Sample-wise Label Confidence Incorporation for Learning with Noisy Labels

Chanho Ahn* Kikyung Kim Ji-won Baek Jongin Lim Seungju Han
Samsung Advanced Institute of Technology (SAIT), Korea
{chanho.ahn, kk87.kim, jw0328.baek, jonny.lim, sj75.han}@samsung.com

Abstract

Deep learning algorithms require large amounts of labeled data for effective performance, but the presence of noisy labels often significantly degrade their performance. Although recent studies on designing a robust objective function to label noise, known as the robust loss method, have shown promising results for learning with noisy labels, they suffer from the issue of underfitting not only noisy samples but also clean ones, leading to suboptimal model performance. To address this issue, we propose a novel learning framework that selectively suppresses noisy samples while avoiding underfitting clean data. Our framework incorporates label confidence as a measure of label noise, enabling the network model to prioritize the training of samples deemed to be noise-free. The label confidence is based on the robust loss methods, and we provide theoretical evidence that our method can reach the optimal point of the robust loss, subject to certain conditions. Furthermore, the proposed method is generalizable and can be combined with existing robust loss methods, making it suitable for a wide range of applications of learning with noisy labels. We evaluate our approach on both synthetic and real-world datasets, and the experimental results demonstrate its effectiveness in achieving outstanding classification performance compared to state-of-the-art methods.

1. Introduction

Recent advances in deep learning have led to remarkable image classification performance that surpasses human ability [8, 9]. However, the challenge is that deep learning models require a substantial amount of labeled data to maintain their desired performance. Image classification benchmark datasets often contain tens of thousands [14] to millions [24] of labeled data. Despite extensive efforts to curate these datasets, mislabeling is inevitable due to the complexity of samples and errors made by experts [1]. This issue is particularly problematic given the capacity of deep networks to overfit data samples [35]. Consequently, preventing overfitting on mislabeled data is a critical challenge.

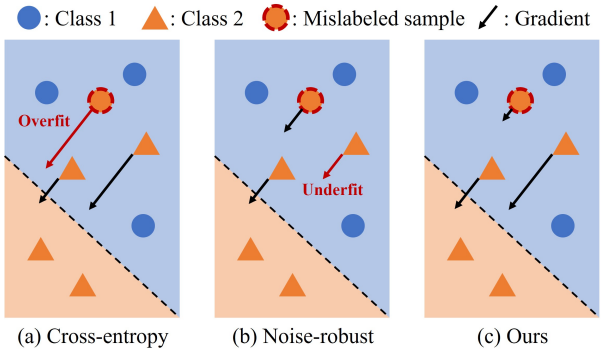


Figure 1. **Comparison with robust loss functions against to label noise.** (a) Cross-entropy loss cannot prevent the learning of mislabeled samples, as it imposes a stronger penalty on predictions that are more misaligned. (b) Noise-robust loss prevents the model from learning mislabeled samples by suppressing strong penalties, but it carries the risk of underfitting. (c) Ours adjusts the penalty based on the confidence of each sample, imposing a strong penalty only on samples with high confidence in their labels.

To tackle this challenge, researchers have focused on developing methods for *learning with noisy labels* in deep learning frameworks. These methods can be categorized into three topics: noise estimation, sample selection, and robust loss. While noise estimation methods [6, 10, 22, 25] assume the availability of prior knowledge about the noise model, obtaining this information in real-world practice can be challenging. Sample selection methods [7, 12, 31, 34, 13] aim to eliminate noisy labels in the dataset and train the network model on the refined set. However, the success of these methods heavily depends on the quality of noise elimination, and the human-designed criteria for configuring the refined set may not generalize well across various datasets [33]. Robust loss methods [4, 5, 30, 37], on the other hand, aim to design loss functions that are theoretically less affected by label noise. However, these methods can also encourage underfitting, resulting in lower performance.

This paper presents a novel training strategy aimed at addressing the underfitting problem commonly encountered

*Corresponding author

by robust loss methods. Cross-entropy loss, a widely-used objective function for image classification tasks, is highly responsive to incorrect predictions. This is because both the loss value and gradient magnitude increase when the predicted probability for a given label decreases. Conversely, robust loss methods reduce the penalty for erroneous predictions to prevent the model from learning mislabeled samples. However, batch-based stochastic gradient descent can often impede convergence to the optimal point of the robust loss function, leading to underfitting issues. To alleviate this challenge, we propose a sample-wise label confidence incorporation into our method. This incorporation leads to reduced penalties for inaccurate predictions of noisy samples with lower confidence levels. The proposed approach involves training using a weighted cross-entropy loss, where the weight is determined based the label confidence. A visual representation of this methodology is in Figure 1.

Furthermore, we also offer theoretical evidence that our proposed training strategy is capable of attaining the optimal point of the robust loss method, subject to certain conditions being satisfied regarding the sample-wise label confidence. Specifically, the calculation of the label confidence is based on the existing robust loss method, and our proposed method approximates the optimal point of the selected robust loss method. Importantly, it's worth noting that our method does not introduce a novel loss function that inherently accounts for label noise robustness. Instead, it provides a training strategy to mitigate the underfitting issue of existing robust loss methods like MAE [5], GCE [37], and JS [4]. Thus, our strategy holds the potential for universal applicability. Additionally, in contrast to sample selection methods that often require the hard-tuning of hyperparameters based on human-designed criteria, our method computes the label confidence by simply selecting the appropriate robust loss.

Our research derives label confidence for a dataset consisting primarily of normal samples, inspired by an application of the robust loss [21]. While the previous application employed a robust loss model to identify unbiased images by learning features that counteract bias, our proposed method employs a robust loss model that effectively learns the majority of normal samples. This approach prevents the imposition of excessive penalties on mislabeled samples. The determination of label confidence plays a pivotal role in achieving this objective. In order for the theoretical framework to hold, the label confidence must satisfy two critical conditions: (1) a strict negative correlation between the label confidence and the robust loss value exists, and (2) samples with a robust loss value surpassing a specific threshold must converge towards a label confidence of 0. These conditions are intuitive and align with our assumption that samples with low robust loss values are less likely to be noisy samples.

