DISC: Learning from Noisy Labels via Dynamic Instance-Specific Selection and Correction

Yifan Li[1,2], Hu Han[1,2,3], Shiguang Shan[1,2,3], Xilin Chen[1,2]
[1]Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
[2]University of the Chinese Academy of Sciences, Beijing 100049, China
[3]Peng Cheng Laboratory, Shenzhen 518055, China
{liyifan20g, hanhu, sgshan, xlchen}@ict.ac.cn

Abstract

Existing studies indicate that deep neural networks (DNNs) can eventually memorize the label noise. We observe that the memorization strength of DNNs towards each instance is different and can be represented by the confidence value, which becomes larger and larger during the training process. Based on this, we propose a Dynamic Instance-specific Selection and Correction method (DISC) for learning from noisy labels (LNL). We first use a two-view-based backbone for image classification, obtaining confidence for each image from two views. Then we propose a dynamic threshold strategy for each instance, based on the momentum of each instance's memorization strength in previous epochs to select and correct noisy labeled data. Benefiting from the dynamic threshold strategy and two-view learning, we can effectively group each instance into one of the three subsets (i.e., clean, hard, and purified) based on the prediction consistency and discrepancy by two views at each epoch. Finally, we employ different regularization strategies to conquer subsets with different degrees of label noise, improving the whole network's robustness. Comprehensive evaluations on three controllable and four real-world LNL benchmarks show that our method outperforms the state-of-the-art (SOTA) methods to leverage useful information in noisy data while alleviating the pollution of label noise. Code is available at https://github.com/JackYFL/DISC.

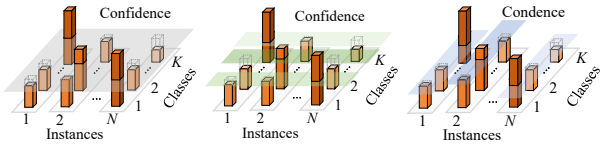
1. Introduction

Label noise is inevitable in image classification model learning, especially for large-scale database annotations through web-crawling [31,40], crowd-sourcing [52], or pre-trained models [12], etc. Recent studies show that DNNs are susceptible to label noise and could fit to the entire data set [2, 55] including the noisy set. Meanwhile, researchers found that DNNs have a memorization effect [2], i.e., the learning process of DNNs follows a curriculum, in which simple patterns are memorized first, followed by more difficult ones like data with noisy labels. Recent studies have explored the use of memorization effect for LNL tasks, with many of these approaches being "early-learning"-based methods [1, 7, 11, 17, 23, 24, 29, 30, 32, 44, 53, 57, 58]. These methods leverage an early-stage DNN to improve the model robustness and generalization ability.

Early-learning-based LNL methods can be divided into three main directions: sample selection [7,17,23,24,29,53], label correction [1,30,44,57] and regularization [11,32,37, 56,58,60]. Sample selection-based methods usually utilize the early-stage DNN's losses or confidence to select reliable

Figure 1. An illustration of DNN's increasing memorization strength during network training. The class prototypes are weights of the DNN classifier, and for simplicity, we only take a two-class case as an example. (a) In the beginning, DNN first fits clean data whose features are closer to class prototypes than noisy data, which is more sparsely distributed in feature space. (b) As training progresses, DNN begins to fit slightly noisy data, some of which can also be classified to its labeled class with relatively high confidence. (c) By the end of the training, the DNN has greatly increased its memorization strength, and even extremely noisy data can also be grouped into its labeled class with high confidence.

This research was supported in part by the National Key R&D Program of China (grant 2021ZD0111901), and the National Natural Science Foundation of China (grant 62176249).



(a) (b) (c)

Figure 2. An illustration of different threshold strategies, including (a) the global threshold, (b) the class-wise threshold, and (c) the proposed dynamic instance-specific threshold.

instances, which are utilized to update the network. Some of these methods require a predefined threshold [24, 30] or prior knowledge about the noise label rate [17, 53] to select instances. Such a global predefined threshold, as shown in Fig. 2 (a), is usually difficult to determine, and may require prior knowledge about the noise label rate to avoid excessive or insufficient noisy data selection, which will further lead to over-fitting and confirmation bias issues [17]. Label correction-based methods try to learn [44] or generate pseudo-labels [30, 57] to replace the original noisy ones. Many of these methods employ the semi-supervised learning (SSL) techniques to pseudo-label the noisy data, and most of them use global (say MixMatch [3], FixMatch [39]) or class-wise threshold (say FlexMatch [54]) to recalibrate labels. As shown in Fig. 2 (b), while a class-wise threshold considers the fitting difficulty of different classes, it still applies a uniform threshold for individual instances in each class. This remains sub-optimum if we consider an example of face images, in which profile face images are more difficult to fit (relatively lower classification confidence) than frontal ones (relatively higher classification confidence) of the same subject. Regularization-based methods aim to design robust loss functions [11, 32, 37, 58, 60, 61] or regularization techniques such as augmentation [56] that can utilize all instances to improve the model robustness against label noise. While these methods work well on moderately noisy data, they may have poor generalization ability under extremely noisy data (see Table 1), since all instances are utilized during the training process.

