Dirichlet-Based Prediction Calibration for Learning with Noisy Labels

Chen-Chen Zong, Ye-Wen Wang, Ming-Kun Xie, Sheng-Jun Huang*

College of Computer Science and Technology/Artificial Intelligence, Nanjing University of Aeronautics and Astronautics
MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing, China
{chencz, linuswangg, mkxie, huangsj}@nuaa.edu.cn

Abstract

Learning with noisy labels can significantly hinder the generalization performance of deep neural networks (DNNs). Existing approaches address this issue through loss correction or example selection methods. However, these methods often rely on the model's predictions obtained from the softmax function, which can be over-confident and unreliable. In this study, we identify the translation invariance of the softmax function as the underlying cause of this problem and propose the *Dirichlet-based Prediction Calibration* (DPC) method as a solution. Our method introduces a calibrated softmax function that breaks the translation invariance by incorporating a suitable constant in the exponent term, enabling more reliable model predictions. To ensure stable model training, we leverage a Dirichlet distribution to assign probabilities to predicted labels and introduce a novel evidence deep learning (EDL) loss. The proposed loss function encourages positive and sufficiently large logits for the given label, while penalizing negative and small logits for other labels, leading to more distinct logits and facilitating better example selection based on a large-margin criterion. Through extensive experiments on diverse benchmark datasets, we demonstrate that DPC achieves state-of-the-art performance. The code is available at https://github.com/chenchenzong/DPC.

Introduction

Large-scale datasets with high-quality annotations are crucial for achieving remarkable performance in deep neural networks (DNNs). However, collecting a large number of accurately annotated data is often costly and time-consuming. Recently, crowdsourcing labeling has become a mainstream solution to this problem due to its cost-effectiveness (Hossain and Kauranen 2015; Chen et al. 2014; Huang et al. 2021). While the labeling cost is significantly reduced, it often introduces noisy labels unavoidably due to the various levels of expertise possessed by different labelers. Directly learning with such noisy labels can easily degrade the generalization performance of DNNs (Zhang et al. 2021). Therefore, training robust DNNs with noisy labels has become a challenge that attracted significant attention in recent years (Han et al. 2018; Huang et al. 2019; Li, Socher, and Hoi 2020; Liu et al. 2020; Zong et al. 2022; Karim et al. 2022).

*Corresponding author.

Existing methods can be broadly classified into two categories: loss correction (Patrini et al. 2017; Ma et al. 2018; Zhang and Sabuncu 2018; Shu et al. 2019; Ma et al. 2020) and example selection (Garcia, de Carvalho, and Lorena 2016; Malach and Shalev-Shwartz 2017; Huang et al. 2019; Li, Socher, and Hoi 2020; Zhou, Wang, and Bilmes 2021; Karim et al. 2022). The former aims to correct the loss by estimating the noise transition matrix, adjusting the example labels or weights. The latter attempts to separate clean examples from noisy ones based on the small-loss criterion (Li, Socher, and Hoi 2020), where the examples with low loss are assumed to have clean labels, and further consider recognized mislabeled examples as unlabeled ones to perform semi-supervised learning. Despite these two kinds of methods making great progress in dealing with noisy labels, most of them often suffer from the unreliability issue since standard DNNs can easily produce over-confident but incorrect predictions (Sensoy, Kaplan, and Kandemir 2018; Wang, Feng, and Zhang 2021; Xie et al. 2023). For example, to demonstrate the over-confidence issue of DNNs, Sensoy, Kaplan, and Kandemir (2018) experimented on handwritten digit recognition and observed that when the digit "1" is rotated at an angle greater than 60 degrees, the model tends to output the digit "2" with very high confidence.

In this paper, we first disclose that the occurrence of the over-confidence phenomenon is caused by the translation invariance of the softmax function, *i.e.*, adding or subtracting a constant from all logits does not alter the softmax values, and this actually elevates the probability of misclassifying mislabeled examples as clean ones in noisy label learning. Based on these findings, we develop a Dirichlet-based Prediction Calibration (DPC) method to calibrate the softmax probabilities and thus solve the over-confidence issue. Specifically, to obtain reliable confidence, a suitable constant is added to the exponent term of the softmax function to break its translation invariance. To solve the gradient shrinking issue caused by the calibration, a Dirichlet training scheme is developed to enforce the logits on the given label and other labels to be as separated as possible. With more distinguishable logits, we further design a large-margin criterion to achieve better performance on example selection.

The main contributions are summarized as follows:

- We disclose that softmax's translation invariance leads to over-confidence problems and may amplify the risk of

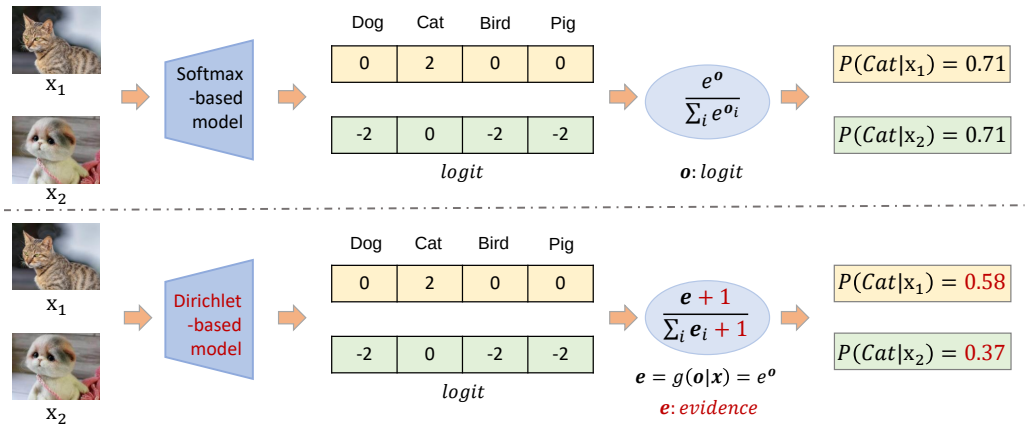


Figure 1: A specific case comparing the softmax-based model and our proposed calibrated Dirichlet-based model. The softmax function has translation invariance, *i.e.*, can only reflect the relative relationship between logits, and gives the same prediction for $\mathbf{x}_1$ and $\mathbf{x}_2$, which contradicts our subjective intuition. We break the translation invariance by placing a suitable constant on the exponent term and proposing a corresponding Dirichlet-based training method.

misclassifying mislabeled examples as clean ones.