In summary, the proposed approach offers a novel learning strategy to address the underfitting problem of robust loss methods, which can be combined with existing robust loss methods in a versatile manner. Moreover, our theoretical findings emphasize the potential for designing label confidence to adhere to two crucial conditions. This insight can serve as valuable guidance for future research endeavors aimed at calculating label confidence through alternative approaches. Finally, we evaluate the performance of the proposed method on synthetic datasets and real-world datasets. On synthetic datasets, our method significantly outperforms existing robust loss methods, particularly as the noise ratio increases. In addition, our method shows consistent performance across different hyperparameters, which is a significant advantage over existing learning with noisy label methods that require hard-tuning. On real-world datasets, our method also achieves state-of-the-art performance, demonstrating its effectiveness in real-world scenarios. Our contributions are three-fold:

- The proposed approach reduces underfitting issues of robust loss methods by incorporating sample-wise label confidence, leading to improved convergence to the optimal point of the robust loss function.
- The proposed method is generally applicable to robust loss methods, and label confidence can be simply computed by selecting the appropriate robust loss.
- Theoretical evidence supports the proposed method by providing guidelines for designing label confidence that approximates the optimal point of the selected robust loss method, subject to certain conditions being satisfied regarding sample-wise label confidence.

2. Related Works

Early studies on learning with noisy labels [6, 10, 22, 25, 29, 32] assume that prior knowledge about the noise rate or clean data is available. Since we address a more practical problem where this prior knowledge is not given, these methods are not the focus of this paper.

Classical studies proposing robust loss functions against label noise replace a given label with the smoothed label [28] or the network prediction [23]. The Mean Absolute Error (MAE) [5] proposes a symmetric loss with a constant sum of losses for all labels for each sample to provide theoretical validity. However, MAE has a significant underfitting problem, which is highly correlated with its poor performance. Recent methods [37] design a relaxed symmetric loss or a linear combination of existing loss functions [19, 30] to alleviate the underfitting effects of the symmetric loss. While they attempt to avoid the underfitting problem, these methods are still reported to have an underfitting issue compared to the cross-entropy loss [4]. Even the

method [4] of addressing the underfitting problem proposes augmentation-invariant regularization and does not directly solve the underfitting problem.

Sample selection methods refine the given dataset and train the network model on the refined set to address the noisy label problem. MentorNet [12] employs a pre-trained teacher network to identify clean samples. Decoupling [20] and Co-teaching+ [34] train two models jointly and compare their predictions to determine the samples for training each model. While Decoupling and Co-teaching+ focus on identifying prediction disagreements, JoCoR [31] encourages the agreement of predictions. CAR [18] combines the provided labels with early learning predictions. Despite their compromised performance, the human-designed proportion of the refined set and the sample selection within the mini-batch require complex tuning that is dependent on the dataset, which limits their general applicability [33].

Other studies [26, 33] on learning with noisy labels address the open-set noisy label problem, which can occur in data collected based on web searching. They design the model to find out-of-distribution samples to show satisfactory performance on the benchmark dataset. Another noisy label method [11] finds that Mixup [36] augmentation can be an appropriate regularization for learning with noisy labels. However, the methods that use prior information that out-of-distribution samples are in the dataset or use a specific augmentation are outside the scope of this paper.

3. Proposed Method

Preliminaries. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ denote a clean dataset, where $x_i \in \mathcal{X}$ and $y_i \in (0, c]$ are the i-th image and its corresponding true label, respectively, $\mathcal{X}$ is the image space, and c is the number of classes. Given a classifier parameterized by θ denoted as $f(\cdot;\theta) : \mathcal{X} \to \mathbb{R}^c$, the empirical risk of classification commonly used in deep learning is defined as $\mathcal{R}_{\mathcal{L}}(\theta;\mathcal{D}) = \mathbb{E}_{(x,y)\in\mathcal{D}}[\mathcal{L}(f(x;\theta), y)]$.

In [5], Ghosh *et al.* defines a symmetric loss that satisfies the condition that the sum of the loss function for all labels is constant, as follows: $\sum_{i=1}^{c} \mathcal{L}(f(x;\theta), i) = \text{Const.}$ for $\forall x$. Then, the minimizer of $\mathcal{R}_{\mathcal{L}}(\cdot;\mathcal{D})$ is also the minimizer of $\mathcal{R}_{\mathcal{L}}(\cdot;\mathcal{D}_n)$, where $\mathcal{D}_n$ is the dataset containing noisy labels with a noise ratio less than a certain threshold [5].

However, the critical underfitting problem of the symmetric loss is reported in [37]. To address this issue, recent studies propose a loss function that relaxes the condition of the symmetric loss that the sum of the loss function for all labels is bounded [4, 37] or define the loss as a linear combination of various loss functions [19, 30]. Nevertheless, the noise-robust loss functions still suffer from slow convergence and underfitting compared to the cross-entropy loss [4]. In this paper, we refer to the relaxed symmetric loss proposed in [19, 30, 37] as the *noise-robust loss*.

3.1. Label Confidence Incorporation

To address the underfitting issue of the noise-robust loss, we introduce the utilization of cross-entropy loss. Our investigations reveal that simple combinations, such as a linear combination of cross-entropy loss and noise-robust loss, significantly deteriorate classification performance (as represented in Figure 4 and Section 4.1). This performance degradation can be attributed to the inherent nature of cross-entropy loss. The cross-entropy loss is learned to predict high probabilities for incorrect labels, since the loss and gradient values tend towards infinity when the predicted probability for a given label approaches zero ($\lim_{p\to 0} -\log p, \nabla_p(-\log p) \to -\infty$). In order to circumvent the adverse impact of cross-entropy loss, which can induce overfitting to label noise, we apply cross-entropy loss adaptively for each sample. This adaptive approach ensures robust learning that remains less-affected by label noise-induced overfitting.

