Revisiting Interpolation for Noisy Label Correction

Yuanzhuo Xu[1], Xiaoguang Niu[1*], Jie Yang[1], Ruiyi Su[1], Jian Zhang[1], Shubo Liu[1], Steve Drew[2]

[1]School of Computer Science, Wuhan University, China
[2] Department of Electrical and Software Engineering, University of Calgary, Canada
{xyzxyz, xgniu, csyangjie, ruiyisu, jzhang, liu.shubo}@whu.edu.cn, steve.drew@ucalgary.ca

Abstract

Label correction methods are popular for their simple architecture in learning with noisy labels. However, they suffer severely from false label correction and achieve subpar performance compared with state-of-the-art methods. In this paper, we revisit the label correction methods through theoretical analysis of gradient scaling and demonstrate that the sample-wise dynamic and class-wise uniformity of interpolation weight prevents memorization of the mislabeled samples. We then propose DULC, a simple yet effective label correction method that uses the normalized Jensen-Shannon divergence (JSD) metric as the interpolation weight to promote sample-wise dynamic and class-wise uniformity. Additionally, we provide theoretical evidence that sharpening predictions in label correction facilitates the memorization of true class, and we achieve it by employing the augmentation strategy along with the sharpening function. Extensive experiments on CIFAR-10, CIFAR-100, TinyImageNet, WebVision and Clothing1M datasets demonstrate substantial improvements over state-of-the-art methods.

Code — https://github.com/kovelxyz/DULC.

Introduction

Deep neural networks (DNNs) have proven effective in various tasks (He et al. 2016; Song, Kim, and Lee 2019; Wang et al. 2021; Srinivas et al. 2021; Song et al. 2021). The effectiveness relies heavily on the collection of datasets with high-quality annotations. However, collections of datasets and manual annotations are challenging and expensive. As an alternative, most large-scale datasets focus on open-source data that can be automatically annotated by inexpensive strategies, such as adopting web crawling and leveraging search engines (Le and Yang 2015; Li et al. 2017). These alternative methods inevitably introduce numerous noisy samples. Prior art (Arpit et al. 2017) has revealed that deep networks suffer from dramatic degradation in the generalization due to the tendency to overfit to noisy labels.

To tackle this problem, numerous methods (Arpit et al. 2017; Han et al. 2018; Li, Socher, and Hoi 2020; Wei et al. 2020; Li et al. 2022; Lu and He 2022; Karim et al. 2022; Liu, Cheng, and Zhang 2023; Wei et al. 2023; Zhang et al. 2021) have been proposed for learning with noisy labels. These approaches focus on label correction. Part representative group of methods (Patrini et al. 2017; Hendrycks et al. 2018; Liu, Cheng, and Zhang 2023) propose to reverses noisy labels to clean ones with estimation of the noise transition matrix, which is challenging for high numbers of classes and in high noise scenarios. Another group of label-correction-based methods (Reed et al. 2014; Arazo et al. 2019; Lu and He 2022) propose to generate soft targets by performing a convex combination of noisy labels and predictions according to interpolation weight. The core of these methods lies in the construction of the weight for interpolation. Bootstrapping (Reed et al. 2014) employs a static weight without accounting for sample differences. In subsequent works (Arazo et al. 2019; Lu and He 2022), dynamic weights are introduced to evaluate different samples. These dynamic weights are usually formulated based on loss criteria (e.g., CE loss) and, as a result, may be non-uniform by differences in the distribution of losses between easy and hard classes (Karim et al. 2022).

The design of interpolation weights lacking dynamic and class uniformity is susceptible to false corrections as the hard samples and classes are less likely to be corrected. Consequently, label correction methods have gradually lost their competitiveness against semi-supervised methods relying on the clean sample selection (Li, Socher, and Hoi 2020; Karim et al. 2022; Lu and He 2022; Li et al. 2022; Hu et al. 2023; Feng, Ren, and Xie 2023). Then the questions naturally arise: *Can carefully designed dynamic weights rejuvenate interpolation schemes?*

In this paper, we first revisit the interpolation scheme in label correction from the perspective of gradient scaling. The core idea behind gradient scaling is to promote the memorization of true classes and diminish the impact of mislabeled samples on gradients. The theoretical analysis demonstrates that the interpolation weight should adhere to two properties in order to prevent memorization of the mislabeled samples: i) sample-wise dynamic that indicates the weight is dynamic for each sample and aligned with its label cleanliness; ii) the class-wise uniformity, emphasizing that the weight of samples from different classes should be aligned to reduce the inconsistent gradients caused by class difficulty. Besides, we provide theoretical evidence

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

to demonstrate that sharpening predictions in label correction can further facilitate the memorization of true classes.

We then propose a *Dynamic and Uniform Label Correction* (DULC) method which enjoys simplicity and effectiveness. Specifically, we measure the Jensen-Shannon divergence (JSD) between the predictions and noisy labels as the interpolation weight and demonstrate its sample-wise dynamic and alignment with label cleanliness. Then, we adopt max-min normalization on JSDs within the classes to promote class-wise uniformity. We finally sharpen the prediction for the combination with noisy labels by employing the augmentation strategy and sharpening function. DULC outperforms state-of-the-art (SOTA) selection-based semi-supervised methods with much lower complexity under various noise settings, even in the presence of very high label noise (see Table 1). Our contributions are summarized as follows:

- We are the first to revisit the interpolation scheme for label correction from the perspective of gradient scaling. We provide theoretical evidence for two properties of an ideal interpolation weight: sample-wise dynamic and class-wise uniformity. Besides, we theoretically ensure the effectiveness of prediction sharpening in label correction.
- We propose a simple yet effective label correction method named DULC, utilizing the normalized Jensen-Shannon divergence (JSD) to measure the interpolation weight in label correction, ensuring sample-wise dynamic and class-wise uniformity. Furthermore, we sharpen the prediction through the augmentation and sharpening function.
- By providing comprehensive experimental results, we show that DULC, with a much simpler architecture, significantly outperforms SOTA methods on both simulated and real-world noisy datasets. Furthermore, extensive ablation studies are conducted to validate the effectiveness of different components in DULC.