- We propose a Dirichlet-based prediction calibration method to overcome over-confidence problems. The method calibrates the softmax function by breaking its translation invariance and improves the distinctiveness of predicted logits by designing an EDL loss.
- We drive a large-margin example selection criterion that is more suitable to the calibrated model. Compared to the small-loss criterion, our large-margin criterion is able to take full advantage of the distinguishable logits and thus achieve better example selection.
- We conduct extensive experiments on benchmark and real-world datasets to demonstrate that DPC can achieve competitive performance compared with state-of-the-art methods.

Related Work

Loss Correction. This category of methods is implemented mainly in three ways: noise transition estimation, label correction, and example reweighting. For the first type, Goldberger and Ben-Reuven (2017) estimated the noise transition matrix by adding an additional linear layer on the top of the neural networks to correct the loss function. Patrini et al. (2017) proposed to exploit anchor points (data points that belong to a specific class almost surely) to obtain a pre-calculated Backward or Forward noise transition matrix. Label correction aims to correct the noisy labels. Reed et al. (2014) first proposed to update example labels by using their pseudo-labels in each training epoch. Huang, Zhang, and Zhang (2020) introduced the exponential moving average into the label refurbishment process to alleviate the instability issue caused by the instantaneous prediction. Example reweighting tries to eliminate the effects of noise labels by assigning small weights to the possibly mislabeled examples. Chang, Learned-Miller, and McCallum (2017) used the example prediction variance as its weight to emphasize examples with inconsistent predictions. Shu et al. (2019) introduced meta-learning to automatically learn an explicit loss-weight function based on an additional clean dataset.

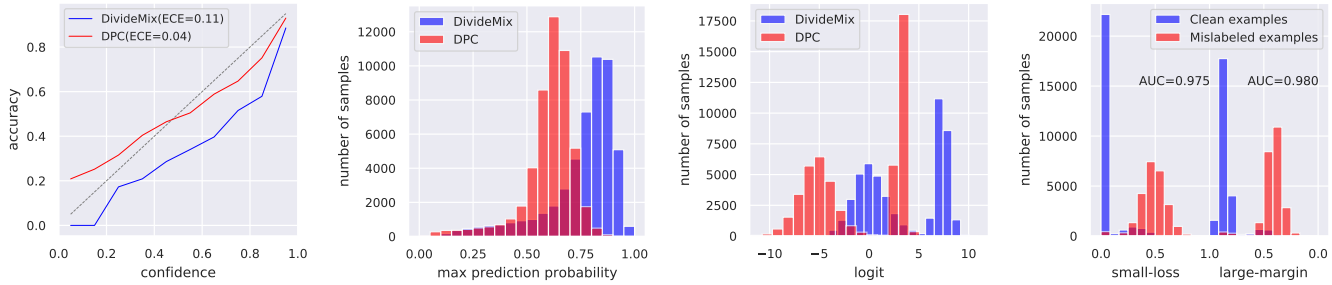
Example Selection. This category of methods attempts to directly identify potentially noisy examples and then learn only based on clean examples or learning in a semi-supervised manner. Huang et al. (2019) proposed a straightforward noisy label detection approach named O2U-net, which requires adjusting the learning rate to keep the network transferring from overfitting to underfitting cyclically and then distinguishes examples by their accumulated loss values. Jiang et al. (2018) firstly trains a teacher network and then uses it to select small loss examples as clean examples for guiding the training of the student network. Co-teaching (Han et al. 2018) and Co-teaching+ (Yu et al. 2019) maintain two networks simultaneously and let them select training examples for each other. DivideMix (Li, Socher, and Hoi 2020) leverages the Gaussian mixture model to distinguish clean and mislabeled data and introduces a semi-supervised technique called MixMatch (Berthelot et al. 2019) to leverage recognized mislabeled examples. Liu et al. (2020) and Bai et al. (2021) further improve this by proposing an early learning regularization term and a progressive early stopping technique to prevent the model from memorization noisy labels, respectively.

Analysis of the Over-Confidence Phenomenon

Formally, let $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$ denotes the corrupted training set with N training examples, where each $\boldsymbol{x}_i$ represents an example and $\boldsymbol{y}_i \in \{0,1\}^C$ is the corrupted one-hot label over C classes. Given a model $f(\cdot;\boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta}$, we fit the training data by minimizing the cross entropy loss:

$$\mathcal{L}_{ce} = \frac{-1}{N}\sum_{i=1}^{N} \boldsymbol{y}_i \log \boldsymbol{\rho}_i = \frac{-1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic} \log \frac{e^{o_{ic}}}{\sum_{j=1}^{C} e^{o_{ij}}},$$

where $\boldsymbol{o}_i = f(\boldsymbol{x}_i;\boldsymbol{\theta}) = [o_{i1}, o_{i2}, \ldots, o_{iC}]$ denotes the logit vector, $\boldsymbol{\rho}_i = [\rho_{i1}, \rho_{i2}, \ldots, \rho_{iC}]$ is the softmax predicted probability vector by applying the softmax function to $\boldsymbol{o}_i$.



(a) (b) (c) (d)

Figure 2: Training on CIFAR-10 with a 50% symmetric noise rate. All the figures are plotted based on the results of the last epoch. (a) The Expected Calibration Error (ECE) results of the test data. A smaller ECE value is better, and correspondingly, a line closer to the dashed line is preferred. We can see that the softmax-based DivideMix tends to produce over-confident predictions. (b) The distribution of the maximum predicted probability for training examples. (c) The distribution of the given label logit for training examples. (d) The comparison of the example selection criterion. We can see that the proposed large-margin criterion can produce more discriminative results.