Determining which samples to learn with the cross-entropy loss is based on the theoretical foundation of noise-robust loss. An optimal parameter set exists that satisfies both the optimal points of the noise-robust loss on clean and noisy datasets [5]. It is clear that a model trained on clean data produces high loss values for noisy labels, assuming that the network model has sufficient capacity to learn clean data. Consequently, we can extrapolate that a model trained with noise-robust loss on noisy datasets also yield elevated loss values for noisy samples. Hence, the model cannot reach the desired optimal point when it overfits the samples with high values of noise-robust loss.

Building upon this observation, we propose an adaptive sample selection approach based on the probability of a given label being correct, termed *label confidence*. Our label confidence is formulated by the noise-robust loss and a differentiable mapping function as follows:

$$C(x, y) \coloneqq P((x, y) \in \mathcal{D}_{ce}) = h(\mathcal{L}(f(x;\theta), y)), \quad (1)$$

where $\mathcal{D}_{ce}$ denotes the set of samples to be trained with the cross-entropy loss, $\mathcal{L}$ is the noise-robust loss function, and $h(\cdot)$ is the mapping function between the robust loss value and the label confidence. The label confidence provides a probabilistic measure of the reliability of the associated label. This allows for dynamic adjustments in the selection of samples to be trained using cross-entropy loss, based on their respective label confidences. Note that deterministic sample selection can lead to learning instability, particularly as robust loss values for each sample can exhibit frequent fluctuations during batch learning. As a solution, our adaptive sample selection strategy serves to enhance learning stability. To be more precise, we compute the expectation value of the cross-entropy loss weighted by the label

confidence as:

$$\mathcal{R}_{\mathrm{CE}}(\theta;\mathcal{D}_{ce}) \coloneqq \mathbb{E}_{(x,y)\in\mathcal{D}_{ce}}[\mathrm{CE}(f(x;\theta),y)] = \sum_{(x,y)\in\mathcal{D}_{ce}} \frac{h(\mathcal{L}(f(x;\theta),y))}{|\mathcal{D}_{ce}|}\mathrm{CE}(f(x;\theta),y), \quad (2)$$

where $|\mathcal{D}_{ce}|$ is $\sum_{(x,y)\in\mathcal{D}_{ce}} h(\mathcal{L}(f(x;\theta),y))$.

To design the mapping function, $h(\cdot)$, it should (1) exhibit a monotonically decreasing behavior concerning the loss value, and (2) yield an output of 0 for the loss values larger than the particular threshold value. Satisfying the first condition is essential for aligning with the behavior of noise-robust loss function values at their optimal points. This implies that samples with higher loss values should possess lower probabilities of being learned with cross-entropy loss. The second condition ensures that samples associated with exceedingly high loss values are excluded from learning with the cross-entropy loss. A detailed definition of the threshold is provided in Section 3.2, where we present a formal proof of the theorem.

3.2. Theoretical analysis

In our proposed framework, the model is trained with a weighted cross-entropy loss that incorporates label confidence while simultaneously being trained with the noise-robust loss to calculate label confidence. However, it cannot be guaranteed that the proposed training process can reach the optimal point of the noise-robust loss, which is independent of the issue of underfitting. To address this uncertainty, we provide theoretical evidence that the proposed framework can approximate the minimization of the noise-robust loss by minimizing the entire loss function provided by the framework. Specifically, we prove that there exists a linear combination of the noise-robust loss and the weighted cross-entropy loss with label confidence, which can serve as a lower bound for the noise-robust loss:

Theorem 1. *Let us assume that* $\mathcal{L}(f(x;\theta),y) \coloneqq g(f(x;\theta)_y)$, *where* $g:[0,1]\rightarrow\mathbb{R}^+$. *Given* $\alpha>0$ *such that* $\lim_{p\rightarrow 1}\nabla_p(g(p)+\alpha\log p)<0$, *there exists a value of* $\tau<1$ *that satisfies the following inequality:*

$$\mathcal{R}_{\mathcal{L}}(\theta;\mathcal{D}) \geq \frac{n-|\mathcal{D}_{ce}|}{n}\mathcal{R}_{\mathcal{L}}(\theta;\mathcal{D}) + \alpha\frac{|\mathcal{D}_{ce}|}{n}\mathcal{R}_{CE}(\theta;\mathcal{D}_{ce}), \quad (3)$$

where n *is the number of samples in* $\mathcal{D}$, $h(\cdot)$ *satisfies the two conditions introduced in Section 3.1, and the condition* $h(l)=0$ *for* $l>g(\tau)$ *holds.*

Proof. The proof is in supplementary materials. □

Given that the cross-entropy loss is always non-negative, the following inequality is clearly satisfied:

$$\mathcal{R}_{\mathcal{L}}(\theta;\mathcal{D}) \leq \mathcal{R}_{\mathcal{L}}(\theta;\mathcal{D}) + \alpha\frac{|\mathcal{D}_{ce}|}{n-|\mathcal{D}_{ce}|}\mathcal{R}_{\mathrm{CE}}(\theta;\mathcal{D}_{ce}). \quad (4)$$

Based on the above two inequalities, we can conclude that minimizing the right-hand side (RHS) of (4) can approximate the minimization of the noise-robust loss, provided that $|\mathcal{D}_{ce}|$ is bounded. Furthermore, as $|\mathcal{D}_{ce}|$ approaches zero, the RHS of (4) becomes a more accurate proxy for the noise-robust loss. In particular, when $|\mathcal{D}_{ce}|$ is equal to zero, no sample is trained using the cross-entropy loss, and the RHS of (4) is exactly equal to the noise-robust loss. Thus, our framework can be considered as a generalization of the training process with a noise-robust loss.

