among neighbors, and contrastive

learning (Zheltonozhskii et al. 2022; Li et al. 2022; Yi et al. 2022; Xue, Whitecross, and Mirzasoleiman 2022; Peng et al. 2023) enhances robust representation. CS-Isolate (Lin et al. 2023) disentangling style and content in the representation space, distancing hard samples from the decision boundary to ease learning. Other regularization techniques including MixUp (Zhang 2017), label smoothing (Szegedy et al. 2016; Wei et al. 2022a; Ding et al. 2024; Fan et al. 2024), and early stopping (Bai et al. 2021) further improve noise tolerance.

Sample selection methods. These methods identify mislabeled samples using model predictions, representations, or training dynamics. *Model prediction-based* methods typically regard samples with small losses as correctly labeled. Co-teaching (Han et al. 2018) and its variants (Yu et al. 2019; Wei et al. 2020; Xia et al. 2022) simultaneously train two collaborating networks, selecting small-loss samples for each other to reduce confirmation bias. These methods need to dynamically adjust the select ratio in each iteration, which can be tricky in practice. A more flexible method is fitting a two-component Beta/Gaussian Mixture Model (BMM/GMM) on per-sample loss to differentiate correct and incorrect labels (Arazo et al. 2019; Li, Socher, and Hoi 2020; Nishi et al. 2021; Zhao et al. 2022). *Representation-based* approaches exploit latent features to distinguish clean from noisy data. CURST (Mirzasoleiman, Cao, and Leskovec 2020) selects samples that provide an approximately low-rank Jacobian matrix, which helps the network learn fast and generalize well. TopoFilter (Wu et al. 2020) assumes clean data clusters together while corrupted data is spread out in the feature representation, using high-order topological information to identify correct labels. FINE (Kim et al. 2021) detects mislabeled samples through the principal components of latent representations made by eigendecomposition. *Training dynamics-based* methods, utilize model predictions over multiple iterations to generate more accurate sample selections. AUM (Pleiss et al. 2020) and recently proposed HMW (Zhang et al. 2024) rank samples using the average logit margin to select correct labels. L2D (Jia et al. 2023) trains a noise detector based on training dynamics in a supervised manner, avoiding manual designing of the sample selection criterion. However, the pre-trained noise detector may not perform consistently well across different datasets with varying noise ratios, and fine-tuning the noise detector on target datasets requires additional clean data. DIST (Li et al. 2023) considered the fitting difficulty of different samples and proposed an instance-dependent sample selection criterion. Unlike previous methods that set a global or class-dependent threshold to select correct labels, they use the momentum maximum confidence of each instance computed across all previous epochs as the threshold value.

Our method belongs to sample selection methods based on training dynamics. Existing approaches typically set thresholds on per-sample loss, confidence, or logit margins to select correct labels, focusing primarily on the value in model predictions. These methods struggle to distinguish between correct and incorrect labels in high-loss data. In contrast, our method emphasizes the trend of confidence gaps in model predictions, enabling the identification of correct labels even within high-loss data. Combining our approach with existing methods can mitigate the trade-off between precision and recall in sample selection, selecting more correctly labeled yet hard-to-learn samples, resulting in stronger performance.

3 Method

Problem Setup

This paper considers the k-class classification problem in the presence of noisy labels. The noisy training set consists of n examples $\mathcal{D} = \{\mathbf{x}^{[i]}, \hat{\mathbf{y}}^{[i]}\}_{i=1}^{n}$, where $\mathbf{x}^{[i]} \in \mathbb{R}^d$ is the ith input and $\hat{\mathbf{y}}^{[i]} \in \{0,1\}^k$ is a one-hot vector indicating the annotated class. We use the non-bold letters $\hat{y}^{[i]}$ and $y^{[i]}$ to represent the annotated and ground truth classes of the ith input, respectively. For simplicity, we denote a deep neural network as $f(\cdot;\theta) : \mathbb{R}^b \mapsto \mathbb{R}^k$, which maps $\mathbf{x}^{[i]}$ to the conditional probability $\mathbf{p}^{[i]} \in \mathbb{R}^k$ for each class. We use $f(\mathbf{x}^{[i]};\theta)_c$ to represent the conditional probability $\mathcal{P}(y^{[i]} = c \mid \mathbf{x}^{[i]})$ predicted by the model. Typically, the parameters $\theta \in \mathbb{R}^p$ are optimized by minimizing the cross-entropy loss:

$$\mathcal{L}_{\mathrm{CE}} = \frac{1}{n}\sum_{i=1}^{n} \ell_{\mathrm{CE}}(\hat{\mathbf{y}}^{[i]}, \mathbf{p}^{[i]}). \tag{1}$$

$$\ell_{\mathrm{CE}}(\hat{\mathbf{y}}^{[i]}, \mathbf{p}^{[i]}) = -\sum_{c=1}^{k} \hat{\mathbf{y}}_c^{[i]} \log f(\mathbf{x}^{[i]};\theta)_c. \tag{2}$$

We consider the common stochastic gradient descent method, where the dataset is divided into multiple mini-batches $\mathcal{B} = \{\mathrm{b}_i\}_{i=1}^{|\mathcal{B}|}$, and perform gradient descent on each batch:

$$\mathcal{L}_{\mathrm{b}_t}(\theta) = -\frac{1}{|\mathrm{b}_t|}\sum_{(\mathbf{x}^{[i]},\mathbf{y}^{[i]})\in \mathrm{b}_t}\sum_{c=1}^{k} \hat{\mathbf{y}}_c^{[i]} \log f(\mathbf{x}^{[i]};\theta)_c, \tag{3}$$

$$\theta_{t+1} = \theta_t - \eta g_t = \theta_t - \eta \nabla \mathcal{L}_{\mathrm{b}_t}(\theta_t), \tag{4}$$

where η denotes the learning rate and θ_t, b_t represent the parameters and the sampled mini-batch at timestep t, respectively. Following previous work (Cheng et al. 2021; Zhou et al. 2021), we only consider *clean-labels-dominant* dataset which means training samples are more likely to be annotated with true semantic labels than with any other class labels.