Related Work

A variety of methods have been proposed to improve the robustness of DNNs on noisy datasets. Here, we mainly introduce label correction relevant to our work and sample selection, which becomes the SOTA baseline.

Label correction is mainly based on noise transition matrices or model predictions. The former category of methods (Patrini et al. 2017; Hendrycks et al. 2018; Liu, Cheng, and Zhang 2023) try to estimate the transition matrix from noisy labels to clean labels but are often limited in high noise ratios. The latter category of methods (Tanaka et al. 2018; Zhang et al. 2021; Zheng, Awadallah, and Dumais 2021; Reed et al. 2014; Arazo et al. 2019; Lu and He 2022) gradually adjusts the assigned label based on the model's prediction. Bootstrapping (Reed et al. 2014) proposes to generate the new labels by convexly combining model predictions and assigned labels with fixed weights. M-correction (Arazo et al. 2019) uses instead dynamic weights defined in terms of the sample's training loss values. Follow-up work (Lu and He 2022) proposes to use of the ensemble prediction of multiple epochs to avoid the possible bias of correction. However, although dynamic weight design (Arazo et al. 2019; Lu and He 2022) makes sense compared to fixed weight (Reed et al. 2014), these methods still lack the careful design of weights without consideration of the class-wise uniformity, which we discuss later.

Sample selection identifies the noisy samples, e.g., using a small-loss selection to separate them from the clean ones. Early works (Han et al. 2018; Wei et al. 2020; Yao et al. 2021; Xu et al. 2023) perform small loss selection to filter out clean samples with a known noise ratio and train on them. Follow-up methods (Li, Socher, and Hoi 2020; Karim et al. 2022; Li et al. 2022; Hu et al. 2023; Feng, Ren, and Xie 2023; Zhang et al. 2024; Wang, Fu, and Sun 2024) remove the dependence on the noise prior and design a more precise division scheme to divide the dataset into clean and noisy subsets. The clean set is typically used for conventional supervised learning and the noisy samples are treated as unlabeled data for semi-supervised learning (Berthelot et al. 2019). Crosssplit (Kim et al. 2023) does not separate clean and noisy samples, but randomly divides the dataset and still performs semi-supervised training. To prevent overfitting to noisy samples, the co-training strategy of peer networks is usually applied (Li, Socher, and Hoi 2020; Karim et al. 2022; Hu et al. 2023; Kim et al. 2023). Label correction can be leveraged in sample selection methods (Li, Socher, and Hoi 2020; Karim et al. 2022; Kim et al. 2023) to alleviate the repercussions of incorrect selection. However, it is not deployed to the entire dataset but exclusively to a subset.

Other deep learning methods including: 1)regularization (Zhang et al. 2017; Liu et al. 2020); 2)robust loss (Lu, Bo, and He 2022; Wei et al. 2023); 3)contrastive learning (Kim et al. 2021; Ortego et al. 2021); 4)representation learning (Iscen et al. 2022; Tu et al. 2023). Compared with them, label correction methods exhibit a simpler structure that is efficient and easier to deploy.

Our objective is to elevate label correction to the level of competitiveness seen in SOTA methods relying on sample selection, making it not only simple but also effective. We formulate sample-wise dynamic and class-wise uniformity for interpolation weight from the perspective of gradient scaling and then propose our DULC that ensures them.

Preliminaries

Classification with Noisy Labels Consider the K-class classification task in the noisy-label scenario, the ground truth (clean) label y is unobservable. We only have a noisy training set $\mathcal{D} = \{\boldsymbol{x}_i, \tilde{y}_i\}_{i=1}^{N}$, where $\boldsymbol{x}_i$ is an input and $\tilde{y}_i \in \{1, \cdots, K\}$ is the corresponding noisy label. We denote $\tilde{\boldsymbol{y}}_i \in \{0, 1\}^K$ as one-hot vector of noisy label $\tilde{y}_i$. A DNN $\mathcal{N}_\theta$ maps an input $\boldsymbol{x}_i$ to a K-dimensional logits $\boldsymbol{z}_i$ and then feeds the logits to a softmax function to obtain the predictions $\boldsymbol{p}_i$ of the conditional probability of each class. θ denotes the parameters of the DNN and $\boldsymbol{z}_i \in \mathbb{R}^{K\times 1}$ denotes the logits. We have $\boldsymbol{z}_i = \mathcal{N}_\theta(\boldsymbol{x}_i)$ and $\boldsymbol{p}_i = \mathrm{softmax}(\boldsymbol{z}_i)$. Our task is to obtain a classifier that is robust to label noise without knowing joint probability distribution $P(\boldsymbol{x}, y)$.

Early Learning Phenomenon When training DNNs with the typical cross-entropy (CE) loss in noisy-label scenarios, it has been observed that the DNNs preferentially fit easy (clean) samples before overfitting hard (noisy) samples (Arpit et al. 2017). Since the memorization of DNNs has a preference for easy (clean) samples, the predictive power of a sample's representation aligns with its label cleanliness in the early training stage.

Label Correction Methods Label-correction methods utilize the early learning phenomenon as the model tends to generate clean predictions for each sample. They typically try to generate soft targets by interpolating between the noisy labels and model prediction for each sample $\boldsymbol{x}_i$ by:

$$\hat{\boldsymbol{y}}_i = \alpha_i \hat{\boldsymbol{p}}_i + (1-\alpha_i)\tilde{\boldsymbol{y}}_i \tag{1}$$

where $\alpha_i \in [0,1]$ is the interpolation weights. $\hat{\boldsymbol{p}}_i$ is obtained by performing certain operations (e.g., copy or ensemble) on $\boldsymbol{p}_i$ and its gradient is typically frozen. Thus the empirical training cross-entropy loss becomes:

$$\mathcal{L}_{ce} = -\frac{1}{N}\sum_{i=1}^{N} \ell_{ce}(\hat{\boldsymbol{y}}_i, \boldsymbol{p}_i) = -\frac{1}{N}\sum_{i=1}^{N} \hat{\boldsymbol{y}}_i^{\top} \log(\boldsymbol{p}_i) \tag{2}$$

The key to the label correction methods lies in the design of interpolation weights α in Equation (1).