Our method introduces an approach to enhance the learning capacity within multi-objective optimization scenarios. This involves training two distinct models independently, each aiming to optimize a different objective function; specifically, the noise-robust loss and the weighted cross-entropy loss. Given that both models are derived from a shared underlying model, we impose a penalty on the difference between the two models to encourage them to include similar parameters:

$$\begin{aligned}&\frac{n-|\mathcal{D}_{ce}|}{n}\mathcal{R}_{\mathcal{L}}(\theta;\mathcal{D}) + \alpha\frac{|\mathcal{D}_{ce}|}{n}\mathcal{R}_{\mathrm{CE}}(\theta;\mathcal{D}_{ce}) \geq \\ &\min_{\theta^*}\frac{n-|\mathcal{D}_{ce}|}{n}\mathcal{R}_{\mathcal{L}}(\theta;\mathcal{D}) + \alpha\frac{|\mathcal{D}_{ce}|}{n}\mathcal{R}_{\mathrm{CE}}(\theta^*;\mathcal{D}_{ce}) \\ &\quad +\lambda\mathcal{L}_p(\theta,\theta^*),\end{aligned} \quad (5)$$

where $\lambda>0$ and $\mathcal{L}_p(\theta,\theta^*)$ is a penalty function that takes its minimum value of 0 when θ equals θ^*. The validity of the inequality can be deduced the fact that the minimum value of the right-hand term in (5) is always smaller than or equal to itself under the constraint, $\theta^*=\theta$. Similar to the inequality in (4), all terms on the right-hand side of (5) are positive. In the same vein, we note that optimizing both models via this method approximates the optimization of a single model with noise-robust loss.

3.3. Overall framework

The proposed method consists of two models, namely the noise-robust model and the noise-free model, both jointly trained with distinct objectives, as shown in Figure 2. The noise-robust model is trained to minimize the noise-robust loss on the entire dataset, while the noise-free model is trained to minimize the weighted cross-entropy loss with label confidence incorporation. At inference time, only the noise-free model is utilized due to its superior performance, as explained in Section 4.1. In our framework, the prediction probability value of the image corresponding to the label is first calculated by the noise-robust model. Then, the noise-robust loss value and the label confidence are sequentially derived using the designed mapping function, denoted as $h(\cdot)$. To ensure that the function meets the proposed conditions outlined in Section 3.1, we define it as follows:

$$h(f(x;\theta)_y) = \sigma(0.5*(-\mathcal{L}(f(x;\theta),y)+\mu+m)), \quad (6)$$

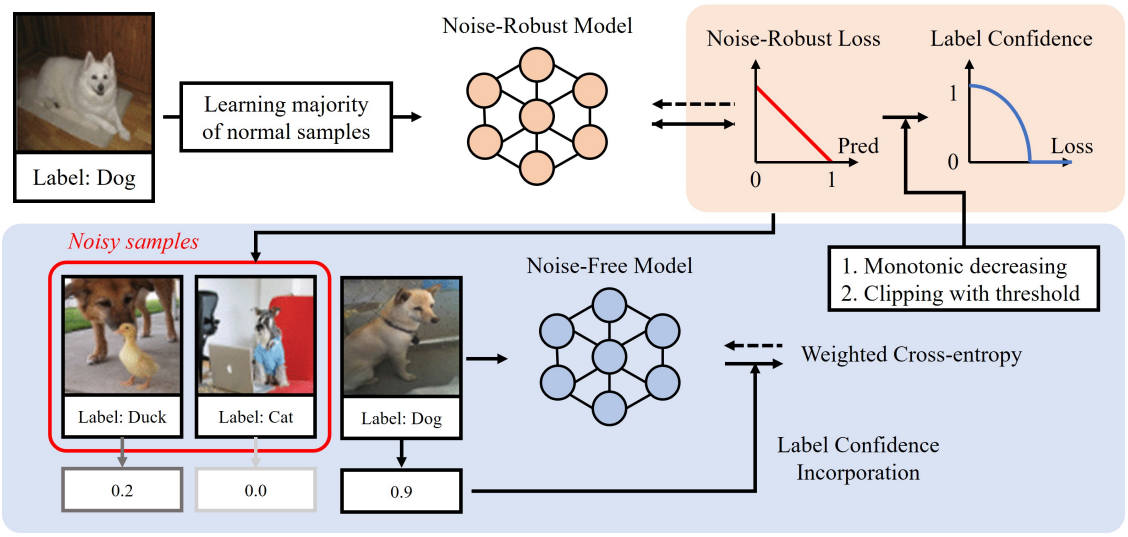


Figure 2. **Overall framework of the proposed method.** Our framework comprises two models, namely the noise-robust model and the noise-free model. The noise-robust model is trained on the entire dataset using a noise-robust loss function to calculate label confidence. The noise-free model is trained using a cross-entropy loss function, incorporating sample-wise label confidence to avoid overfitting on noisy samples. The sample-wise label confidence is obtained using the noise-robust loss values and the proposed mapping function, shown in the upper right of the figure. The mapping function satisfies the proposed conditions on the right, enabling our framework to approximate the optimization of the noise-robust loss function.

where $\sigma(\cdot)$ signifies the sigmoid function, μ represents the average loss value, and m is a variable that affects the size of $|\mathcal{D}_{ce}|$. As m increases, both $|\mathcal{D}_{ce}|$ and the effect of the cross-entropy loss grow. We investigate this phenomenon through an ablation study in Section 4.1. In order to simplify the description of $h(\cdot)$ and reduce the number of hyperparameters, we employ soft thresholding using the sigmoid function, as opposed to the use of hard thresholding.