Our goal is to train a sufficiently good model based on the corrupted dataset $\mathcal{D}$. Towards this goal, many current methods (Li, Socher, and Hoi 2020; Bai et al. 2021; Karim et al. 2022) first partitioned the whole dataset into clean and mislabeled subsets and then performing semi-supervised learning by treating the former as labeled, while the latter as unlabeled. For example, DivideMix (Li, Socher, and Hoi 2020), as a representative method, fits a two-component Gaussian Mixture Model (GMM) (Permuter, Francos, and Jermyn 2006) to the example loss and uses the posterior probability of the component with a smaller mean to partition clean and mislabeled subsets. Then, a semi-supervised learning method called MixMatch (Berthelot et al. 2019) is adopted.

It is noteworthy that both example selection and label correction highly depend on the quality of model predictions. If the predicted probabilities are closer to true ones, then the model would select clean examples more precisely. We perform an experiment to show that the model trained by DivideMix often suffers from the over-confidence issue based on Expected Calibration Error (ECE) (Guo et al. 2017), which is the most commonly used metric to quantify how well a deep learning model's predicted probabilities align with the actual probabilities of the events it predicts. Figure 2a illustrates the ECE results of DivideMix on the test data of CIFAR-10 with 50% symmetric noise. We can see that the confidence of the DivideMix model is much higher than the accuracy, *i.e.*, the model becomes overconfident and produces unreliable predicted probability.

To illustrate the intuition behind the phenomenon, we provide a specific case in Figure 1 where examples $\mathbf{x}_1$ with logits [0, 2, 0, 0] and $\mathbf{x}_2$ with logits [-2, 0, -2, -2] have equal probabilities of being predicted as "cat". Since the softmax function can only reflect the relative relationship among logits of different classes, it would not change the softmax value by adding/subtracting a constant from all logits. This "either-or" prediction strategy does not exhibit errors on common closed-set datasets with clean labels. However, it becomes unreliable when confronted with abnormal examples and may yield erroneous results. For example, $\mathbf{x}_2$ lacks typical "cat" characteristics compared to $\mathbf{x}_1$; still, it receives a high-confidence "cat" label due to its dissimilarity to other categories. In essence, the model's confidence in $\mathbf{x}_2$ isn't based on its clear "cat"-like features. In noisy label learning, we tend to identify the most likely clean examples and leverage them to assist in exploiting all the available data. Obviously, $\mathbf{x}_1$, with unmistakable "cat" features, is more likely to be a clean example and would be expected to have a larger probability than $\mathbf{x}_2$. This is also supported by Dempster–Shafer Theory of Evidence (DST) (Dempster 1968) that a smaller/larger logit means that there is less/more evidence to support belonging to that class.

Based on the preceding findings, we can conclude that the translation invariance of the softmax function would lead to the over-confidence phenomenon in noisy label learning, *e.g.*, potential mislabeled examples with logits like $\mathbf{x}_2$ may end up with probabilities equal to those of clean examples with logits like $\mathbf{x}_1$, despite their logits being significantly smaller than $\mathbf{x}_1$. This makes it challenging for the model to effectively perform label correction or example selection.

The Proposed Method

To meet the challenge, we first propose a *Dirichlet-based Prediction Calibration* (DPC) approach. As shown in Figure 1, by placing a suitable constant on the exponent term of the softmax function, we can easily break softmax's translation invariance and calibrate the model prediction. Compared to DivideMix, DPC can significantly reduce ECE, *i.e.* the red line is closer to the dashed line in Figure 2a, and avoid producing over-confident model predictions (Figure 2b). Meanwhile, we can see in Figure 2c that DPC gives a more specific meaning to the logit, where a logit greater/less than 0 indicates that there is evidence/no evidence for the example belonging to that class. Therefore, we drive a large-margin criterion for example partitioning. In Figure 2d, the AUC value of the large-margin criterion is larger than the small-loss criterion and achieves better distinguishable results.

Dirichlet-Based Prediction Calibration

As mentioned above, the softmax operator that converts the logit vector $\boldsymbol{o}$ to the probability vector $\boldsymbol{\rho}$ often leads the model predictions to be over-confident. To obtain more reliable output probabilities, we propose a calibrated softmax function and express the predicted probability for $\boldsymbol{x}_i$ as:

$$\hat{\rho}_{ic} = \frac{e^{o_{ic}} + \gamma}{\sum_{j=1}^{C} \left(e^{o_{ij}} + \gamma\right)}, \quad c = 1, 2, \ldots, C, \tag{1}$$

where γ is a constant. This simple transformation can break the translation invariance of the softmax function and can significantly alleviate the over-confidence issue. However, since the gradient difference before and after calibration is constantly greater than 0 on the complementary labels (labels other than the given label), the calibration leads the commonly used cross-entropy loss to suffer from the gradient shrinking issue, which can be demonstrated as:

$$\begin{aligned} &\frac{\partial \mathcal{L}_{ce|\boldsymbol{\rho}_i}}{\partial o_{ic}}|_{\boldsymbol{x}_i} - \frac{\partial \mathcal{L}_{ce|\hat{\boldsymbol{\rho}}_i}}{\partial o_{ic}}|_{\boldsymbol{x}_i} \\ &= \frac{\gamma C e^{o_{ic}}}{\sum_{j=1}^{C} e^{o_{ij}} \sum_{j=1}^{C} \left(e^{o_{ij}} + \gamma\right)} > 0, \ \forall y_{ic} = 0. \end{aligned} \tag{2}$$

This motivates us to design a specific loss to compress the probabilities of complementary labels. However, simply adding a regularization term to the model output (*e.g.*, KL divergence or L2 regularization) provides excessively strong constraints and can not achieve this while ensuring optimal model performance. Evidential deep learning (EDL) (Sensoy, Kaplan, and Kandemir 2018; Xie et al. 2023) regards the probability as a random variable and indirectly optimizes it by optimizing the parameters of the distribution, which is a softer constraint and can help us achieve this goal. Based on this, we propose a novel training strategy to seamlessly incorporate EDL with our calibrated softmax function. Below, we provide a detailed description of the training strategy.