An Empirical Analysis of Model Learning Process

Before detailing our method, we first analyze why the trend in confidence gaps can serve as a criterion for identifying correct labels in clean-labels-dominant datasets. We begin by examining how gradient descent over batches influences the classification model's prediction on a specific input. Typically, the parameters θ do not change significantly in a single gradient descent step. Thus, the nonlinear function $f(\mathbf{x};\theta_{t+1})_c$ can be approximated by its first-order Taylor expansion:

$$f(\mathbf{x};\theta_{t+1})_c \approx f(\mathbf{x};\theta_t)_c + \langle \nabla_{\theta_t} f(\mathbf{x};\theta_t)_c, \theta_{t+1} - \theta_t \rangle. \tag{5}$$

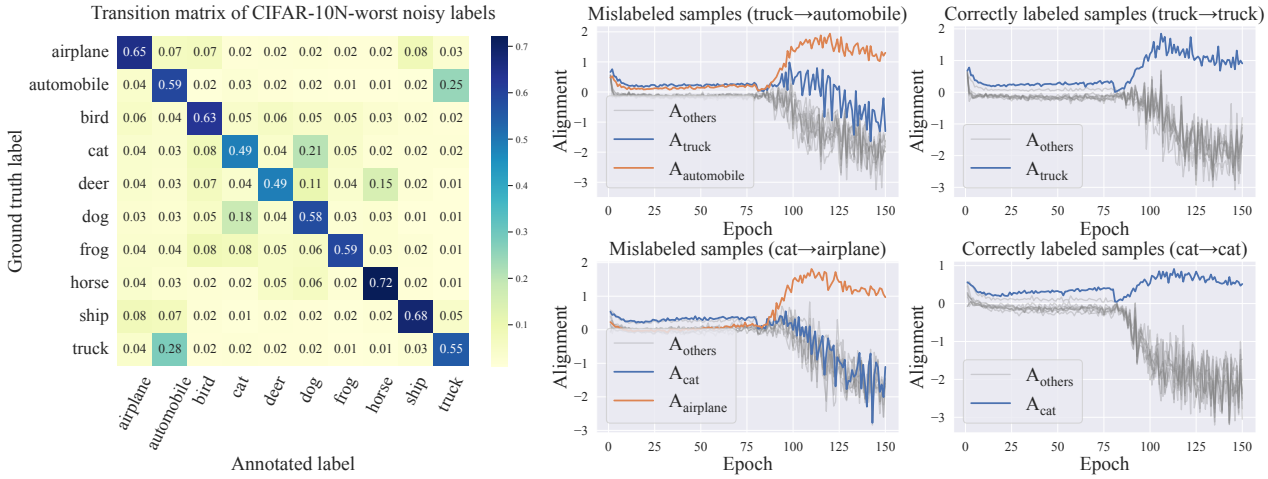

Figure 2: Left: Transition matrix of CIFAR-10N-Worst noisy labels. Right: Gradient alignment (Equation 7) for different types of samples when training a PreActResNet-18 using cross-entropy loss and an SGD optimizer on CIFAR-10N-Worst.

For input $\mathbf{x}$, changes in model's prediction confidence for class c after a gradient descent step on mini-batch b_t can be modeled as follows:

$$f(\mathbf{x};\theta_{t+1})_c - f(\mathbf{x};\theta_t)_c \propto \langle \nabla_{\theta_t} \ell_{\mathrm{CE}}(\mathbf{c},\mathbf{x}), \nabla \mathcal{L}_{\mathrm{b}_t}(\theta_t) \rangle. \quad (6)$$

The degree of alignment between $\nabla_{\theta_t} \ell_{\mathrm{CE}}(\mathbf{c},\mathbf{x})$ and $\nabla \mathcal{L}_{\mathrm{b}_t}(\theta_t)$ determines the direction of change in the model's prediction confidence. Intuitively, if $\nabla \mathcal{L}_{\mathrm{b}_t}(\theta_t)$ is aligned with $\nabla_{\theta_t} \ell_{\mathrm{CE}}(\mathbf{c},\mathbf{x})$, the gradient descent step on mini-batch b_t will decrease $\ell_{\mathrm{CE}}(\mathbf{c},\mathbf{x})$ and increase the model's prediction confidence for class c given input $\mathbf{x}$.