Dataset	CIFAR-10		CIFAR-100	
Noise ratio	92%	95%	92%	95%
UNICON	90.08	85.94	32.24*	19.37*
DULC	**92.94**	**92.04**	**45.32**	**25.61**

Table 1: Performance under extreme label noise on CIFAR10 and CIFAR100. (*) denotes the results we obtain by rerunning their publicly available code.

Revisit the Interpolation via Gradients

In this section, we introduce a gradient analysis of Eq.(2) to motivate our scheme of interpolation weight. Despite simplifying the actual model and training process, the analysis leads to some interesting implications and provides insight into how the interpolation weight should be set.

For clarity of explanation, we denote the true label of sample $\boldsymbol{x}$ as $y \in \{1, \cdots, K\}$. We then denote the distribution over ground-truth labels for sample $\boldsymbol{x}$ as $q(y|\boldsymbol{x})$, and $\sum_{k=1}^{K} q(k|\boldsymbol{x}) = 1$. Similarly, the prediction probability is defined as $p(k|\boldsymbol{x})$ and $\sum_{k=1}^{K} p(k|\boldsymbol{x}) = 1$. In the case of a single ground-truth label y, we have $q(y|\boldsymbol{x}) = 1$ and $q(k|\boldsymbol{x}) = 0$ for all $k \neq y$. For notation simplicity, we denote p_j, q_j, q_y as abbreviations for $p(j|\boldsymbol{x})$, $q(j|\boldsymbol{x})$ and $q(y|\boldsymbol{x})$, where j represents j-th entry. Based on the early learning phenomenon, We assume that samples tend to have a higher posterior probability $p(y|\boldsymbol{x})$ of ground-truth labels in the early training stage. We then derive the following theorem and the proof is offered in Appendix A.1:

Theorem 0.1. *Given the cross-entropy loss $\mathcal{L}_{ce}$ in Eq.(2), we rewrite the sample-wise loss $\ell_{ce} = -\sum_{k=1}^{K}(\alpha\hat{p}_k + (1-\alpha)q_k)\log p_k$. Its gradient with respect to z_j is*

$$\frac{\partial \ell_{ce}}{\partial z_j} = \begin{cases} \alpha(p_j - \hat{p}_j) + (1-\alpha)(p_j - 1), & q_j = 1 \quad (3a) \\ \alpha(p_j - \hat{p}_j) + (1-\alpha)p_j, & q_j = 0 \quad (3b) \end{cases}$$

where $\alpha \in [0,1]$ and z_j is the j-th entry of logits z.

Theorem 0.1 indicates that learning on true class persists when training with Eq.(2) and the gradient of two terms in ℓ_{ce} is scaled by a positive multiplier term α. In addition, Theorem 0.1 has the following interpretations:

- **Sharpening on $\boldsymbol{p}_i$.** The gradient term $p_j - \hat{p}_j$ in Eq.(3a) and Eq.(3b) is independent of q_j and becomes 0 if $\hat{p}_j$ is a copy of p_j. Simply setting this term to 0 is not advisable as we notice that the true class typically has a higher posterior probability (i.e., $p_y > p_j$ for $j \neq y$). This fact inspires us to apply sharpening operation on $\boldsymbol{p}_i$, making $\hat{p}_y > p_y$ and $\hat{p}_j < p_j$ for $j \neq y$. Thus, the gradient of $p_j - \hat{p}_j$ is negative in true class and positive in other class, effectively promoting memorization of the true classes.
- **Sample-wise dynamic of α.** For the samples with true class j, the gradient term $p_j - 1$ of clean samples in Eq.(3a) tend to vanish after the early learning stage, causing mislabeled samples in Eq.(3b) to dominate the gradient. The multiplier $1-\alpha$ should be sample-wise dynamic and we expect α to reflect the label cleanliness of sample $\boldsymbol{x}_i$ under prediction $\boldsymbol{p}_i$: $\alpha \to 0$ for clean samples and $\alpha \to 1$ for mislabeled samples. Thus by multiplying $1-\alpha$, it counteracts the effect of gradient dominating by mislabeled samples. For the samples that j is not the true class, the gradient term $p_j - 1$ in Eq.(3a) is negative, and p_j in Eq.(3b) is positive. Multiplying the dynamic $1-\alpha$ effectively reduces the magnitudes of coefficients on mislabeled samples, thereby mitigating their impact on the gradient.
- **Class-wise uniformity of α.** It should also be considered that the distribution of prediction p_j over different classes is uneven. Higher prediction probability p_j tends to be skewed towards easier classes, as clean samples from hard classes (e.g.,cats and dogs in CIFAR10) may not have been memorized yet (Karim et al. 2022). In addition to the inherent bias in gradients across classes in Eq.(3a) and Eq.(3b), the non-uniformity of predictions p_j leads to the inter-class bias of α because the dynamic α is related to p_j, which further exacerbates the bias of gradients on different classes. Therefore, we recommend incorporating a mechanism of class-wise uniformity in α to align the gradient scaling of different classes.

Despite their importance, the dynamic and uniformity have hardly been considered or substantiated by theoretical analysis in previous methods. Bootstrapping (Reed et al. 2014) overlooks the sample-wise dynamic and applies the static weight (e.g., $\alpha = 0.6$) to all samples indiscriminately. M-correction (Arazo et al. 2019) makes α dynamic by BMM estimation on standard CE losses of all samples without consideration of class-wise uniformity. Besides, SELC (Lu and He 2022) assigns the same weight to all samples within one

epoch, emphasizing accurate ensemble predictions by aggregating predictions over multiple epochs using exponential moving averages, while overlooking sample-wise dynamic.

We propose DULC to achieve both sample-wise dynamic and class-wise uniformity, which is a simple and effective label correction method. Details of DULC are presented in the following section.

Our Algorithm: DULC

Augmentation and Sharpening

As the discussion for Theorem 0.1, the prediction $\boldsymbol{p}_i$ should be sharpened to promote memorization of the true classes. In other words, the prediction $\hat{\boldsymbol{p}}_i$ involved in label correction should be more confident than the prediction $\boldsymbol{p}_i$ in gradient backpropagation. In this paper, We adopt two mechanisms to achieve this.