EDL terms evidence $\boldsymbol{e}_i = g(\boldsymbol{o}_i)$ as a measure of the amount of support collected from data in favor of an example to be classified into a certain class, where $g(\cdot)$ is a function (*e.g.* exponential function) to ensure non-negative $\boldsymbol{e}_i$. Unlike traditional DNNs which give a point estimate of $\boldsymbol{\rho}$, EDL regards $\boldsymbol{\rho}$ as a random variable and places a Dirichlet distribution over $\boldsymbol{\rho}$ to represent the probability density of each possible $\boldsymbol{\rho}$. Specifically, for a given example $\boldsymbol{x}_i$, the probability density function of $\boldsymbol{\rho}_i$ is denoted as:

$$\begin{aligned} &p\left(\boldsymbol{\rho}_i | \boldsymbol{x}_i, \boldsymbol{\theta}\right) = Dir\left(\boldsymbol{\rho}_i | \boldsymbol{\alpha}_i\right) \\ &= \begin{cases} \frac{\Gamma\left(\sum_{j=1}^{C} \alpha_{ij}\right)}{\Pi_{j=1}^{C} \Gamma(\alpha_{ij})} \Pi_{j=1}^{C} \rho_{ij}^{\alpha_{ic}-1}, & \text{if } \boldsymbol{\rho}_i \in \triangle^C, \\ 0, & \text{otherwise}, \end{cases} \end{aligned}$$

where $\boldsymbol{\alpha}_i$ denotes the parameters of the Dirichlet distribution for $\boldsymbol{x}_i$, $\Gamma(\cdot)$ is the Gamma function and $\triangle^C = \left\{\boldsymbol{\rho}_i \mid \sum_{j=1}^{C} \rho_{ij} = 1 \text{ and } \forall j \ 0 \leq \rho_{ij} \leq 1\right\}$ is a C-dimensional unit simplex. Here, we define $\boldsymbol{e}_i = \gamma\left(\boldsymbol{\alpha}_i - 1\right)$. By marginalizing over $\boldsymbol{\rho}_i$, we can obtain the predicted probability for a given class c as:

$$\begin{aligned} P(y = c | \boldsymbol{x}_i, \boldsymbol{\theta}) &= \int p(y = c | \boldsymbol{\rho}_i) p(\boldsymbol{\rho}_i | \boldsymbol{x}_i, \boldsymbol{\theta}) d\boldsymbol{\rho}_i \\ &= \frac{\alpha_{ic}}{\sum_{j=1}^{C} \alpha_{ij}} = \frac{g(o_{ic}) + \gamma}{\sum_{j=1}^{C} \left(g(o_{ij}) + \gamma\right)}. \end{aligned} \tag{3}$$

Specifically, we can bridge the connection with Equation 1 by defining $g(\cdot)$ as an exponential function. Following that, we train the EDL model by driving it to produce a sharp Dirichlet distribution situated at the corner of $\triangle^C$ for all labeled data. To ensure this, on one hand, we minimize the negative logarithm of the marginal likelihood ($\mathcal{L}_{nll}$) to ensure the correctness of prediction:

$$\begin{aligned} \mathcal{L}_{nll} &= -\frac{1}{N} \sum_{i=1}^{N} \log\left[P(y = c | \boldsymbol{x}_i, \boldsymbol{\theta})\right] \\ &= \frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{ic} \left[\log\left(\sum_{j=1}^{C} \alpha_{ij}\right) - \log \alpha_{ic}\right]. \end{aligned} \tag{4}$$

On the other hand, to cope with the problem posed by Equation 2 and regularize the predictive distribution, a KL-divergence term $\mathcal{L}_{kl}$ is adopted by penalizing the evidence of the complementary labels to approach 0:

$$\begin{aligned} \mathcal{L}_{kl} &= \frac{1}{NC} \sum_{i=1}^{N} D_{KL}\left(Dir\left(\boldsymbol{\rho}_i | \tilde{\boldsymbol{\alpha}}_i\right) \parallel Dir\left(\boldsymbol{\rho}_i | \mathbf{1}\right)\right) \\ &= \frac{1}{NC} \sum_{i=1}^{N} \left[\log\left[\frac{\Gamma\left(\sum_{j=1}^{C} \tilde{\alpha}_{ij}\right)}{\Gamma(C) \Pi_{j=1}^{C} \Gamma\left(\tilde{\alpha}_{ij}\right)}\right] \right. \\ &\quad \left. + \sum_{c=1}^{C} \left(\tilde{\alpha}_{ic} - 1\right) \left[\psi\left(\tilde{\alpha}_{ic}\right) - \psi\left(\sum_{j=1}^{C} \tilde{\alpha}_{ij}\right)\right]\right], \end{aligned} \tag{5}$$

where $\tilde{\boldsymbol{\alpha}}_i = \boldsymbol{y}_i + (1 - \boldsymbol{y}_i) \odot \boldsymbol{\alpha}_i$ can be seen as the Dirichlet parameters after removal of the given label evidence from predicted parameters $\boldsymbol{\alpha}$, $\mathbf{1}$ is a vector consisting of C ones and $\psi(\cdot)$ represents the digamma function. Then, the overall training loss can be formulated as:

$$\mathcal{L}_{edl} = \mathcal{L}_{nll} + \beta \mathcal{L}_{kl},$$

where β is used to balance the two terms.

Large-Margin Example Selection Criterion

To achieve the same predicted probability on a single example, the calibrated softmax needs to provide a logit distribution with greater differentiation than the commonly used softmax function. This leads the proposed calibration method to produce output logits distribution more distinguishable. Thus, we propose a large-margin example selection criterion and define the margin for a given example $\boldsymbol{x}_i$ as the difference of predicted logits between the given class and the largest probable classes of complementary labels:

$$margin\left(\boldsymbol{x}_i\right) = o_{ic} - \max_{j \neq c} o_{ij}, \text{ where } c = arg\max \boldsymbol{y}_i.$$

A larger margin yields that the model is more confident that the example belongs to the class c. Unlike the small-loss