Previous studies on Coherent Gradients (Chatterjee 2020) indicate that gradients from similar examples are alike, and the overall gradient is stronger in directions where these reinforce each other. Consider a randomly initialized model that outputs random guesses on all inputs. During the early stages of training, the model has not yet fit the given annotations, and the gradients from correct and incorrect labels typically have similar magnitudes. Additionally, the gradients from correct labels are coherent since correctly labeled samples usually share similar patterns. As a result, in the clean-labels-dominant datasets, the correct labels will dominate the overall gradients in the early training stage. This means $\nabla \mathcal{L}_{\mathrm{b}_t}(\theta_t)$ tends to be most aligned with $\nabla_{\theta_t} \ell_{\mathrm{CE}}(\mathbf{y},\mathbf{x})$. In other words, the model's prediction confidence for ground truth labels tends to increase faster than for other classes in the early training stage. However, as the model gradually fits the correctly labeled examples, their gradients tend to diminish. Then the gradients from the under-fitted mislabeled examples will take over the gradient descent process, leading to the memorization of incorrect labels. To verify our analysis, we randomly sample a subset $\mathcal{N} = \{(\mathbf{x}^{[j]}, \mathbf{y}^{[j]}, \hat{\mathbf{y}}^{[j]})\}_{j=1}^{|\mathcal{N}|}$ consisting of examples from CIFAR-10N-Worst (Wei et al. 2022b), CIFAR-10 with noisy human annotations from Amazon Mechanical Turk, and report the following metrics in different training iterations:

$$A_c = \frac{1}{|\mathcal{N}|} \frac{1}{|\mathcal{B}|} \sum_{t=1}^{|\mathcal{B}|} \sum_{j=1}^{|\mathcal{N}|} \langle \nabla_{\theta_t} \ell_{\mathrm{CE}}(\mathbf{c},\mathbf{x}^{[j]}), \nabla \mathcal{L}_{\mathrm{b}_t}(\theta_t) \rangle. \quad (7)$$

Here, t indicates the gradient descent step in one iteration, increasing from 1 to the number of batches in the dataset. A_c measures the degree of alignment between $\nabla_{\theta_t} \ell_{\mathrm{CE}}(\mathbf{c},\mathbf{x})$ and the gradients over batches in one iteration. The larger A_c is, the faster the model's prediction confidence increases in category c. Figure 2 shows the transition matrix of CIFAR-10N-worst noisy labels and the trajectories of A_c on different kinds of samples. The experimental results are consistent with our analysis: A_y is highest on both correctly labeled and mislabeled samples in the early training stage. Only after a period of training, $A_{\hat{y}}$ will have a larger value than A_y. These empirical findings align with the *early learning* phenomenon (Liu et al. 2020), also known as memorization effect (Arpit et al. 2017), which suggests the deep neural networks optimized by SGD typically learn from correct labels before overfitting noisy data. This phenomenon has already been proved under high-dimensional linear classification (Liu et al. 2020). We provide additional evidence to illustrate why this occurs in deep neural networks. Note this phenomenon only suggests that the model's prediction confidence for ground truth labels usually rises fastest among all classes in the early training stage. However, it does not guarantee that the model will output high-confidence predictions for ground truth labels since some samples may be difficult to fit, causing the confidence to rise slowly and resulting in correctly labeled examples with high loss, just like the hard-to-learn dogs' images in Figure 1. This motivates us to design a novel sample selection method that considers trends in the model's prediction confidence rather than relying solely on loss values.

Sample Selection by Confidence Tracking

Our observations indicate that during the early stages of training, only for the correctly labeled samples, the model's prediction confidence for annotated labels usually increases faster than for any other classes. Based on this phenomenon, we introduce a novel sample selection method. Generally, for an input $\mathbf{x}^{[i]}$, if the model's prediction confidence increases faster for class c_1 than for class c_2, then the confidence gap between c_1 and c_2 should increase:

$$\begin{aligned} &\mathbf{p}_{c_1}^{[i]}(t+1) - \mathbf{p}_{c_1}^{[i]}(t) > \mathbf{p}_{c_2}^{[i]}(t+1) - \mathbf{p}_{c_2}^{[i]}(t) \\ \Rightarrow &\mathbf{p}_{c_1}^{[i]}(t+1) - \mathbf{p}_{c_2}^{[i]}(t+1) > \mathbf{p}_{c_1}^{[i]}(t) - \mathbf{p}_{c_2}^{[i]}(t), \end{aligned} \tag{8}$$

where $\mathbf{p}_c^{[i]}(t)$ denotes the model's prediction confidence for $\mathbf{x}^{[i]}$ in class c at iteration t. If the confidence gaps between the annotated label and other labels all tend to increase, the confidence should rise fastest on the annotated label. We regard such samples as potentially correctly labeled. To implement this idea, for each example $(\mathbf{x}^{[i]}, \hat{\mathbf{y}}^{[i]})$ in the noisy training set, we first collect the following confidence gaps:

$$d_c^{[i]}(t) = \mathbf{p}_{\hat{y}^{[i]}}^{[i]}(t) - \mathbf{p}_c^{[i]}(t). \tag{9}$$

Gathering those confidence gaps over different training iterations, we obtain the following series:

$$D_c^{[i]}(t) = \{d_c^{[i]}(1), d_c^{[i]}(2), ..., d_c^{[i]}(t)\}. \tag{10}$$

To judge the trend of these series, we utilize the Mann-Kendall Trend Test (Mann 1945; Kendall 1975). This is a non-parametric method and robust against extreme values. We use MK-Test$(\cdot)$ to represent the Mann-Kendall testing process, which takes a series of data as input and outputs the standardized test statistic Z. Our alternative hypothesis points to an upward trend in the confidence gaps series. Formally, our sample selection criterion is:

$$\min_{c \in \{1,2,...,k\} \setminus \{\hat{y}^{[i]}\}} \text{MK-Test}(D_c^{[i]}(t)) > Z_{1-\alpha}, \tag{11}$$

where α is the chosen significance level and $Z_{1-\alpha}$ is the $100(1-\alpha)$th percentile of the standard normal distribution. If MK-Test$(D_c^{[i]}(t)) > Z_{1-\alpha}$, the probability of $D_c^{[i]}(t)$ has no trend is less than α so we accept the alternative hypothesis. Therefore, if an example $(\mathbf{x}^{[i]}, \hat{\mathbf{y}}^{i})$ satisfies Equation 11, it suggests that all confidence gaps have an upward trend, so we regard it as potentially correctly labeled.