Firstly, we utilize the "Weak and Strong Augmentation" strategy to alleviate this problem, which is widely used in semi-supervised learning task (Berthelot et al. 2019; Sohn et al. 2020) and label noise learning (LNL) methods (Li, Socher, and Hoi 2020; Kim et al. 2023). Trivially, we generate strong and weak augmentation sets of the entire dataset, denoted $\mathcal{D}_s$ and $\mathcal{D}_w$ respectively. DULC exploits predictions on weak augmentation set $\mathcal{D}_w$ for label correction and train the network on strong augmentation set $\mathcal{D}_s$. The network produces more confident predictions on the weak augmented set than on the strong augmented set (Cubuk et al. 2018), resulting in sharpening predictions. Besides, we apply a sharpening function with temperature coefficient T on the prediction $\boldsymbol{p}_i$ on $\mathcal{D}_w$ directly reduce its temperature:

$$\hat{\boldsymbol{p}}_j = {p_j}^{\frac{1}{T}} \Big/ \sum_{k=1}^{K} {p_k}^{\frac{1}{T}}, \text{ for } j = 1, 2, \cdots, K. \tag{4}$$

JSD metric for Sample-wise Dynamic

In DULC, we seek to utilize a new metric to dynamize α. Jo-SRC (Yao et al. 2021) and UNICON (Karim et al. 2022) propose to adopt the Jensen-Shannon divergence (JSD) to quantify the difference between prediction probability distribution $\boldsymbol{p}_i$ and the noisy labels distribution $\tilde{\boldsymbol{y}}_i$. JSD is naturally bounded in $[0, 1]$, and we derive the following Theorem to prove that JSD can serve as an ideal metric of α:

Theorem 0.2. *Given the noisy label of sample $\boldsymbol{x}$ as $\tilde{y} \in \{1, \cdots, K\}$. We denote the prediction and one-hot label as $\boldsymbol{p}$ and $\tilde{\boldsymbol{y}} \in \{0, 1\}^K$, respectively. Then the JS-divergence between the $\boldsymbol{p}$ and $\tilde{\boldsymbol{y}}$ becomes:*

$$\mathrm{JSD}(\tilde{\boldsymbol{y}}, \boldsymbol{p}) = \frac{1}{2} p_{\tilde{y}} \log p_{\tilde{y}} - \frac{1}{2}(1 + p_{\tilde{y}}) \log(1 + p_{\tilde{y}}) + 1 \tag{5}$$

Where $p_{\tilde{y}}$ is the abbreviation of $p(\tilde{y}|\boldsymbol{x})$.

Theorem.0.2 shows that in a single-classification scenario, the JSD value between $\tilde{\boldsymbol{y}}$ and $\boldsymbol{p}$ is monotonically decreasing on $p_{\tilde{y}}$. As the network tends to have a larger posterior probability on true class y, $p_{\tilde{y}} \to 1$ or 0 indicates that the sample $\boldsymbol{x}$ is more likely to be clean (i.e., $\tilde{y} = y$) or mislabeled (i.e., $\tilde{y} \neq y$). To verify the effectiveness of JSD metric

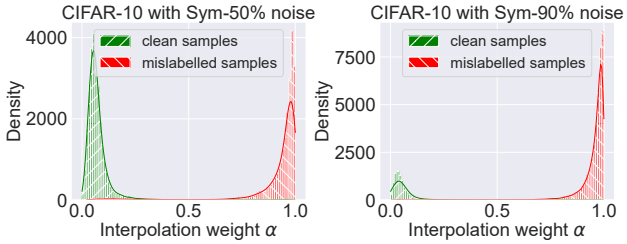


Figure 1: The distribution of interpolation weight α on CIFAR-10 with Sym-50% and Sym-90% label noise using PreAct ResNet-18, and the α is achieved throught normalized JSD metric.

in discriminating the mislabeled samples from clean samples, we empirically analyze the JSD value distribution of clean and mislabeled samples with symmetric noise. Figure 1 shows the results on CIFAR-10. We observe that the JSD values of clean samples exhibit a peak close to 0, while the mislabeled samples are mostly significantly approaching 1, verifying the effectiveness of JSD metric. Thus the JSD metric satisfies the sample-wise dynamic of α we discussed before. We then perform label correction by linearly combining the noisy label $\tilde{\boldsymbol{y}}_i$ with the sharpening prediction $\hat{\boldsymbol{p}}_i$, guided by $\alpha_i = \mathrm{JSD}\left(\tilde{\boldsymbol{y}}_i, \boldsymbol{p}_i\right)$,

$$\hat{\boldsymbol{y}}_i = \mathrm{JSD}\left(\tilde{\boldsymbol{y}}_i, \boldsymbol{p}_i\right) \hat{\boldsymbol{p}}_i + \left(1 - \mathrm{JSD}\left(\tilde{\boldsymbol{y}}_i, \boldsymbol{p}_i\right)\right) \tilde{\boldsymbol{y}}_i \tag{6}$$

JSD Normalization for Class-wise Uniformity

DULC utilizes $\alpha_i = \mathrm{JSD}\left(\tilde{\boldsymbol{y}}_i, \boldsymbol{p}_i\right)$ and ensures the sample-wise dynamic in label correction. However, as we analyzed before, the standard JSD still metric lacks class-wise uniformity as it relies on the posterior probability of noisy class $\tilde{y}_i$ in Eq.(6). Figure 2(a) shows that the JSD metric exhibits a broader range in easy classes (e.g., class 8 and 9) compared to difficult ones (e.g., class 2,3 and 5). To this end, we employ the normalization on standard JSD to align the range of JSD values of each class.