Method \ Dataset	CIFAR-10						CIFAR-100					
	Symmetric			Asymmetric			Symmetric			Asymmetric		
	20%	50%	80%	10%	30%	40%	20%	50%	80%	10%	30%	40%
Cross-Entropy	88.8	81.7	76.1	88.8	81.7	76.1	61.8	37.3	8.8	68.1	53.3	44.5
Mixup (Zhang et al. 2017)	93.3	83.3	77.7	93.3	83.3	77.7	72.4	57.6	48.1	72.4	57.6	48.1
PENCIL (Yi and Wu 2019)	92.4	89.1	77.5	93.1	92.9	91.6	69.4	57.5	31.1	76.0	59.3	48.3
JPL (Kim et al. 2021)	93.5	90.2	35.7	94.2	92.5	90.7	70.9	67.7	17.8	72.0	68.1	59.5
DivideMix (Li, Socher, and Hoi 2020)	95.7	94.4	92.9	93.8	92.5	91.7	76.9	74.2	59.6	71.6	69.5	55.1
PES (Bai et al. 2021)	95.9	95.1	93.1	-	-	-	77.4	74.3	61.6	-	-	-
ELR+ (Liu et al. 2020)	95.8	94.8	93.3	95.4	**94.7**	93.0	77.6	73.6	60.8	77.3	74.6	73.2
MOIT+ (Ortego et al. 2021)	94.1	91.1	75.8	94.2	94.1	93.2	75.9	70.1	51.4	77.4	75.1	74.0
DPC	**96.1**	**95.2**	**93.5**	**95.5**	94.5	**93.6**	**79.4**	**76.5**	**63.0**	**79.0**	**77.6**	**74.1**
UniCon (Karim et al. 2022)	96.0	95.6	93.9	95.3	94.6	94.1	78.9	77.6	63.9	78.2	75.6	74.8
DPC*	**96.5**	**95.9**	**94.8**	**95.4**	**95.3**	**94.9**	**81.0**	**78.5**	**66.4**	**80.5**	**79.7**	**75.6**

Table 1: Test accuracy (%) comparison with state-of-the-art methods on CIFAR-10 and CIFAR-100 with synthetic noise. For previous techniques, results are copied from their respective papers. For our method, results are reported over 3 random runs.

criterion, which requires a suitable loss function for example selection, our large-margin criterion only relates to the model itself without introducing any external information.

After obtaining margin values for all training examples, we fit a two-component GMM to model per-example margin distribution and use its prediction on the Gaussian component with a larger mean (larger margin) to divide examples.

Combining with Semi-Supervised Learning

Following the partitioning of the training set $\mathcal{D}$ into the clean subset $\mathcal{X}$ and the mislabeled subset $\mathcal{U}$, we adopt MixMatch (Berthelot et al. 2019) as the semi-supervised learning framework for the subsequent training similar to (Li, Socher, and Hoi 2020; Karim et al. 2022).

Specifically, for a pair of examples $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ in $\mathcal{X} \cup \mathcal{U}$ with their given labels $(\boldsymbol{y}_1, \boldsymbol{y}_2)$ and model predictions $(\boldsymbol{\rho}_1, \boldsymbol{\rho}_2)$, we can obtain the mixed $(\boldsymbol{x}', \boldsymbol{y}', \boldsymbol{\rho}')$ by:

$$\lambda \sim Beta\left(\alpha\right), \lambda' = \max\left(\lambda, 1-\lambda\right),$$

$$\boldsymbol{t}' = \lambda' \boldsymbol{t}_1 + (1-\lambda')\boldsymbol{t}_2, \text{ where } \boldsymbol{t} \in \{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\rho}\}.$$

We can thus drive a mixed clean set $\mathcal{X}'$ and a mixed mislabeled set $\mathcal{U}'$. To deeply integrate our method with MixMatch, for a given example $(\boldsymbol{x}', \boldsymbol{y}')$ in $\mathcal{X}'$, the supervised loss $\mathcal{L}_{sup}$ is defined as:

$$\mathcal{L}_{sup}\left(\boldsymbol{x}', \boldsymbol{y}'\right) = \lambda' \mathcal{L}_{edl}\left(\boldsymbol{x}_1, \boldsymbol{y}_1\right) + (1-\lambda')\mathcal{L}_{edl}\left(\boldsymbol{x}_2, \boldsymbol{y}_2\right).$$

The unsupervised loss $\mathcal{L}_{uns}$ for another given example $(\boldsymbol{x}', \boldsymbol{\rho}')$ in $\mathcal{U}'$ is defined as:

$$\mathcal{L}_{uns}\left(\boldsymbol{x}', \boldsymbol{\rho}'\right) = \left\|\boldsymbol{\rho}' - \boldsymbol{\rho}(f(\boldsymbol{x}'; \boldsymbol{\theta}))\right\|_2^2.$$

Eventually, by using the balancing factor λ_{uns} to control the strength of $\mathcal{L}_{uns}$, the overall loss can be expressed as:

$$\mathcal{L}_{total} = \sum_{\mathcal{X}'} \mathcal{L}_{sup}(\boldsymbol{x}', \boldsymbol{y}') + \lambda_{uns} \sum_{\mathcal{U}'} \mathcal{L}_{uns}(\boldsymbol{x}', \boldsymbol{\rho}').$$

Note that previous works (Li, Socher, and Hoi 2020; Karim et al. 2022) calculate the supervised loss directly based on $\boldsymbol{x}'$, while our supervised loss is indirectly obtained by combining $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. This can ensure that models give a positive and sufficiently large logit on the given label and penalize other labels with a negative and as small as possible logit. Considering that the mixed examples with unsupervised loss may break this rule and drive the model to provide smooth outputs, we adopt a two-head network architecture, where one head is allocated for supervised loss and the other for unsupervised loss. Detailed derivation of Equation 2,3,4 and 5 and more detailed implementations are available in the supplementary file.

Experiments

Datasets. We experimentally demonstrate the effectiveness of DPC on both synthetic noise datasets (CIFAR-10 and CIFAR-100 (Krizhevsky, Hinton et al. 2009)) and real-world noise datasets (CIFAR-10N, CIFAR-100N (Wei et al. 2021) and WebVision (Li et al. 2017)). All the CIFAR datasets contain 50K training images and 10K test images of size 32×32. For CIFAR-10 and CIFAR-100, we manually inject two types of label noise: symmetric and asymmetric noise, where the noise rate is set to 20%, 50%, and 80% for symmetric noise and 10%, 30%, and 40% for asymmetric noise. CIFAR-10/100N is a re-annotation version of CIFAR-10/100 by human workers. Specifically, each image in CIFAR-10N owns three submitted labels denoted as "rand1", "rand2", and "rand3", and two ensembled labels denoted as "aggre", and "worst". While in CIFAR-100N, each image only contains a single submitted label represented as "noisy100". WebVision comprises 2.4 million images obtained from web crawling using 1K concepts included in ImageNet ILSVRC12. Here, following the previous studies (Li, Socher, and Hoi 2020; Karim et al. 2022), we only use the first 50 classes of the Google image subset to construct the training set.