In practice, we combine Confidence Tracking (CT) with existing sample selection methods to enhance performance. Specifically, we use the union of samples selected by CT and other methods for training. Current methods reliably select correct labels from small-loss samples, while CT identifies correct labels from high-loss samples. This combination maintains precision and improves recall for sample selection. Let C_t denote the selected examples at iteration t, we assign zero weight to samples not in C_t in the loss function:

$$\mathcal{L}_t = \frac{1}{n} \sum\nolimits_{i=1}^{n} \mathbb{I}((\mathbf{x}^{[i]}, \hat{\mathbf{y}}^{[i]}) \in \mathcal{C}_t) \ell_{\text{CE}}(\hat{\mathbf{y}}^{[i]}, \mathbf{p}^{[i]}). \tag{12}$$

Following the common practice in sample selection (Li, Socher, and Hoi 2020; Kim et al. 2021; Li et al. 2023), we first warm up the network using standard cross-entropy loss for several epochs. Then we begin to select a potentially clean subset at the end of each epoch and apply Equation 12 for training. Generally, the selected subset likely excludes some mislabeled data, aligning the overall gradient towards memorizing ground truth labels, thereby promoting correct label memorization. This helps to generate better sample selection results in the following iterations. Such a positive feedback loop gradually eliminates mislabeled data from the noisy training set. Section C in the technical appendix provides the details of the Mann-Kendall Trend Test and the pseudo-code of our algorithm.

4 Experiment

Experimental settings

We evaluate our approach on four benchmarks, namely CIFAR-10, CIFAR-100 (Krizhevsky 2009), WebVision (Li et al. 2017), and Food-101N (Lee et al. 2018). For CIFAR-10 and CIFAR-100, we experiment with both simulated and human-annotated real-world noisy labels. For simulated noise, we follow the previous setups (Patrini et al. 2017; Liu et al. 2020) and experiment with two types of label noise: *symmetric* and *asymmetric*. Symmetric noise is generated by randomly replacing the original labels with all other possible classes. Asymmetric noise is a more realistic setting where labels are replaced by similar classes. We generate asymmetric noise following the same schema with previous works (Liu et al. 2020; Kim et al. 2021). For CIFAR-10, we map Turck $\rightarrow$ Automobile, Bird $\rightarrow$ Airplane, Deer $\rightarrow$ Horse, Cat $\leftrightarrow$ Dog. For CIFAR-100, we create 20 five-size superclasses and replace the original label with the next class within super-classes circularly. For human-annotated real-world noise, we experiment with the CIFARN (Wei et al. 2022b) dataset (CIFAR-10/100 dataset with noisy human annotations from Amazon Mechanical Turk). Unlike simulated class-dependent noise, real-world noise patterns are instance-dependent making it more challenging to identify correct labels.

WebVision and Food-101N are two real-world noisy datasets. WebVision dataset contains 2.4 million images crawled from the web and its estimated noise rate is about 20% (Li et al. 2017). Following previous work (Liu et al. 2020), we use the mini WebVision dataset for training, which contains the first 50 classes from the Google image subset (about 66 thousand images), and evaluate model performance on both WebVision and ImageNet ILSVRC12 validation sets (Deng et al. 2009) using InceptionResNetV2 (Szegedy et al. 2017). The Food-101N dataset contains about 310,009 images of food recipes classified in 101 categories and its estimated noise rate is about 20%. Food-101N and the Food-101 dataset (Bossard, Guillaumin, and Van Gool 2014) share the same 101 classes, whereas Food-101N has much more images and is more noisy. Following previous work (Lee et al. 2018), we use ResNet50 (He et al. 2016) pretrained on ImageNet (Deng et al. 2009) and evaluate model performance on the Food-101 test set.

Dataset	CIFAR-10				CIFAR-100				Avg
Noise type	Sym. 20%	Sym. 50%	Asym. 40%	Real. 40%	Sym. 20%	Sym. 50%	Asym. 40%	Real. 40%	
CE	86.51±0.22	77.41±0.65	83.78±1.76	77.94±0.91	61.37±0.12	46.82±1.32	45.70±0.39	52.82±0.30	66.54
L2D	92.25±0.12	87.27±0.55	82.57±1.31	84.50±0.44	71.05±0.47	60.82±0.59	47.94±0.63	59.44±0.33	73.23
Co-teaching	91.88±0.21	87.58±0.41	87.72±1.00	85.22±0.28	70.45±0.36	64.07±0.47	58.95±0.91	62.32±0.26	76.03
CNLCU	91.92±0.36	87.58±0.73	88.14±0.61	86.08±0.39	70.61±0.34	63.87±0.11	55.94±0.73	62.28±0.20	75.80
HMW	92.02±0.21	87.81±0.22	87.37±0.29	85.28±0.29	72.01±0.22	65.22±0.36	64.69±0.13	61.52±0.28	76.99
GMM	91.60±0.28	88.07±0.08	89.45±0.85	86.63±0.34	69.59±0.32	63.95±0.44	65.29±0.42	60.22±0.23	76.85
GMM+CT	**92.57±0.12**	**89.11±0.21**	**90.55±0.22**	**87.33±0.38**	**71.37±0.67**	**65.17±0.39**	**68.84±0.47**	**62.73±0.14**	78.46
FINE	89.13±0.48	85.66±0.32	82.56±2.00	80.09±0.45	70.96±0.45	58.58±0.48	49.48±0.77	56.87±0.25	71.67
FINE+CT	**92.48±0.21**	**87.56±0.13**	**86.92±0.51**	**84.22±0.49**	**71.17±0.32**	**58.76±0.41**	**53.16±0.88**	**58.52±0.15**	74.10
AUM	92.31±0.13	87.80±0.24	88.21±0.54	86.22±0.11	72.50±0.44	64.90±0.28	61.25±0.41	61.75±0.38	76.87
AUM+CT	**92.45±0.13**	87.91±0.40	**89.70±0.40**	**87.29±0.16**	72.56±0.18	64.99±0.41	**63.80±0.37**	**62.05±0.18**	77.59
DIST	92.63±0.15	88.43±0.24	90.00±0.42	86.39±0.54	72.73±0.32	65.59±0.24	66.74±0.81	60.97±0.20	77.93
DIST+CT	92.43±0.31	88.35±0.16	**90.58±0.21**	**87.01±0.43**	72.78±0.27	65.51±0.16	**69.05±0.47**	**62.18±0.17**	78.49