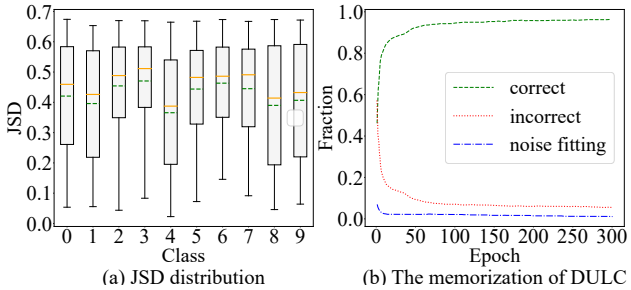
At the beginning of each training epoch, we compute the maximum and minimum JSD values within a class c, which can be expressed as,

$$\begin{aligned} \mathrm{JSD}_c^{\max} &= \max_{\{i|\tilde{\boldsymbol{y}}_i = c\}} \mathrm{JSD}(\tilde{\boldsymbol{y}}_i, \boldsymbol{p}_i) \\ \mathrm{JSD}_c^{\min} &= \min_{\{i|\tilde{\boldsymbol{y}}_i = c\}} \mathrm{JSD}(\tilde{\boldsymbol{y}}_i, \boldsymbol{p}_i) \end{aligned} \tag{7}$$

Then we perform min-max normalization for each sample $(\boldsymbol{x}_i, \tilde{\boldsymbol{y}}_i)$ with Equation (7):

$$\mathrm{JSD}_{\mathrm{norm}}(\tilde{\boldsymbol{y}}_i, \boldsymbol{p}_i) = \frac{\mathrm{JSD}(\tilde{\boldsymbol{y}}_i, \boldsymbol{p}_i) - \mathrm{JSD}_{\tilde{\boldsymbol{y}}_i}^{\min}}{\mathrm{JSD}_{\tilde{\boldsymbol{y}}_i}^{\max} - \mathrm{JSD}_{\tilde{\boldsymbol{y}}_i}^{\min}} \tag{8}$$

Through min-max normalization, we align the interpolation weight α_i of the samples in each class to [0,1]. Nevertheless, the interpolation weight α_i for samples from hard classes (especially in high noise ratios) should not be set as high as 1, as their predictions $\boldsymbol{p}_i$ are not entirely reliable in the early stages. We further adopt a linear decay strategy to constrain



(a) JSD distribution (b) The memorization of DULC

Figure 2: The results on the CIFAR-10 with symmetric-90%-noise. Plot (a) shows the JSD value range of the 10 classes before normalization (yellow: median, green: mean). Plots (b) show the fraction of mislabeled samples of correct prediction (green), noise fitting (i.e. the prediction equals the wrong label, shown in blue), and the incorrect predictions that are neither true nor noise labels (red).

the JSD values to a smaller range in the early stage, $[a, b]$ (e.g., [0.2,0.8]). The final label correction function in Equation (1) becomes:

$$\hat{\boldsymbol{y}}_i = \alpha_i \boldsymbol{p}_i + (1 - \alpha_i)\, \tilde{\boldsymbol{y}}_i \tag{9}$$

$$\alpha_i = (1 - \beta)\mathrm{JSD}_{\mathrm{norm}}\left(\tilde{\boldsymbol{y}}_i, \boldsymbol{p}_i\right) + \gamma\beta \tag{10}$$

where γ is the scaling factor and β is the coefficient that decay linearly with current training epochs t as follow:

$$\beta_t = \mathbb{I}_{\{t \leq N\}} a \left(1 - \frac{t}{N}\right) \tag{11}$$

The linear decay strategy allows us to maintain a certain proportion of prediction weight in the early stages. We perform ablation experiments and provide more details in the Appendix B.4. we further combines MixUp augmentation (Zhang et al. 2017) and contrastive learning loss $\mathcal{L}_{\mathrm{c}}$ (Karim et al. 2022) to mitigate noisy label memorization. The overall training objective is expressed as

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{\mathrm{c}} \tag{12}$$

where λ is the contrastive loss coefficient and $\mathcal{L}_{ce}$ is the cross-entropy loss in Eq.2.

Experiment

Datasets and Implementation Datails

Extensive experiments are conducted on three manually corrupted datasets with different noisy types (i.e. CIFAR-10/100 (Krizhevsky, Hinton et al. 2009) and Tiny-ImageNet (Le and Yang 2015)) and two real-world noisy datasets (i.e., WebVision (Li et al. 2017) and Clothing1M (Xiao et al. 2015)), to demonstrate the effectiveness of DULC. Both CIFAR-10 and CIFAR-100 contain 50K training images and 10K test images. Tiny-ImageNet is a subset of the ImageNet, featuring 200 classes, each with 500 images, totaling 100K images at a size of 64×64. WebVision comprises 2.4 million images sourced from Flickr and Google, categorized into the same 1,000 classes as ImageNet ILSVRC12. Consistent with previous studies (Li, Socher, and Hoi 2020; Karim et al. 2022), we utilize the initial 50 classes from the Google image subset as the training data. Clothing1M is an unbalanced real-world noisy dataset that contains 1M images with about 38.46% noisy labels for training and 10K images with clean labels for testing, and its most populated class contains almost 5 times more instances than the smallest one. For CIFAR-10/100, we conduct two types of commonly simulated noisy labels: symmetric noise (Patrini et al. 2017) rates of 20%, 50%, 80%, and 90%[1] and asymmetric noise (Li et al. 2019) rates of 10%, 30%, and 40%. Symmetric noise is generated by uniformly flipping the label to the opposite class. Asymmetric noise simulates fine-grained classification, with label flipping limited to similar classes (e.g., dog → cat). For CIFAR-100, label flips are applied within each class to transition to the next one within the super-classes. For Tiny-ImageNet, we consider the symmetric noise rates of 20% and 50%.

We use the PreAct ResNet-18 (He et al. 2016) architecture for CIFAR10/100 and TinyImageNet in line with other methods. For WebVision and Clothing1M, we take a ResNet50 (He et al. 2016) instead of a more complex InceptionResNetV2 network (Szegedy et al. 2017) in other methods. To obtain strongly augmentated images, we follow the Auto-augment policy described in (Cubuk et al. 2018). For CIFAR10, CIFAR100 and TinyImageNet, we apply CIFAR10-Policy. For WebVision, we use ImageNet-Policy. Additional details of training and parameter setting, as well as more experimental results and discussions, can be found in the *Appendix*.