Training Details. For CIFAR-10 and CIFAR-100, we use PreAct ResNet18 (He et al. 2016b) as the base model and train it by stochastic gradient descent (SGD) optimizer with momentum 0.9, weight decay 0.0005, and batch size 128 for 300 epochs. The initial learning rate is set to 0.02 and reduced by a factor of 10 after 150 epochs. The warm-

Method \ Dataset	CIFAR-10					CIFAR-100
	aggre	rand1	rand2	rand3	worst	noisy100
Cross-Entropy	87.77±0.38	85.02±0.65	86.46±1.79	85.16±0.61	77.69±1.55	55.50±0.66
GCE (Zhang and Sabuncu 2018)	87.85±0.70	87.61±0.28	87.70±0.56	87.58±0.29	80.66±0.35	56.73±0.30
Co-teaching (Han et al. 2018)	91.20±0.13	90.33±0.13	90.30±0.17	90.15±0.18	83.83±0.13	60.37±0.27
PES (Bai et al. 2021)	94.66±0.18	95.06±0.15	95.19±0.23	95.22±0.13	92.68±0.22	70.36±0.33
ELR+ (Liu et al. 2020)	94.83±0.10	94.43±0.41	94.20±0.24	94.34±0.22	91.09±1.60	66.72±0.07
CORES (Cheng et al. 2020)	95.25±0.09	94.45±0.14	94.88±0.31	94.74±0.03	91.66±0.09	61.15±0.73
DivideMix (Li, Socher, and Hoi 2020)	95.01±0.71	95.16±0.19	95.23±0.07	95.21±0.14	92.56±0.42	71.13±0.48
SOP (Liu et al. 2022)	95.61±0.13	95.28±0.13	95.31±0.10	95.39±0.11	93.24±0.21	67.81±0.23
DPC	**95.77±0.23**	**95.97±0.07**	**95.92±0.14**	**95.90±0.09**	**93.82±0.31**	**71.42±0.23**

Table 2: Comparison with state-of-the-art methods in test accuracy (%) on CIFAR-N. The corresponding noise rate is "aggre" (9.03%), "rand1" (17.23%), "rand2" (18.12%), "rand3" (17.64%), "worst" (40.21%) and "noisy100" (40.20%). The results of comparing methods are copied from (Wei et al. 2021). The results (mean±std) of our method are reported over 5 random runs.

up epoch is 10 for CIFAR-10 and 30 for CIFAR-100. For CIFAR-10N and CIFAR-100N, ResNet34 (He et al. 2016a) is adopted and the warm-up epoch is changed to 30 for CIFAR-10 and 40 for CIFAR-100, respectively. For WebVision, we train InceptionResNetV2 (Szegedy et al. 2017) from scratch with changed parameters of batch size 32, warm-up epoch 1, and training epoch 100. The initial learning rate is changed to 0.01 and reduced by a factor of 10 after 50 epochs. For all experiments, we set β as 0.5 and have $\gamma = \frac{10}{C}$.

Note that methods employing stronger data augmentation techniques or integrating advanced technologies like self-supervision consistently push the limits of state-of-the-art performance. Our approach, however, operates orthogonally with these advanced techniques and is expected to yield improved performance in combined with them. In this paper, we mainly compare with state-of-the-art methods not using these techniques. Additionally, we compare with UniCon (Karim et al. 2022) to verify the effectiveness of DPC when adopting strong data augmentation for training. For a fair comparison, we apply a consistency regularization term to implement the same strong data augmentation for our method (denoted as DPC*). More detailed implementations can be found in the supplementary file.

Experimental Results

Table 1 shows the averaged test accuracy over the last 10 epochs on CIFAR-10 and CIFAR-100 with different levels of symmetric and asymmetric label noise. For methods without strong augmentation, DPC achieves the best performance in most cases among all the compared methods. Especially for CIFAR-100 with symmetric noise 20% and 50%, DPC gains at least 1.8% performance improvement. For methods with strong augmentation, we can see that DPC* still outperforms UniCon by a large margin across all noise ratios. These experiments not only confirm that the over-confidence phenomenon of the model is detrimental to the task of learning with noisy labels, but also demonstrate the effectiveness of our proposed method.

Table 2 reports the final round accuracy on CIFAR-10N and CIFAR-100N with realistic label noise. Although the model architecture is changed in this setting, our method still maintains a performance gain over all comparing methods.

Method	WebVision		ILSVRC12	
	top1	top5	top1	top5
D2L (Ma et al. 2018)	62.7	84.0	57.8	81.4
MentorNet (Jiang et al. 2018)	63.0	81.4	57.8	79.9
Co-teaching (Han et al. 2018)	63.6	85.2	61.5	84.7
ICV (Chen et al. 2019)	65.2	85.3	61.6	85.0
ELR+	77.8	91.7	70.3	89.8
DivideMix	77.3	91.6	75.2	90.8
DPC	**79.2**	**93.0**	**75.8**	**92.5**
UniCon	77.6	93.4	75.3	93.7
DPC*	**81.1**	**93.5**	**78.0**	**93.8**

Table 3: Accuracy comparison on the WebVision validation set and the ImageNet ILSVRC12 validation set.

We further report the performance of our method on another real-world noise dataset WebVision. As shown in Table 3, DPC and DPC* both achieve the best results on the WebVision validation set and the ImageNet ILSVRC12 validation set.