Finally, the overall loss function of the proposed method is given by:

$$\mathcal{R}_{\mathcal{L}}(\theta;\mathcal{D}) + \alpha\frac{|\mathcal{D}_{ce}|}{n-|\mathcal{D}_{ce}|}\mathcal{R}_{\mathrm{CE}}(\theta^*;\mathcal{D}_{ce}) + \lambda\mathcal{L}_p(\theta,\theta^*). \quad (7)$$

Here, the term training with the cross-entropy loss can be reformulated as follows:

$$\begin{aligned}\alpha\frac{|\mathcal{D}_{ce}|}{n-|\mathcal{D}_{ce}|}\mathcal{R}_{\mathrm{CE}}(\theta^*;\mathcal{D}_{ce}) = \\ \alpha\sum_{(x,y)\in\mathcal{D}}\frac{h(\mathcal{L}(f(x;\theta),y))}{n-|\mathcal{D}_{ce}|}\mathrm{CE}(f(x;\theta^*),y)).\end{aligned} \quad (8)$$

Indeed, calculating $n-|\mathcal{D}_{ce}|$ at every iteration can be computationally demanding. To address this, we approximate it by considering a unit batch as follows: $n-|\mathcal{D}_{ce}| \approx n/|\mathcal{B}|\sum_{(x,y)\in\mathcal{B}}(1-h(\mathcal{L}(f(x;\theta),y)))$, where $\mathcal{B}$ is a batch. Note that the stochastic gradient descent (SGD) method calculates the average of the batch unit loss instead of the overall average loss, so the $n/|\mathcal{B}|$ term can be ignored. The penalty function, $\mathcal{L}_p(\theta,\theta^*)$, in (7) can be any function that attains its minimum value of 0 when $\theta=\theta^*$. We suggest employing the Euclidean distance between two parameter sets as the penalty function, although an alternative could involve using a distillation loss between two models.

Recent learning with noisy labels strategies [4, 11] applied to real-world scenarios have integrated augmentation-invariant regularizers to learn semantic information that reduces the impact of incorrect labels. Additionally, such approaches [26, 33] often employ label correction techniques that modify labels during the learning process. Inspired by these practices, we apply similar heuristic techniques to train the noise-free model using the cross-entropy loss. Remarkably, in multiple instances, these techniques exhibit a notably positive influence on our proposed method.

The **augmentation-invariant regularizer** is a loss function that involves augmenting a single image into two different views, with the objective of aligning the predictions of these augmented images. We utilize the Jensen-Shannon Divergence to quantify the dissimilarity between the predictions made for these augmented views.

Incorporating the **label correction method** entails adjusting labels during the training process based on predictions generated by an oracle model. Within our proposed framework, we assume the oracle model to be the noise-robust model. The practical execution of label correction involves a linear combination of the original ground truth labels and the predicted labels obtained from the noise-robust model. This combination occurs after a specific number of training epochs have been completed.

3.4. Training procedure

The proposed loss function applied to a batch is formulated as follows:

$$\mathcal{L}_{\mathcal{B}} \coloneqq \frac{1}{|\mathcal{B}|} \sum_{(x,y)\in\mathcal{B}} \mathcal{L}(f(x';\theta),y) + C(x',y)\cdot \mathrm{CE}(f(x';\theta^*),\tilde{y}) + \rho\cdot \mathrm{JSD}(f(x';\theta^*), f(x'';\theta^*)) + \lambda||\theta-\theta^*||_F^2, \tag{9}$$

where JSD symbolizes the Jensen-Shannon divergence, $\tilde{y}$ signifies the corrected label, and ρ operates as a scaling factor, influencing the intensity of the augmentation-invariant regularization term. In the equation above, certain hyperparameters, including α and those governing label confidence, have been omitted. Guidelines for omitted hyperparameters can be accessed within the supplementary materials. Both models are trained jointly using $\mathcal{L}_{\mathcal{B}}$. For a clear procedural overview of the training methodology within a single batch, please refer to Algorithm 1, which outlines the sequence of steps.

Algorithm 1 Label Confidence Incorporation for LNL

1: **Input**: λ, ρ, n
2: **Initialize**: $\mu \leftarrow 0$
3: **for** iter = 1:n
4: **Batch sampling:** $\mathcal{B} = \{(x_i, y_i)\} \leftarrow \mathcal{D}$
5: **Data augmentation:** $x', x'' \leftarrow \mathcal{T}(x)$ for $x \in \mathcal{B}$
6: **If** iter $> \lceil n/4 \rceil$**:**
7: $\hat{y} = 0.5\cdot y + 0.5\cdot f(x;\theta)$ for $(x,y) \in \mathcal{B}$
8: **Average loss in (6):** $\mu \leftarrow 0.7\cdot\mu + 0.3\cdot\sum \mathcal{L}(f(x;\theta),y)/|\mathcal{B}|$
9: **Label confidence:** $C(x,y) = h(f(x;\theta)_y)$ for $(x,y) \in \mathcal{B}$
10: **Derive the gradient of $\mathcal{L}_{\mathcal{B}}$ w.r.t.** θ, θ^*
11: **Update parameters with gradient descent,** θ and θ^*

In Algorithm 1, lines 6 to 7 illustrate the label correction process. We formulate this procedure to commence label correction approximately at the 25% point of the overall iterations. This aspect is adaptable, contingent upon the training pace of the noise-robust model. Line 8 entails the computation of the average loss value of the noise-robust model, a pivotal step in deriving the mapping function delineated in equation (6). For computational convenience, we employ a moving average mechanism.

4. Experiments

We conducted experiments to evaluate the effectiveness of our work in learning with noisy labels on both synthetic and real-world datasets. On synthetic datasets, we verified that the proposed method addresses the underfitting issue of the noise-robust loss under various noise ratios. Moreover, we conducted ablation studies to gain insights into the characteristics of our framework. On real-world datasets, we compared the performance of our proposed method with the state-of-the-art methods for learning with noisy labels, and demonstrated its competitive performance.

Experimental setup. We followed the experimental settings of [4] to create the noisy synthetic datasets by adding label noise to the CIFAR-10 and CIFAR-100 datasets [14]. The synthetic datasets consisted of two types of label noise: symmetric and asymmetric noise. Symmetric noise assumes that all samples have an equal probability of being mislabeled, while asymmetric noise assumes label-dependent noise; that is, samples of a particular class are easily mislabeled to a certain class. We trained ResNet-34 [9] with a momentum SGD optimizer as the backbone model, and it was trained from scratch. For detailed experimental settings, please refer to [4]. To evaluate the proposed method's effectiveness in learning with noisy labels on real-world datasets, we used the mini-WebVision [16] and Clothing1M [32] datasets, which are popular benchmarks. We trained ResNet-18 and ResNet-50 [9] with a momentum SGD optimizer on Clothing 1M, and ResNet-50 on WebVision. Also, we used an ImageNet pre-trained model for the Clothing1M dataset. To build the pre-trained model, we performed self-supervised learning [2] instead of using the label information. In all experiments, we used the GCE method [37] as the noise-robust model in our framework. The hyperparameters of the proposed method are reported in the supplementary materials.