Table 1: Test accuracy (%) of different methods on CIFAR-10 and CIFAR-100 with symmetric, asymmetric, and real-world noisy labels (CIFAR-10N-Worst and CIFAR-100N-Noisy). We implement all methods based on public code and report mean accuracy and standard deviation over five random seeds. We use the Wilcoxon signed-rank test with a confidence level of 0.05 to compare the performance and bold the results where CT brings significant improvements.

Noise setting	Sym. 20%			Sym. 50%			Asym. 40%			Real. 40%			Average		
Metric	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
GMM	99.86	71.11	83.06	98.97	69.46	81.62	98.64	63.84	77.51	89.52	72.82	80.31	96.75	69.30	80.62
GMM+CT	99.71	83.36	90.80	98.33	79.61	87.99	96.46	82.12	88.72	87.04	84.61	85.81	95.39	82.43	**88.33**
FINE	97.93	93.23	95.52	85.68	95.14	90.16	69.98	72.13	71.03	73.23	83.68	78.11	81.70	86.05	83.70
FINE+CT	97.65	96.34	96.99	85.52	96.16	90.53	72.76	85.90	78.78	72.97	93.99	82.15	82.22	93.10	**87.11**
AUM	99.09	92.84	95.86	96.15	86.67	91.17	86.26	72.03	78.51	84.48	84.67	84.57	91.50	84.05	87.53
AUM+CT	99.10	92.99	95.95	96.14	87.06	91.38	88.82	76.69	82.31	84.58	85.71	85.14	92.16	85.61	**88.69**
DIST	99.04	94.74	96.84	95.56	91.56	93.51	96.87	73.88	83.81	85.80	79.10	82.31	94.32	84.82	89.12
DIST+CT	99.00	94.93	96.93	95.27	92.25	93.73	96.56	87.55	91.84	85.88	86.32	86.09	94.18	90.27	**92.15**

Table 2: Comparisons of sample selection precision, recall, and F1-score to correct labels on CIFAR-100 dataset with symmetric, asymmetric, and real-world noisy labels. Results are averaged over five random seeds.

Integrating with Sample Selection Methods

We mainly experiment with the following advanced sample selection methods: Co-teaching (Han et al. 2018), CNLCU (Xia et al. 2022), GMM (Li, Socher, and Hoi 2020), FINE (Kim et al. 2021), L2D (Jia et al. 2023), AUM (Pleiss et al. 2020), DIST (Li et al. 2023), and HMW (Zhang et al. 2024), which cover representative methods based on model predictions, sample representations, and training dynamics. Section F in the technical appendix provides a more detailed introduction and implementation details of those baselines.

Tabel 1 demonstrates the performance when integrating CT with state-of-the-art sample selection approaches. All methods use the same architecture (PreActResNet-18) and training procedure (please refer to Section F in the technique appendix for more details). The significance level α used in CT is set to 0.01 and the warm-up epoch is 30 for all methods. We retain 10% of the training sets to perform validation and select the model with the best validation performance for the test. CT brings consistent improvement to various baselines. We notice the improvement is more significant under asymmetric and real-world noise. This is because the symmetric noise randomly changes the correct label to another one, making the gradient between noise samples usually incoherent. As a result, the model fits correct labels much faster than incorrect ones, making it easy to distinguish them using the loss values. So sample selection methods based on the small-loss criterion can achieve satisfying performance under symmetric noise. However, under asymmetric or real-world noise, noise labels are more structured (*e.g.,* trucks are generally mislabeled as automobiles rather than other categories), meaning the gradient of this mislabeled data is also coherent. Thus, the model can quickly fit mislabeled data, making the loss of some correctly labeled yet hard-to-learn samples similar to misla-