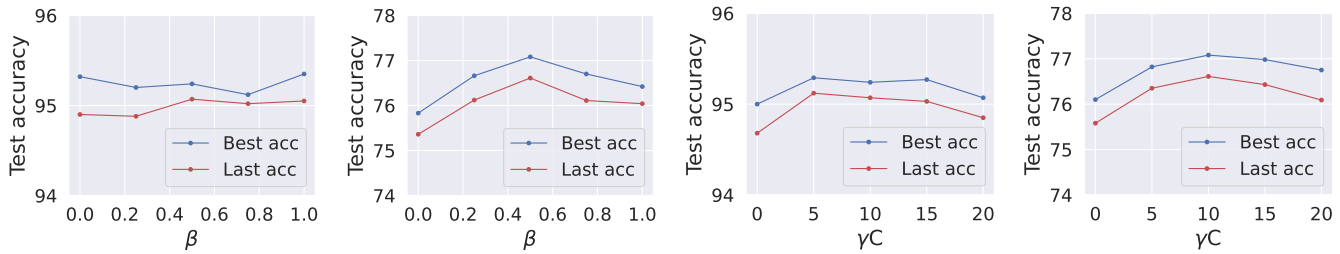
These experiments demonstrate that our method is equally effective for both synthetic and real-world label noise, and consistently performs well across different models.

Ablation Studies

Effect of $\mathcal{L}_{edl}$. Table 4 indicates the impact of the proposed Dirichlet training scheme (denoted as $\mathcal{L}_{edl}$) on the overall performance of DPC. We can see that the performance of DPC without $\mathcal{L}_{edl}$ has a significant drop at various noise rates. This suggests that the softmax-based predicted probabilities do have bias, and it is necessary to calibrate existing noisy label learning methods. Furthermore, to verify the generalizability of the Dirichlet-based prediction calibration method, we also provide the results of UniCon integrated with $\mathcal{L}_{edl}$ in Table 4. As we can see, the proposed method still works effectively and has a certain degree of generality.

Dataset	CIFAR-10						CIFAR-100					
Noise Rate	20%		50%		80%		20%		50%		80%	
Method	Best	Last	Best	Last	Best	Last	Best	Last	Best	Last	Best	Last
DPC w/o $\mathcal{L}_{edl}$	96.1	96.0	94.9	94.7	93.3	93.0	79.4	78.6	75.6	75.1	60.3	60.0
DPC w/o LM	96.3	96.1	95.1	94.8	93.3	93.1	77.8	77.6	76.5	76.0	57.4	57.1
DPC w/o TH	96.3	96.1	95.2	94.9	92.8	92.5	79.8	79.6	76.5	76.1	62.8	62.4
DPC	96.3	96.1	95.3	95.2	93.6	93.5	79.9	79.4	76.9	76.5	63.3	63.0
UniCon	94.2	93.4	95.6	94.9	94.2	93.7	79.0	77.0	77.1	75.9	65.0	64.1
UniCon w $\mathcal{L}_{edl}$	94.6	93.8	95.5	94.8	94.2	93.7	79.5	77.8	77.5	76.0	65.4	64.4
Improve	$\uparrow$ 0.4	$\uparrow$ 0.4	$\downarrow$ 0.1	$\downarrow$ 0.1	$\uparrow$ 0.0	$\uparrow$ 0.0	$\uparrow$ 0.5	$\uparrow$ 0.8	$\uparrow$ 0.4	$\uparrow$ 0.1	$\uparrow$ 0.4	$\uparrow$ 0.3

Table 4: Ablation studies on CIFAR-10 and CIFAR-100 with symmetric noise. LM indices the large-margin example selection criterion. TH means a model with two classification heads. Here, we show the rerun results for UniCon.



(a) (b) (c) (d)

Figure 3: Ablation studies on CIFAR-10 and CIFAR-100 with a 50% symmetric noise rate, respectively. (a) and (b) are the ablation studies of β on CIFAR-10 and CIFAR-100. (c) and (d) are the ablation studies of γ on CIFAR-10 and CIFAR-100.

Note that we have also provided integration results of two other methods with $\mathcal{L}_{edl}$ in the supplementary materials, and the results are consistent.

Effect of Large-Margin Criterion. DPC without large-margin criterion (denoted as LM), *i.e.* with small-loss criterion, also shows unsatisfactory performance (see Table 4). This is because the two items in $\mathcal{L}_{edl}$ have different scales. For CIFAR-100 with noise ratio 80%, $\mathcal{L}_{kl}$ is oscillating and provides unstable loss estimations. The proposed criterion can counter this since it is independent of a specific loss.

Effect of Two Classification Heads. The performance gap between one and two classification heads (denoted as TH) becomes large with the increase of noise rate as shown in Table 4. $\mathcal{L}_{uns}$ tends to make the model produce smooth logit vectors, while $\mathcal{L}_{sup}$ encourages the model to provide distinguishable logit vectors. From an example selection perspective, $\mathcal{L}_{uns}$ and $\mathcal{L}_{sup}$ are in conflict especially under high noise settings since a large noise rate commonly corresponds to a large λ_{uns}.

Hyper-Parameter Sensitivity of β. This parameter controls the tradeoff between $\mathcal{L}_{nll}$ and $\mathcal{L}_{kl}$. We test the sensitivity of β on CIFAR-10 and CIFAR-100 with a 50% symmetric noise rate, respectively. The results are shown in Figure 3a and 3b, where $\beta \in \{0, 0.25, 0.5, 0.75, 1\}$. Since the CIFAR-10 task is relatively simple, the model is less sensitive to the value of β. However, for CIFAR-100, different values of β have a significant impact on the accuracy. Finally, we determined the value of β to be 0.5 and used this setting in all experiments.

Hyper-Parameter Sensitivity of γ. This parameter is the placed constant in the calibrated softmax function. Figure 3c and 3d present the results on CIFAR-10 and CIFAR-100 under 50% symmetric noise rate, where $\gamma C \in \{0, 5, 10, 15, 20\}$. According to the results, we recommend setting γC as 10 for all experiments.