4.1. Synthetic label noise

Comparison with state-of-the-art. We compared the classification accuracy of our method with state-of-the-art noise-robust loss methods to demonstrate that it can effectively reach the desired optimal point. The comparison algorithms included LS [28], BS [23], SCE [30], GCE [37], NCE [19], JS [4], and GJS [4]. We conducted the experiments under the same setting as [4] and the performances of comparison methods were taken from the paper. We measured the average accuracy of the proposed method trained on five different seeds.

Table 1 shows the classification accuracy of various methods in different noise ratios. Our method outperformed all other competitors in all cases. In particular, the performance difference with other methods was more significant in situations with asymmetric noise or high noise ratios. For example, when there was 80% symmetric label noise, our proposed method achieved an accuracy increase of 10.45%p and 11.37%p from the second best on CIFAR-10 and CIFAR-100, respectively. Our proposed method also demonstrated robust classification performance against high ratios of label noise compared to GJS [4] that uses an augmentation-invariant regularizer.

Noise-robust model versus noise-free model. Our proposed framework consists of two models: a noise-robust model and a noise-free model. The noise-robust model is

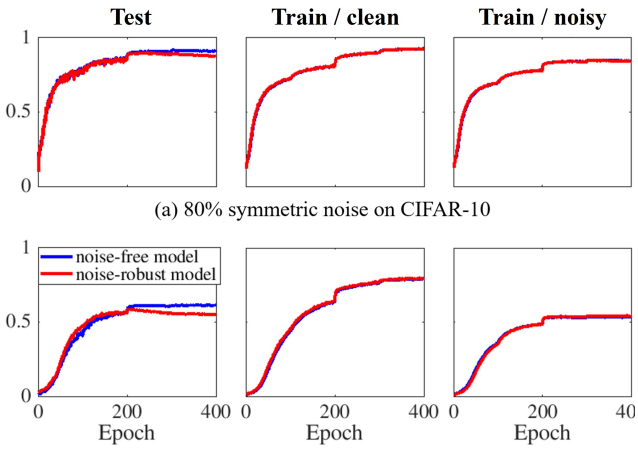
Table 1. **Comparison of classification accuracy (%) on noisy CIFAR-10 and CIFAR-100 datasets.** The noise ratio is indicated under the type of noise. The proposed method is compared with other noise-robust methods using ResNet-34 backbone. The results are averaged over five different seeds. Results in bold indicate the best performance, and underlined text represents the second best performance.

		no noise	symmetric noise				asymmetric noise	
Dataset	Method	0	20	40	60	80	20	40
CIFAR-10	CE	95.77	91.63	87.74	81.99	66.51	92.77	87.12
	BS [23]	94.58	91.68	89.23	82.65	16.97	93.06	88.87
	LS [28]	95.64	93.51	89.90	83.96	67.35	92.94	88.10
	SCE [30]	95.75	94.29	92.72	89.26	80.68	93.48	84.98
	GCE [37]	95.75	94.24	92.82	89.37	79.19	92.83	87.00
	NCE + RCE [19]	95.36	94.27	92.03	87.30	77.89	93.87	86.83
	JS [4]	95.89	94.52	93.01	89.64	76.06	92.18	87.99
	GJS [4]	95.91	95.33	93.57	91.64	79.11	93.94	89.65
	Ours	**96.10**	**95.78**	**95.47**	**94.47**	**91.13**	**95.68**	**93.17**
CIFAR-100	CE	77.60	65.74	55.77	44.42	10.74	66.85	49.45
	BS [23]	77.65	72.92	68.52	53.80	13.83	73.79	64.67
	LS [28]	78.60	74.88	68.41	54.58	26.98	73.17	57.20
	SCE [30]	78.29	74.21	68.23	59.28	26.80	70.86	51.12
	GCE [37]	77.65	75.02	71.54	65.21	49.68	72.13	51.50
	NCE + RCE [19]	74.66	72.39	68.79	62.18	31.63	71.35	57.80
	JS [4]	77.95	75.41	71.12	64.36	45.05	71.70	49.36
	GJS [4]	79.27	78.05	75.71	70.15	44.49	74.60	63.70
	Ours	**79.40**	**78.21**	**75.82**	**71.28**	**61.05**	**77.08**	**68.05**

designed to reach the optimal point of the noise-robust loss, as proven theoretically. Besides, the noise-free model has the potential to fit well-trained samples effectively from the noise-robust model. To help users choose between the two models, we provide training curves for both models, and their performances are measured at every epoch under challenging noise ratios (80%), to show a clear difference.

Figure 3 illustrates the accuracy comparison between the noise-robust model and the noise-free model on CIFAR-10 and CIFAR-100 datasets with 80% of symmetric label noise. Both models share knowledge and show similar performance at every epoch on the train dataset. However, on the test dataset, the difference in accuracy gradually increases as the epoch increases. From the results, it appears that the noise-free model trained with cross-entropy holds better generalization performance. The phenomenon that the noise-robust model loses accuracy in the test set, although the accuracy in the training set is the same as the noise-free model, can be analyzed as an underfitting problem in noise-robust loss. This implies that our noise-free model properly learned both the properties of noise-robust loss and cross-entropy loss. Therefore, we argue that the proposed noise-free model can prevent this underfitting issue and recommend using the noise-free model over the noise-robust model for solving classification problems.