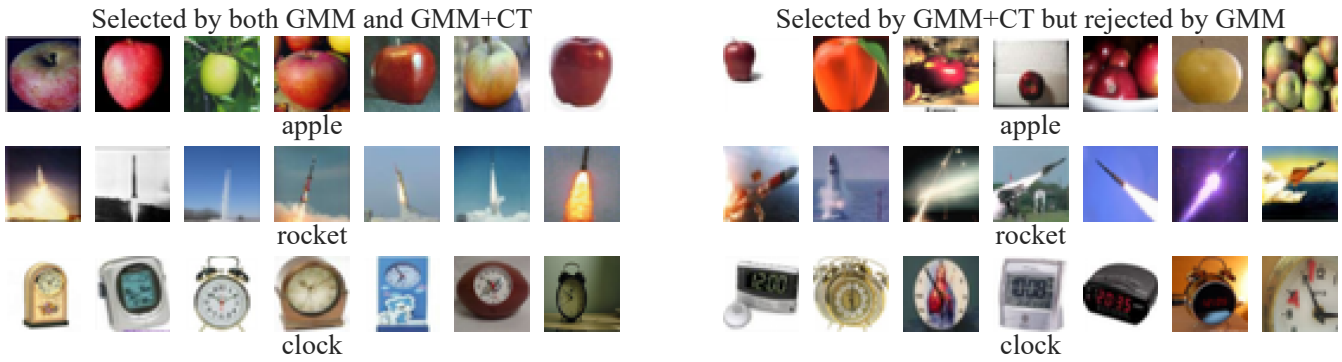

Figure 3: The sample selection results on the CIFAR-100N-noisy dataset. The left graph shows samples selected by both GMM and GMM+CT and the right graph shows samples selected by GMM+CT but rejected by GMM. These samples are chosen randomly, not cherry-picked.

beled ones. The popular small-loss criterion is less effective under such situations, while CT can still distinguish hard correctly labeled data from mislabeled ones. Tabel 2 compares the sample selection precision and recall for correct labels of different methods. Generally, integrating CT with other sample selection methods significantly improves recall and maintains the precision of sample selection which shows CT accurately identifies correct labels from samples rejected by existing sample selection methods, *i.e.,* samples with relatively high loss or logit margins. These results confirm that CT is more powerful than existing sample selection methods in distinguishing correctly labeled yet hard-to-learn and mislabeled samples. Section D in the technical appendix provides additional results on synthetic instance-dependent label noise (Xia et al. 2020).

Figure 3 compares the sample selected by GMM and GMM+CT on the CIFAR-100N-noisy dataset. Compared with images selected by both GMM and GMM+CT, images selected only by GMM+CT show greater inter-class variability. For instance, the apple images encompass different environment settings and perspectives; the rocket images capture various stages of rocket launches; the clock images display diverse designs and time formats. It intuitively shows that introducing CT can select a richer sample set, which helps improve model performance. Due to the page limit, we further analyze the robustness of CT in Section E in the technical appendix, which shows CT is not sensitive to the choice of α and the number of warm-up epochs.

Integrating with Advanced LNL Methods

Existing state-of-the-art methods of learning with noisy labels usually combine sample selection with semi-supervised learning to further improve performance. In this section, we integrate CT with CORSE (Cheng et al. 2021), DivideMix (Li, Socher, and Hoi 2020), f-DivideMix (Kim et al. 2021), and DISC (Li et al. 2023) to analyze whether our sample selection procedure can bring further improvement to these state-of-the-art methods. Table 3 shows CT consistently improves the performance of all baselines on real-world noisy datasets.

Test dataset	Webvision		ILSVRC12		Food-101N
Metric	Top1	Top5	Top1	Top5	ACC
CORSE	71.70	89.02	68.36	88.28	84.38
CORSE+CT	**72.11**	**90.24**	**68.52**	**90.03**	**84.43**
DivideMix	77.44	91.88	74.72	92.12	86.53
DivideMix+CT	**78.12**	**92.26**	**75.23**	**92.19**	**86.76**
f-DivideMix	78.36	92.54	75.25	92.22	86.83
f-DivideMix+CT	**78.81**	**92.91**	**75.66**	**93.22**	**87.05**
DISC	80.07	92.38	77.40	92.38	87.32
DISC+CT	80.07	**92.56**	**78.26**	**92.43**	**87.45**

Table 3: The average test accuracy (%) over the last 10 epochs on the real-world noisy dataset.

5 Conclusion

In this paper, we introduced a novel sample selection method for image classification with noisy labels, termed Confidence Tracking (CT). Unlike existing methods that rely on small-loss criteria, our approach leverages the observation that only for the correctly labeled samples, the model's prediction confidence for annotated labels usually increases faster than for any other classes. By monitoring the trends in confidence gaps between annotated labels and other classes, CT effectively distinguishes correctly labeled samples even when they exhibit high losses during training. Our experimental results demonstrate that CT enhances the performance of existing learning with noisy labels (LNL) methods across various benchmarks, showcasing its robustness and reliability. This method successfully alleviates the trade-off dilemma between precision and recall in sample selection when setting a threshold on pre-sample loss, model prediction confidence, or logit margins, offering a more accurate identification of hard-to-learn yet correctly labeled samples. Future research could explore a deeper theoretical understanding of the early learning phenomenon and the development of more advanced sample selection criteria based on training dynamics.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62276110, No. 62172039, and in part by the fund of Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL). The authors would also like to thank the anonymous reviewers for their comments on improving the quality of this paper.

References

Arazo, E.; Ortego, D.; Albert, P.; O'Connor, N.; and McGuinness, K. 2019. Unsupervised label noise modeling and loss correction. In *ICML*, 312–321. PMLR.

Arpit, D.; Jastrzebski, S.; Ballas, N.; Krueger, D.; Bengio, E.; Kanwal, M. S.; Maharaj, T.; Fischer, A.; Courville, A.; Bengio, Y.; et al. 2017. A closer look at memorization in deep networks. In *ICML*, 233–242. PMLR.

Bai, Y.; Yang, E.; Han, B.; Yang, Y.; Li, J.; Mao, Y.; Niu, G.; and Liu, T. 2021. Understanding and improving early stopping for learning with noisy labels. In *NeurIPS*, 24392–24403.

Berthon, A.; Han, B.; Niu, G.; Liu, T.; and Sugiyama, M. 2021. Confidence scores make instance-dependent label-noise learning possible. In *ICML*, 825–836. PMLR.