Results

CIFAR-10/100: Tabel 2 shows the average test accuracies for CIFAR-10 and CIFAR-100. DULC consistently outperforms the baseline methods in a wide range of noisy ratios. We observe that DULC achieves significant improvements in all noise ratios except for CIFAR-100 with symmetric-50% and asymmetric-40% noise ratios. For the exception, one possible explanation could be that CIFAR-100 has fewer single-class samples (i.e., 500), and the equal number of noisy and clean samples has a higher tolerance for biased division of the dataset, thus bringing greater benefits to the sample selection methods. In particular, we achieve a larger improvement over the SOTA for asymmetric noise. In this more challenging noise ratio, the class labels of the dataset become unbalanced and more consistent with the real scenario. Additionally, following UNICON (Karim et al. 2022), we perform a T-SNE visual comparison on the features learned by the classifier in Appendix C.2. The results show that the features learned by DULC are not more discriminative than UNICON but still promise the SOTA effect. This is due to the smooth transition between interpolated classes, and we provide more discussion in Appendix C.2. It is worth mentioning that DULC can be used as a MixUp-

[1]We follow the same settings as DivideMix (Li, Socher, and Hoi 2020) and 90% symmetric noise means 90% of the samples are randomly allocated.

Dataset	CIFAR-10							CIFAR-100						
Noise type	Symmetric				Asymmetric			Symmetric				Asymmetric		
Methods/Noise ratio	20	50	80	90	10	30	40	20	50	80	90	10	30	40
Standard CE	86.8	79.4	62.9	42.7	88.8	81.7	76.1	62.0	46.7	19.9	10.1	68.1	53.3	44.5
MixUp (2017)	95.6	87.1	71.6	52.2	93.3	83.3	77.7	67.8	57.3	30.8	14.6	72.4	57.6	48.1
ELR (2020)	95.8	94.8	93.3	78.7	95.4	94.7	93.0	77.6	73.6	60.8	33.4	77.3	74.6	73.2
DivideMix (2020)	96.1	94.6	93.2	76.0	93.8	92.5	91.7	77.3	74.6	60.2	31.5	71.6	69.5	55.1
JPL (2021)	93.5	90.2	35.7	23.4	94.2	92.5	90.7	70.9	67.7	17.8	12.8	72.0	68.1	59.5
MOIT (2021)	94.1	91.1	75.8	70.1	94.2	94.1	93.2	75.9	70.1	51.4	24.5	77.4	75.1	74.0
Sel-CL (2022)	95.5	93.9	89.2	81.9	95.6	95.2	93.4	76.5	72.4	59.6	48.8	78.7	76.4	74.2
UNICON (2022)	96.0	95.6	93.9	90.8	95.3	94.8	94.1	78.9	**77.6**	63.9	44.8	78.2	75.6	74.8
MILD (2023)	93.0	88.7	79.1	-	-	-	89.8	67.3	36.0	-	-	69.9	-	-
OT-Filter (2023)	96.0	95.3	94.0	90.5	-	-	95.1	76.7	73.8	61.8	43.8	-	-	**76.6**
HMW (2024)	93.5	95.2	93.7	90.7	93.5	94.7	93.7	76.6	75.8	63.4	43.4	76.7	76.3	72.1
K-SPR (2024)	95.4	-	84.6	-	-	94.5	93.6	77.5	-	30.5	-	-	76.3	73.9
Bootstrapping (2014)	86.8	79.8	63.3	42.9	-	-	-	62.1	46.6	19.9	10.2	-	-	-
M-correction (2019)	94.0	92.0	86.8	69.1	89.6	92.2	91.2	73.9	66.1	48.2	24.3	67.1	58.6	47.4
SELC (2022)	95.0	-	78.6	-	-	-	92.9	76.4	-	37.2	-	-	-	73.6
DULC (ours)	**96.6**	**96.0**	**95.0**	**93.5**	**96.7**	**95.5**	**95.2**	**79.4**	76.4	**67.7**	**52.8**	**79.2**	**77.8**	75.8

Table 2: Comparison with state-of-the-art methods in test accuracy (%) on CIFAR-10 and CIFAR-100 with symmetric noise. The best scores are **boldfaced**, and the second best ones are underlined.

Noise (%)	20		50	
Method	Best	Avg.	Best	Avg.
Standard CE	35.8	35.6	19.8	19.6
F-correction (2017)	44.5	44.4	33.1	32.8
MentorNet (2018)	45.7	45.5	35.8	35.5
Co-teaching+ (2019)	48.2	47.7	41.8	41.2
M-correction (2019)	57.2	56.6	51.6	51.3
NCT (2021)	58.0	57.2	47.8	47.4
OT-Filter (2023)	58.1	57.7	50.9	50.1
UNICON (2022)	**59.2**	58.4	52.7	**52.4**
DULC (ours)	58.9	**58.5**	**52.9**	52.1

Table 3: Test accuracies (%) on Tiny-ImageNet dataset under symmetric noise settings. We report the results for other methods directly from (Karim et al. 2022) with the Best and the average (Avg.) test accuracy (%) over the last 10 epochs.

Dataset	WebVision		ILSVRC12	
Method	*Top1*	*Top5*	*Top1*	*Top5*
MentorNet (2018)	63.0	81.4	57.8	79.9
Co-Teaching (2018)	63.6	85.2	61.5	84.7
Iterative-CV (2018)	65.2	85.3	61.6	85.0
DivideMix (2020)	77.3	91.6	75.2	90.8
ELR (2020)	77.8	91.7	70.3	89.8
MOIT (2021)	78.8	-	-	-
UNICON (2022)	77.6	93.4	75.3	93.7
RCAL+ (2023)	79.6	93.4	76.3	93.7
HMW (2024)	78.0	93.1	71.9	92.2
K-SPR (2024)	78.0	92.3	74.7	92.9
DULC (ours)	**79.9**	**93.7**	**76.9**	**93.9**

Table 4: The results on WebVision and ILSVRC12. All methods are trained on WebVision while evaluated on both Webvsion and ILSVRC12 validation set.

like scheme to benefit the standard classifier, as DULC can achieve higher accuracy than standard CE on clean datasets (a.k.a, with 0 noise ratio), which we conduct more experiments and discussions in the Appendix B.5.