Performance according to the effect of cross-entropy. In Equation (6), we defined a function that maps the noise-robust loss value to the label confidence. The mapping function includes an adjustable variable m, which increases the entire label confidence as m increases, leading to increase



(a) 80% symmetric noise on CIFAR-10

(b) 80% symmetric noise on CIFAR-100

Figure 3. **Comparison of accuracy between the noise-robust model and the noise-free model.** We measure the accuracy on the test set, clean train set, and noisy train set at each epoch. The performance of both models is measured on CIFAR-10 and CIFAR-100 datasets with 80% of symmetric label noise.

the effect of the cross-entropy loss. We argued in the paper that our criterion is a simple and generalizable approach compared to existing sample selection methods. To support this claim, we conducted experiments on the CIFAR-10 and CIFAR-100 datasets with 80% symmetric label noise and verified the performance sensitivity of m. We adjusted m from -5 to 5 in this experiment, where the average test set loss was less than 2. Figure 4 shows the performance of the

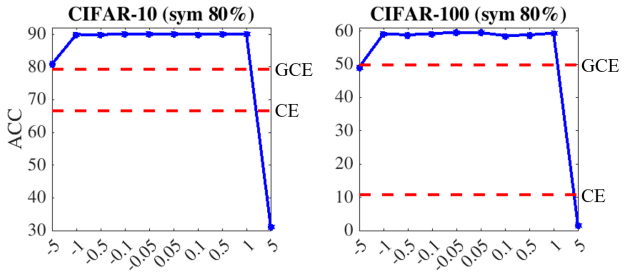


Figure 4. **Performance according to the effect of the cross-entropy loss.** We measure the classification accuracy while changing the effect of the cross-entropy loss by adjusting the variable m in Equation (6). The experiment is conducted on CIFAR-10 and CIFAR-100 datasets with 80% symmetric label noise. The range of m is from -5 to 5. As m increases, the effect of the cross-entropy loss becomes larger, leading to a larger $|\mathcal{D}_{ce}|$.

proposed method as m varies.

The results indicate that the proposed method shows consistent performance while m changes between -1 and 1, demonstrating the advantage of the adaptive sample selection approach trained with cross-entropy loss. Even if mislabeled samples are trained with cross-entropy loss, their label confidence may be low. In this case, their contribution to learning the noise-free model will be reduced. When m is -5, the performance of the proposed method is similar to that of the GCE [37] model, supporting the theoretical analysis that our method approximates the noise-robust loss model more accurately as the size of $|\mathcal{D}_{ce}|$ decreases. Finally, when m is 5, the influence of the cross-entropy loss is the strongest in this experiment. The performance of the proposed method is lower than that of the model trained with the cross-entropy loss, suggesting that the simple combination of the cross-entropy loss and the noise-robust loss can cause significant performance degradation.

Variants of the proposed method. In our paper, we proposed that incorporating label confidence into the cross-entropy loss function can help mitigate the underfitting problem and approximate the learning of a noise-robust loss function. To increase the learning capacity of our model, we extended the single-model approach to a two-model framework, where each model optimizes one of the two objectives. Moreover, we integrated augmentation-invariant regularization and label correction methods into our framework. To evaluate the impact of each of these modifications on the performance of our method, we conducted extensive experiments and compared the results with the baseline method, which uses the GCE loss function [37].

Our experiments on the CIFAR-10 and CIFAR-100 datasets, which involve challenging levels of noise corruption, revealed that our 'Single model' approach, which simultaneously learns with the cross-entropy and noise-robust losses, exhibits similar performance to the baseline method. Unfortunately, both objectives in a single model did not yield significant performance improvements, because of the discrepancies between the cross-entropy loss, which concentrates low-probability predictions, and the noise-robust loss. This phenomenon is clearly demonstrated in Figure 3. In contrast, our 'Two models' approach, which trains two separate models, each optimizing one objective, consistently achieved better performance than the baseline method. Finally, our full method, which leverages augmentation-invariant regularization and label correction, demonstrated superior performance. Supplementary material includes experiments conducted across diverse noise ratios to enable a comprehensive evaluation of module-specific performance.

Table 2. **Ablation studies.** 'Single model' refers to a single model that minimizes the right-hand side of equation (4). 'Two models' refers to our framework without augmentation-invariant regularization and label correction. Bold indicates the best performance except for our method.

	CIFAR-10		CIFAR-100	
Method	sym-80	asym-40	sym-80	asym-40
GCE [37]	79.19	87.00	49.68	51.50
Single model	78.10	90.54	40.57	58.34
Two models	**81.49**	**91.20**	**52.22**	**62.45**
Ours	91.13	93.17	61.05	68.05

4.2. Real-world label noise

We conducted experiments on two datasets to evaluate the effectiveness of our proposed method in real-world scenarios: the Clothing1M dataset [32], which includes 14 categories of clothing, and the mini-WebVision dataset, which consists of samples from the 50 most popular classes in the WebVision dataset [16]. We compared our method with state-of-the-art approaches, including Co-teaching [7], Co-teaching+ [34], JoCor [33], ELR+ [17], DivideMix [15], and GJS [4]. Consistent with GJS, we also presented ensemble results ('(E)' in Table 3) by training two independent networks and ensembling their outputs. For the mini-WebVision experiment, we used a contrastive loss [3] to learn a model quickly during the warm-up step. Additionally, we applied the ColorJitter method as an image augmentation technique, following GJS. In the Clothing1M experiment, we used only 128,000 samples in the training dataset and ensured each category contained an equal number of samples. We adopted the ELR+ settings for our implementation, except for Mixup [36].

In Table 3, our proposed method achieved state-of-the-art performance on the mini-WebVision dataset and competitive results on the Clothing1M dataset, demonstrating its effectiveness in real-world scenarios. Specifically, our method achieved higher accuracy than ELR+ and DivideMix on the mini-WebVision dataset, despite

Table 3. **Comparison of classification performance on real-world noisy datasets.** We compare our proposed method with recent studies on two real-world noisy datasets: mini-WebVision (WebVision) and Clothing1M (Clothing). '(E)' denotes the ensemble performance of two independently trained networks, and 'IResNet2' indicates InceptionResNetV2. We report the highest performance described in each paper, except for the GJS method on Clothing1M. The best performance is indicated in bold.