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101–mining discriminative components with random forests. In *ECCV*, 446–461. Springer.

Chatterjee, S. 2020. Coherent Gradients: An Approach to Understanding Generalization in Gradient Descent-based Optimization. In *ICLR*.

Cheng, D.; Liu, T.; Ning, Y.; Wang, N.; Han, B.; Niu, G.; Gao, X.; and Sugiyama, M. 2022. Instance-dependent label-noise learning with manifold-regularized transition matrix estimation. In *CVPR*, 16630–16639.

Cheng, H.; Zhu, Z.; Li, X.; Gong, Y.; Sun, X.; and Liu, Y. 2021. Learning with Instance-Dependent Label Noise: A Sample Sieve Approach. In *ICLR*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255. Ieee.

Ding, Z.; Wei, W.; Qu, X.; and Chen, D. 2024. Improving Pseudo Labels with Global-Local Denoising Framework for Cross-lingual Named Entity Recognition. In *IJCAI*, 6252–6260.

Englesson, E.; and Azizpour, H. 2021. Generalized jensen-shannon divergence loss for learning with noisy labels. In *NeurIPS*, 30284–30297.

Fan, S.; Wei, W.; Wen, X.; Mao, X.; Chen, J.; and Chen, D. 2024. Personalized Topic Selection Model for Topic-Grounded Dialogue. In *Findings of ACL*, 7188–7202.

Ghosh, A.; Kumar, H.; and Sastry, P. S. 2017. Robust loss functions under label noise for deep neural networks. In *AAAI*, volume 31.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I. W.; and Sugiyama, M. 2018. Co-teaching: robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 8536–8546.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hendrycks, D.; Mazeika, M.; Wilson, D.; and Gimpel, K. 2018. Using trusted data to train deep networks on labels corrupted by severe noise. In *NeurIPS*, 10477–10486.

Iscen, A.; Valmadre, J.; Arnab, A.; and Schmid, C. 2022. Learning with neighbor consistency for noisy labels. In *CVPR*, 4672–4681.

Jia, Q.; Li, X.; Yu, L.; Bian, J.; Zhao, P.; Li, S.; Xiong, H.; and Dou, D. 2023. Learning from training dynamics: Identifying mislabeled data beyond manually designed features. In *AAAI*, volume 37, 8041–8049.

Jiang, Z.; Zhou, K.; Liu, Z.; Li, L.; Chen, R.; Choi, S.-H.; and Hu, X. 2022. An information fusion approach to learning with instance-dependent label noise. In *ICLR*.

Kendall, M. 1975. *Rank Correlation Methods*. Charles Griffin, 4th edition.

Kim, T.; Ko, J.; Cho, S.; Choi, J.; and Yun, S.-Y. 2021. FINE samples for learning with noisy labels. In *NeurIPS*, 24137–24149.

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images. *Master's thesis, University of Tront*.

Lee, K.-H.; He, X.; Zhang, L.; and Yang, L. 2018. Cleannet: Transfer learning for scalable image classifier training with label noise. In *CVPR*, 5447–5456.

Li, J.; Socher, R.; and Hoi, S. C. 2020. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In *ICLR*.

Li, S.; Xia, X.; Deng, J.; Gey, S.; and Liu, T. 2024. Transferring annotator-and instance-dependent transition matrix for learning from crowds. *TPAMI*.

Li, S.; Xia, X.; Ge, S.; and Liu, T. 2022. Selective-supervised contrastive learning with noisy labels. In *CVPR*, 316–325.

Li, W.; Wang, L.; Li, W.; Agustsson, E.; and Gool, L. V. 2017. WebVision Database: Visual Learning and Understanding from Web Data. *CoRR*, abs/1708.02862.

Li, Y.; Han, H.; Shan, S.; and Chen, X. 2023. Disc: Learning from noisy labels via dynamic instance-specific selection and correction. In *CVPR*, 24070–24079.

Lin, Y.; Yao, Y.; Shi, X.; Gong, M.; Shen, X.; Xu, D.; and Liu, T. 2023. CS-isolate: extracting hard confident examples by content and style isolation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 58556–58576.

Liu, S.; Niles-Weed, J.; Razavian, N.; and Fernandez-Granda, C. 2020. Early-learning regularization prevents memorization of noisy labels. In *NeurIPS*, 20331–20342.

Liu, Y.; and Guo, H. 2020. Peer loss functions: Learning from noisy labels without knowing noise rates. In *ICML*, 6226–6236. PMLR.

Ma, X.; Huang, H.; Wang, Y.; Romano, S.; Erfani, S.; and Bailey, J. 2020. Normalized loss functions for deep learning with noisy labels. In *ICML*, 6543–6553. PMLR.

Mann, H. B. 1945. Nonparametric tests against trend. *Econometrica: Journal of the econometric society*, 245–259.

Mirzasoleiman, B.; Cao, K.; and Leskovec, J. 2020. Coresets for robust training of neural networks against noisy labels. In *NeurIPS*, 11465–11477.

Natarajan, N.; Dhillon, I. S.; Ravikumar, P.; and Tewari, A. 2013. Learning with noisy labels. In *NeurIPS*, 1196–1204.

Nishi, K.; Ding, Y.; Rich, A.; and Hollerer, T. 2021. Augmentation strategies for learning with noisy labels. In *CVPR*, 8022–8031.

Patrini, G.; Rozza, A.; Krishna Menon, A.; Nock, R.; and Qu, L. 2017. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*, 1944–1952.

Peng, D.; Wei, W.; Mao, X.-L.; Fu, Y.; and Chen, D. 2023. An Empirical Study on the Language Modal in Visual Question Answering. In *IJCAI*, 4109–4117.