TinyImageNet: We conduct experiments in 20% and 50% symmetric noise ratios. Table 3 presents the performance comparison of DULC and other methods. We count the test accuracy both with the best and average accuracy over the last 10 epochs. The results show that we achieve the best or suboptimal performance on all noise ratios. For the suboptimality of DULC (decreased by the average of 0.3%), our analysis is that the TinyImageNet dataset has more categories and fewer samples per class (500), which leads to more serious prediction confusion in the early stages of the network, resulting in increased errors in label correction. In this scenario, the selection-based method (i.e., UNICON) can alleviate the memorization of noisy labels relatively well.

WebVision: We present our experimental results on this dataset in Table 4. All comparison methods use Inception-ResNetV2 as the backbone, while DULC adopts a simpler network ResNet50. The results demonstrate Top-1 and Top-5 test accuracy on WebVision and ILSVRC12. Despite WebVision's increased complexity as a real-world dataset, we outperform all baselines, achieving 0.3% (top-1) and 0.2% (top-5) improvement over the suboptimal approach. With a simpler ResNet50 backbone, DULC attains superior performance compared to other methods, affirming the effectiveness of our design.

Clothing1M Tabel 6 presents performance comparison on this real world noisy labeled dataset. We achieve 0.11% performance improvement over UNICON (Karim et al. 2022). Clothing1M dataset is unbalanced with greater challenge. The superior result demonstrates that the design of the two criteria to enhancing the robustness of the model for unbalanced datasets.

Noise type	Symmetric				Asymmetric		
Noise ratio	20%	50%	80%	90%	10%	30%	40%
DULC w/o weak&strong Aug.	95.6±0.05	94.8±0.14	93.8±1.89	82.3±2.64	95.9±0.06	94.6±0.21	90.4±0.78
DULC w/o sharpening (T=1)	96.2±0.03	95.8±0.12	86.7±0.18	74.7±2.12	96.5±0.06	<u>94.8</u>±0.23	93.4±0.78
DULC w/o JSD metric	95.9±0.06	95.2±0.15	78.6±2.64	60.7±3.69	90.3±0.16	94.6±0.23	<u>93.1</u>±0.23
DULC w/o JSD normalization	95.2±0.06	90.9±0.04	80.3±0.14	24.6±0.09	96.3±0.11	92.3±0.12	91.5±2.55
DULC w/o linear decay	<u>96.5</u>±0.02	<u>95.9</u>±0.04	<u>94.8</u>±0.3	**93.6**±0.23	<u>96.6</u>±0.17	92.9±0.37	92.2±0.84
DULC	**96.6**±0.02	**96.0**±0.04	**95.0**±0.03	<u>93.5</u>±0.24	**96.7**±0.16	**95.5**±0.34	**95.2**±0.67

Table 5: Ablation study for DULC on CIFAR-10: Test accuracy (%) of different noise ratios. The best scores are **boldfaced**, and the second best ones are <u>underlined</u>.

Method	Accuracy(%)
Cross-Entropy	69.21
JPL (Kim et al. 2021)	74.15
DivideMix (Li, Socher, and Hoi 2020)	74.76
ELR (Liu et al. 2020)	74.81
SELC (Lu and He 2022)	74.01
UNICON (Karim et al. 2022)	74.98
OT-Filter (Feng, Ren, and Xie 2023)	74.50
DISC (Li et al. 2023)	74.79
DULC (ours)	**75.09**

Table 6: Test accuracies (%) on Clothing1M dataset.

Ablation Study and Discussions

To study the impact of each component in DULC, we use CIFAR-10 for the ablation studies. Here, we mainly perform studies on the following components: weak and strong augmentation; dynamic JSD metric, JSD normalization and linear decay. For contrastive learning, we examine and discuss them in Appendix B.6. Table 5 shows the contribution of each component into DUCL.

Discussion of augmentation and sharpening. We first study the effect of augmentation and the sharpening function as they both sharpen the prediction according to our analysis. For augmentation, we perform both label correction and training on weakly augmented datasets as ablation. The result shows a degradation of overall performance. We can observe a 11.2% drop (from 93.5% to 82.3%) in the case of symmetric-90%-noise and a 4.8% drop (from 95.2 %to 90.4%) in the case of asymmetric-40-noise. For sharpening, we set $T = 1$ to remove. The result shows a significant performance decline, particularly at symmetric-90-noise (from 93.5% to 74.7%). We analyze the reason that in high-noise scenarios, the model needs to focus more on learning true classes (i.e., the gradient term $p_j - \hat{p}_j$ in Theorem 0.1) rather than suppressing memorization of mislabeled samples.

Discussion of dynamic JSD metric. The α weight of label correction in DULC adjusts dynamically according to the memorization (JSD metric) of each sample. we use the ensemble strategy in SELC for ablation research, where $\boldsymbol{t}_{[k]} = \alpha^k \hat{\boldsymbol{y}} + \sum_{j=1}^{k} (1-\alpha)\alpha^{k-j} \boldsymbol{p}_{[j]}$ and α is fixed (e.g., 0.9). The results showed similar results to SELC. We observe severe performance degradation of 32.8% in the case of symmetric-90-noise. In high noise ratios, increased differences between samples highlight the deficiency of a static weight design, compromising class-wise uniformity and adversely affecting label correction.

Discussion of JSD normalization. We perform label correction by Equation (6) without normalizing the JSD. The result shows that under high noise ratios (symmetric-90%-noise and asymmetric-45%-noise), the performance drops by as much as 68.9%. The significant decline in performance is a result of heightened non-uniformity among classes in high noise ratios. In this case, hard classes become even harder to rectify, leading to a notable rise in the count of false label corrections. Subsequently, these inaccurate corrections frequently spill over to affect other related classes and create a ripple effect. In contrast, DULC can perform successful label correction and avoids the memorization of noisy labels (see Figure 2(b)).

Discussion of linear decay. We fix β=0 in Equation (10) to remove the linear decay. The results show varied performance degradations, especially severe under asymmetric noise. Specifically, we observe a 3% drop in asymmetric-40%-noise. This is because the class imbalance under this noise ratio leads to lower prediction accuracy for samples in difficult classes. Still, their weights are amplified due to JSD normalization, resulting in false correction. Linear relaxation enables label correction to retain more information about assigned labels during the early stage of low prediction accuracy. We provide more discussion in Appendix B.4.