Method	Backbone	WebVision	Clothing
ELR+ [17]	IResNet2	77.78	**74.81**
DivideMix [15]	IResNet2	77.32	74.76
CE	ResNet-34	70.69	69.88
GJS [4]	ResNet-50	77.99	72.43
Ours	ResNet-50	**78.72**	74.61
GJS (E) [4]	ResNet-50	79.28	-
Ours (E)	ResNet-50	**80.00**	-
Co-teaching [7]	ResNet-18	-	69.21
Co-teaching+ [34]	ResNet-18	-	59.32
JoCor [33]	ResNet-18	-	70.30
Ours	ResNet-18	-	**72.97**

the fact that they used a powerful InceptionResNetV2 backbone [27] and Mixup [36] as an augmentation technique. Similar to our approach, GJS used noise-robust loss and augmentation-invariant regularization, and it achieved higher performance than ours by ensembling two independently trained networks. However, our method also achieved similar performance improvements when it used an ensemble method. In the Clothing1M experiment, methods without Mixup showed relatively low accuracy. Nevertheless, our proposed method achieved a difference in accuracy of only 0.2%p compared to the highest-performing method, whereas GJS had a larger gap. In addition, our method outperformed other sample selection methods using ResNet-18 backbone. These results suggest that our proposed method performs competitively, not only in the presence of synthetic noise but also in real-world noise.

5. Conclusion

In conclusion, this paper proposes a novel training strategy to address the underfitting problem of robust loss methods in deep learning for image classification tasks with noisy labels. Our proposed approach incorporates sample-wise label confidence into the training process, resulting in lower penalties for incorrect predictions of noisy samples with lower confidence levels. Our method is trained on the weighted cross-entropy loss, where the weight is calculated based on the label confidence, and can be applied to existing robust loss methods. Theoretical evidence suggests that our method can achieve the optimal point of the selected robust loss method, provided certain conditions are met regarding the sample-wise label confidence. Experimental results on both synthetic and real-world datasets demonstrate the effectiveness and robustness of our proposed method in handling noisy labels. The proposed method outperforms existing robust loss methods, extremely as the noise ratio increases, and achieves competitive performance compared to state-of-the-art methods. We believe that our theoretical evidence can provide valuable guidance for future research in the field of learning with noisy labels, particularly for those looking to define their own label confidence measures.

References

[1] Görkem Algan and Ilkay Ulusoy. Image classification with deep learning in the presence of noisy labels: A survey. *Knowledge-Based Systems*, 215:106771, 2021. 1

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Neural Information Processing Systems (NIPS)*, 33:9912–9924, 2020. 6

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pages 1597–1607, 2020. 8

[4] Erik Englesson and Hossein Azizpour. Generalized jensen-shannon divergence loss for learning with noisy labels. *Neural Information Processing Systems (NIPS)*, 34:30284–30297, 2021. 1, 2, 3, 5, 6, 7, 8, 9

[5] Aritra Ghosh, Himanshu Kumar, and P Shanti Sastry. Robust loss functions under label noise for deep neural networks. In *AAAI conference on artificial intelligence (AAAI)*, volume 31, 2017. 1, 2, 3

[6] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. 2017. 1, 2

[7] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Neural Information Processing Systems (NIPS)*, 31, 2018. 1, 8, 9

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015. 1

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1, 6

[10] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. *Neural Information Processing Systems (NIPS)*, 31, 2018. 1, 2

[11] Ahmet Iscen, Jack Valmadre, Anurag Arnab, and Cordelia Schmid. Learning with neighbor consistency for noisy labels. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4672–4681, 2022. 3, 5

[12] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning (ICML)*, pages 2304–2313, 2018. 1, 3

[13] Youngdong Kim, Juseung Yun, Hyounguk Shon, and Junmo Kim. Joint negative and positive learning for noisy labels. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9442–9451, 2021. 1

[14] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1, 6

[15] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. 2020. 8, 9

[16] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017. 6, 8

[17] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Neural Information Processing Systems (NIPS)*, 33:20331–20342, 2020. 8, 9

[18] Yangdi Lu, Yang Bo, and Wenbo He. Confidence adaptive regularization for deep learning with noisy labels. *arXiv preprint arXiv:2108.08212*, 2021. 3

[19] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *International Conference on Machine Learning (ICML)*, pages 6543–6553, 2020. 2, 3, 6, 7

[20] Eran Malach and Shai Shalev-Shwartz. Decoupling" when to update" from" how to update". *Neural Information Processing Systems (NIPS)*, 30, 2017. 3

[21] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Neural Information Processing Systems (NIPS)*, 33:20673–20684, 2020. 2

[22] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1944–1952, 2017. 1, 2

[23] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. 2015. 2, 6, 7

[24] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 1

[25] Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. Training convolutional networks with noisy labels. 2015. 1, 2

[26] Zeren Sun, Fumin Shen, Dan Huang, Qiong Wang, Xiangbo Shu, Yazhou Yao, and Jinhui Tang. Pnp: Robust learning from noisy labels by probabilistic noise prediction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5311–5320, 2022. 3, 5

[27] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI conference on artificial intelligence (AAAI)*, 2017. 9

[28] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016. 2, 6, 7

[29] Arash Vahdat. Toward robustness against label noise in training deep discriminative neural networks. *Neural Information Processing Systems (NIPS)*, 30, 2017. 2

[30] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *International Conference on Computer Vision (ICCV)*, pages 322–330, 2019. 1, 2, 3, 6, 7

[31] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13726–13735, 2020. 1, 3

[32] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2691–2699, 2015. 2, 6, 8

[33] Yazhou Yao, Zeren Sun, Chuanyi Zhang, Fumin Shen, Qi Wu, Jian Zhang, and Zhenmin Tang. Jo-src: A contrastive approach for combating noisy labels. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5192–5201, 2021. 1, 3, 5, 8, 9

[34] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *International Conference on Machine Learning (ICML)*, pages 7164–7173, 2019. 1, 3, 8, 9

[35] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021. 1

[36] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. 2018. 3, 8, 9

[37] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Neural Information Processing Systems (NIPS)*, 31, 2018. 1, 2, 3, 6, 7, 8