Pleiss, G.; Zhang, T.; Elenberg, E.; and Weinberger, K. Q. 2020. Identifying mislabeled data using the area under the margin ranking. In *NeurIPS*, 17044–17056.

Qu, Y.; Mo, S.; and Niu, J. 2021. Dat: Training deep networks robust to label-noise by matching the feature distributions. In *CVPR*, 6821–6829.

Reed, S.; Lee, H.; Anguelov, D.; Szegedy, C.; Erhan, D.; and Rabinovich, A. 2014. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*.

Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 31.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *CVPR*, 2818–2826.

Tanaka, D.; Ikami, D.; Yamasaki, T.; and Aizawa, K. 2018. Joint optimization framework for learning with noisy labels. In *CVPR*, 5552–5560.

Tu, Y.; Zhang, B.; Li, Y.; Liu, L.; Li, J.; Wang, Y.; Wang, C.; and Zhao, C. R. 2023. Learning from noisy labels with decoupled meta label purifier. In *CVPR*, 19934–19943.

Wang, Y.; Ma, X.; Chen, Z.; Luo, Y.; Yi, J.; and Bailey, J. 2019. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*, 322–330.

Wei, H.; Feng, L.; Chen, X.; and An, B. 2020. Combating noisy labels by agreement: A joint training method with co-regularization. In *CVPR*, 13726–13735.

Wei, J.; Liu, H.; Liu, T.; Niu, G.; Sugiyama, M.; and Liu, Y. 2022a. To Smooth or Not? When Label Smoothing Meets Noisy Labels. In *ICML*, 23589–23614. PMLR.

Wei, J.; and Liu, Y. 2021. When Optimizing f-Divergence is Robust with Label Noise. In *ICLR*.

Wei, J.; Zhu, Z.; Cheng, H.; Liu, T.; Niu, G.; and Liu, Y. 2022b. Learning with Noisy Labels Revisited: A Study Using Real-World Human Annotations. In *ICLR*.

Wu, P.; Zheng, S.; Goswami, M.; Metaxas, D.; and Chen, C. 2020. A topological filter for learning with label noise. In *NeurIPS*, 21382–21393.

Xia, X.; Liu, T.; Han, B.; Gong, M.; Yu, J.; Niu, G.; and Sugiyama, M. 2022. Sample Selection with Uncertainty of Losses for Learning with Noisy Labels. In *ICLR*.

Xia, X.; Liu, T.; Han, B.; Wang, N.; Gong, M.; Liu, H.; Niu, G.; Tao, D.; and Sugiyama, M. 2020. Part-dependent label noise: towards instance-dependent label noise. In *NeurIPS*, 7597–7610.

Xiao, T.; Xia, T.; Yang, Y.; Huang, C.; and Wang, X. 2015. Learning from massive noisy labeled data for image classification. In *CVPR*, 2691–2699.

Xu, Y.; Cao, P.; Kong, Y.; and Wang, Y. 2019. L_DMI: A novel information-theoretic loss function for training deep nets robust to label noise. In *NeurIPS*, 6225–6236.

Xue, Y.; Whitecross, K.; and Mirzasoleiman, B. 2022. Investigating why contrastive learning benefits robustness against label noise. In *ICML*, 24851–24871. PMLR.

Yang, S.; Yang, E.; Han, B.; Liu, Y.; Xu, M.; Niu, G.; and Liu, T. 2022. Estimating instance-dependent bayes-label transition matrix using a deep neural network. In *ICML*, 25302–25312. PMLR.

Yao, Y.; Liu, T.; Gong, M.; Han, B.; Niu, G.; and Zhang, K. 2021. Instance-dependent label-noise learning under structural causal models. In *NeurIPS*, 4409–4420.

Yao, Y.; Liu, T.; Han, B.; Gong, M.; Deng, J.; Niu, G.; and Sugiyama, M. 2020. Dual T: reducing estimation error for transition matrix in label-noise learning. In *NeurIPS*, 7260–7271.

Yi, K.; and Wu, J. 2019. Probabilistic end-to-end noise correction for learning with noisy labels. In *CVPR*, 7017–7025.

Yi, L.; Liu, S.; She, Q.; McLeod, A. I.; and Wang, B. 2022. On learning contrastive representations for learning with noisy labels. In *CVPR*, 16682–16691.

Yu, X.; Han, B.; Yao, J.; Niu, G.; Tsang, I.; and Sugiyama, M. 2019. How does disagreement help generalization against label corruption? In *ICML*, 7164–7173. PMLR.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2021. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3): 107–115.

Zhang, H. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhang, S.; Li, Y.; Wang, Z.; Li, J.; and Liu, C. 2024. Learning with Noisy Labels Using Hyperspherical Margin Weighting. In *AAAI*, volume 38, 16848–16856.

Zhang, Z.; and Sabuncu, M. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *NeurIPS*, 31.

Zhao, G.; Li, G.; Qin, Y.; Liu, F.; and Yu, Y. 2022. Centrality and Consistency: Two-Stage Clean Samples Identification for Learning with Instance-Dependent Noisy Labels. In *ECCV*, 21–37.

Zheltonozhskii, E.; Baskin, C.; Mendelson, A.; Bronstein, A. M.; and Litany, O. 2022. Contrast to divide: Self-supervised pre-training for learning with noisy labels. In *WACV*, 1657–1667.

Zhou, X.; Liu, X.; Jiang, J.; Gao, X.; and Ji, X. 2021. Asymmetric loss functions for learning with noisy labels. In *ICML*, 12846–12856. PMLR.