Conclusion

In this paper, we propose DPC to combat label noise by calibrating the predicted probability involved in the example selection and label correction procedure. DPC consists of two important components: Dirichlet-based prediction calibration and large-margin example selection criterion. The former includes a calibrated softmax function that can convert more accurate predicted probability from logit and a corresponding Dirichlet-based model training loss to ensure sufficient training. Through this, the model can avoid outputting over-confident predictions, and produce more distinguishable logit outputs which drive us to propose the latter, *i.e.* the large-margin example selection criterion. Extensive experiments across multiple datasets demonstrate that our proposed method consistently exhibits substantial performance gains compared to the state-of-the-art methods.

Acknowledgments

This work was supported by the National Key R&D Program of China (2020AAA0107000), the Natural Science Foundation of Jiangsu Province of China (BK20222012, BK20211517), and NSFC (62222605).

References

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. *Advances in Neural Information Processing Systems*, 34: 24392–24403.

Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; and Raffel, C. A. 2019. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32.

Chang, H.-S.; Learned-Miller, E.; and McCallum, A. 2017. Active bias: Training more accurate neural networks by emphasizing high variance samples. *Advances in Neural Information Processing Systems*, 30.

Chen, P.; Liao, B. B.; Chen, G.; and Zhang, S. 2019. Understanding and utilizing deep neural networks trained with noisy labels. In *International Conference on Machine Learning*, 1062–1070. PMLR.

Chen, Z.; Fu, R.; Zhao, Z.; Liu, Z.; Xia, L.; Chen, L.; Cheng, P.; Cao, C. C.; Tong, Y.; and Zhang, C. J. 2014. gmission: A general spatial crowdsourcing platform. *Proceedings of the VLDB Endowment*, 7(13): 1629–1632.

Cheng, H.; Zhu, Z.; Li, X.; Gong, Y.; Sun, X.; and Liu, Y. 2020. Learning with instance-dependent label noise: A sample sieve approach. *arXiv preprint arXiv:2010.02347*.

Dempster, A. P. 1968. A generalization of Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2): 205–232.

Garcia, L. P.; de Carvalho, A. C.; and Lorena, A. C. 2016. Noise detection in the meta-learning level. *Neurocomputing*, 176: 14–25.

Goldberger, J.; and Ben-Reuven, E. 2017. Training deep neural-networks using a noise adaptation layer. In *International conference on learning representations*.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International conference on machine learning*, 1321–1330. PMLR.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016a. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016b. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, 630–645. Springer.

Hossain, M.; and Kauranen, I. 2015. Crowdsourcing: a comprehensive literature review. *Strategic Outsourcing: An International Journal*, 8(1): 2–22.

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*, 3326–3334.

Huang, L.; Zhang, C.; and Zhang, H. 2020. Self-adaptive training: beyond empirical risk minimization. *Advances in neural information processing systems*, 33: 19365–19376.

Huang, S.-J.; Zong, C.-C.; Ning, K.-P.; and Ye, H.-B. 2021. Asynchronous Active Learning with Distributed Label Querying. In *IJCAI*, 2570–2576.

Jiang, L.; Zhou, Z.; Leung, T.; Li, L.-J.; and Fei-Fei, L. 2018. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International conference on machine learning*, 2304–2313. PMLR.

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*, 9676–9686.

Kim, Y.; Yun, J.; Shon, H.; and Kim, J. 2021. Joint negative and positive learning for noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9442–9451.

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

Li, J.; Socher, R.; and Hoi, S. C. 2020. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*.

Li, W.; Wang, L.; Li, W.; Agustsson, E.; and Van Gool, L. 2017. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*.

Liu, S.; Niles-Weed, J.; Razavian, N.; and Fernandez-Granda, C. 2020. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33: 20331–20342.

Liu, S.; Zhu, Z.; Qu, Q.; and You, C. 2022. Robust training under label noise by over-parameterization. In *International Conference on Machine Learning*, 14153–14172. 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 *International conference on machine learning*, 6543–6553. PMLR.

Ma, X.; Wang, Y.; Houle, M. E.; Zhou, S.; Erfani, S.; Xia, S.; Wijewickrema, S.; and Bailey, J. 2018. Dimensionality-driven learning with noisy labels. In *International Conference on Machine Learning*, 3355–3364. PMLR.

Malach, E.; and Shalev-Shwartz, S. 2017. Decoupling" when to update" from" how to update". *Advances in neural information processing systems*, 30.

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, 6606–6615.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1944–1952.

Permuter, H.; Francos, J.; and Jermyn, I. 2006. A study of Gaussian mixture models of color and texture features for image classification and segmentation. *Pattern recognition*, 39(4): 695–706.

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*.

Sensoy, M.; Kaplan, L.; and Kandemir, M. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31.

Shu, J.; Xie, Q.; Yi, L.; Zhao, Q.; Zhou, S.; Xu, Z.; and Meng, D. 2019. Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in neural information processing systems*, 32.

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

Wang, D.-B.; Feng, L.; and Zhang, M.-L. 2021. Rethinking calibration of deep neural networks: Do not be afraid of overconfidence. *Advances in Neural Information Processing Systems*, 34: 11809–11820.

Wei, J.; Zhu, Z.; Cheng, H.; Liu, T.; Niu, G.; and Liu, Y. 2021. Learning with noisy labels revisited: A study using real-world human annotations. *arXiv preprint arXiv:2110.12088*.

Xie, M.; Li, S.; Zhang, R.; and Liu, C. H. 2023. Dirichlet-based Uncertainty Calibration for Active Domain Adaptation. *arXiv preprint arXiv:2302.13824*.

Yi, K.; and Wu, J. 2019. Probabilistic end-to-end noise correction for learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7017–7025.

Yu, X.; Han, B.; Yao, J.; Niu, G.; Tsang, I.; and Sugiyama, M. 2019. How does disagreement help generalization against label corruption? In *International Conference on Machine Learning*, 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.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhang, Z.; and Sabuncu, M. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31.

Zhou, T.; Wang, S.; and Bilmes, J. 2021. Robust curriculum learning: from clean label detection to noisy label self-correction. In *International Conference on Learning Representations*.

Zong, C.-C.; Cao, Z.-T.; Guo, H.-T.; Du, Y.; Xie, M.-K.; Li, S.-Y.; and Huang, S.-J. 2022. Noise-Robust Bidirectional Learning with Dynamic Sample Reweighting. *arXiv preprint arXiv:2209.01334*.