Conclusion

In this paper, we revisit the interpolation scheme for label correction from the perspective of gradient scaling and provide theoretical evidence for two properties of an ideal interpolation weight: sample-wise dynamic and class-wise uniformity. We then use normalized JSD metric as the interpolation weight to meet the two properties and propose a simple yet effective label correction method DULC. Besides, we theoretically ensure the effectiveness of prediction sharpening in label correction and and successfully implemented it. We demonstrate the SOTA performance of DULC with extensive experiments on multiple noisy datasets. Furthermore, we conduct an ablation study to illustrate the individual contributions of each component to DULC.

Acknowledgements

This work was supported in part by the Key Research and Development Project of Hubei Province (2022BCA057). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

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

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

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.

Cubuk, E. D.; Zoph, B.; Mane, D.; Vasudevan, V.; and Le, Q. V. 2018. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*.

Feng, C.; Ren, Y.; and Xie, X. 2023. OT-Filter: An Optimal Transport Filter for Learning With Noisy Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16164–16174.

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. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hendrycks, D.; Mazeika, M.; Wilson, D.; and Gimpel, K. 2018. Using trusted data to train deep networks on labels corrupted by severe noise. *Advances in neural information processing systems*, 31.

Hu, C.; Yan, S.; Gao, Z.; and He, X. 2023. MILD: Modeling the Instance Learning Dynamics for Learning with Noisy Labels. *arXiv preprint arXiv:2306.11560*.

Iscen, A.; Valmadre, J.; Arnab, A.; and Schmid, C. 2022. Learning with neighbor consistency for noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4672–4681.

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, J.; Baratin, A.; Zhang, Y.; and Lacoste-Julien, S. 2023. CrossSplit: mitigating label noise memorization through data splitting. In *International Conference on Machine Learning*, 16377–16392. PMLR.

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.

Le, Y.; and Yang, X. 2015. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7): 3.

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

Li, J.; Wong, Y.; Zhao, Q.; and Kankanhalli, M. S. 2019. Learning to learn from noisy labeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5051–5059.

Li, S.; Xia, X.; Ge, S.; and Liu, T. 2022. Selective-supervised contrastive learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 316–325.

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*.

Li, Y.; Han, H.; Shan, S.; and Chen, X. 2023. Disc: Learning from noisy labels via dynamic instance-specific selection and correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24070–24079.

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, Y.; Cheng, H.; and Zhang, K. 2023. Identifiability of label noise transition matrix. In *International Conference on Machine Learning*, 21475–21496. PMLR.

Lu, Y.; Bo, Y.; and He, W. 2022. Noise Attention Learning: Enhancing Noise Robustness by Gradient Scaling. *Advances in Neural Information Processing Systems*, 35: 23164–23177.

Lu, Y.; and He, W. 2022. SELC: self-ensemble label correction improves learning with noisy labels. *arXiv preprint arXiv:2205.01156*.

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.

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*.

Sarfraz, F.; Arani, E.; and Zonooz, B. 2021. Noisy concurrent training for efficient learning under label noise. In *Proceedings of the IEEE/CVF Winter Conference on applications of computer vision*, 3159–3168.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33: 596–608.

Song, D.; Wang, Y.; Chen, H.; Xu, C.; Xu, C.; and Tao, D. 2021. Addersr: Towards energy efficient image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15648–15657.

Song, H.; Kim, M.; and Lee, J.-G. 2019. Selfie: Refurbishing unclean samples for robust deep learning. In *International Conference on Machine Learning*, 5907–5915. PMLR.

Srinivas, A.; Lin, T.-Y.; Parmar, N.; Shlens, J.; Abbeel, P.; and Vaswani, A. 2021. Bottleneck transformers for visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16519–16529.

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.

Tanaka, D.; Ikami, D.; Yamasaki, T.; and Aizawa, K. 2018. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5552–5560.

Tu, Y.; Zhang, B.; Li, Y.; Liu, L.; Li, J.; Zhang, J.; Wang, Y.; Wang, C.; and Zhao, C. R. 2023. Learning with Noisy labels via Self-supervised Adversarial Noisy Masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16186–16195.

Wang, P.; Han, K.; Wei, X.-S.; Zhang, L.; and Wang, L. 2021. Contrastive learning based hybrid networks for long-tailed image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 943–952.

Wang, Y.; Fu, Y.; and Sun, X. 2024. Knockoffs-SPR: Clean Sample Selection in Learning With Noisy Labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wang, Y.; Liu, W.; Ma, X.; Bailey, J.; Zha, H.; Song, L.; and Xia, S.-T. 2018. Iterative learning with open-set noisy labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8688–8696.

Wei, H.; Feng, L.; Chen, X.; and An, B. 2020. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13726–13735.

Wei, H.; Zhuang, H.; Xie, R.; Feng, L.; Niu, G.; An, B.; and Li, Y. 2023. Mitigating memorization of noisy labels by clipping the model prediction. In *International Conference on Machine Learning*, 36868–36886. PMLR.

Xiao, T.; Xia, T.; Yang, Y.; Huang, C.; and Wang, X. 2015. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2691–2699.

Xu, Y.; Niu, X.; Yang, J.; Drew, S.; Zhou, J.; and Chen, R. 2023. USDNL: Uncertainty-based single dropout in noisy label learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 10648–10656.

Yao, Y.; Sun, Z.; Zhang, C.; Shen, F.; Wu, Q.; Zhang, J.; and Tang, Z. 2021. Jo-src: A contrastive approach for combating noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5192–5201.

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, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhang, M.; Zhao, X.; Yao, J.; Yuan, C.; and Huang, W. 2023. When noisy labels meet long tail dilemmas: A representation calibration method. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15890–15900.

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

Zhang, Y.; Zheng, S.; Wu, P.; Goswami, M.; and Chen, C. 2021. Learning with feature-dependent label noise: A progressive approach. *arXiv preprint arXiv:2103.07756*.

Zheng, G.; Awadallah, A. H.; and Dumais, S. 2021. Meta label correction for noisy label learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 11053–11061.