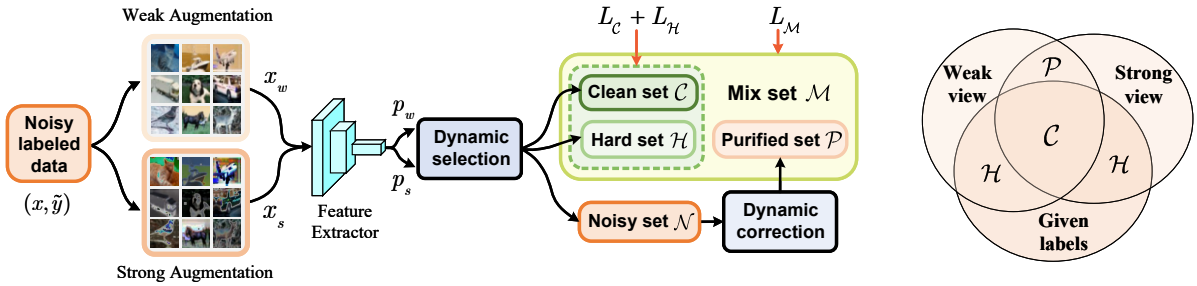
Based on the observation that the memorization strength for individual instances increases during network training, we argue that neither a global threshold nor a class-wise threshold is optimum for LNL. Therefore, we propose a Dynamic Instance-specific Selection and Correction (DISC) approach (see Fig. 3 (a)) for LNL. DISC leverages a dynamic instance-specific threshold strategy (Fig. 2 (c)) following a memorization curriculum to select reliable instances and correct noisy labels. Each threshold is obtained through the momentum of each instance's memorization strength in previous epochs. Such a dynamic threshold strategy can determine a reasonable threshold for each instance according to its memorization strength by the network. Inspired by previous methods of RRL [30], AugDisc [35] and FixMatch [39], DISC also adopts weak and strong augmentations to produce two different views for image classification via a shared-weight model. Unlike previous methods, which use predictions from one view to select reliable instances or generate pseudo-labels for unlabeled data, DISC considers the consistency and discrepancy of two views and divides the noisy data into reliable instances (clean set), hard instances (hard set), and recalibrate noisy labeled instances (purified set), reflecting different degrees of label noise. By dividing the noisy data into three different subsets, DISC can alleviate the contamination of noisy labels to LNL model learning by conquering them via different regularization strategies. As a result, the method can better make full use of the whole noisy dataset. The contributions of this paper include:

- We observe the memorization strength of DNNs towards individual instances can be represented by confidence value, which increases along with training. We provide evidence and experimental analyses to validate this claim.
- Based on the insight of memorization strength, we propose a simple yet effective dynamic instance-specific threshold strategy of LNL that selects reliable instances and recalibrates noisy labels following an easy to hard curriculum.
- Additionally, we leverage the dynamic threshold strategy to group noisy data into three subsets based on predictions from two views generated from weak and strong augmentations. We then adopt different regularization strategies to handle individual subsets.

2. Related Work

Memorization of DNNs. Zhang et.al. [55] observe that the capacity of DNNs is sufficient for memorizing the entire data set. Arpit et. al. [2] propose that DNNs prioritize learning sample patterns first. Furthermore, they also suggest that the notions of the DNNs' capacity are not likely to explain DNNs' memorization degree. Nevertheless, DNNs' memorization degree remains to be studied. Towards the goal of depicting the memorization degree of DNNs during the training process, we propose the *memorization strength* of DNNs defined by the confidence of each instance.

Learning from Noisy Labels. LNL can be roughly categorized into two directions depending on whether or not an additional clean dataset is used. Previous works [20, 21, 28, 45, 46, 50] usually require an additional small clean dataset to learn a robust model, while recent studies focus on a more challenging scenario where only a noisy dataset is provided. However, a majority of these methods are co-training based [17, 29, 32, 34, 47, 53], which require high computation and memory costs. Co-teaching [17] and Co-teaching+ [53] select a fraction of small-loss instances as clean set to teach the other peer network. However, the



(a) (b)

Figure 3. (a) The framework of DISC. DISC first employs a two-view-based network with shared weight for image classification, in which two views are obtained via two diverse augmentations. Then, a dynamic instance-specific threshold strategy is designed by considering both the discrepancy and consistency of two views' predictions in previous epochs, which is used to divide the noisy data into three subsets, i.e., clean set $\mathcal{C}$, hard set $\mathcal{H}$ and purified set $\mathcal{P}$. Finally, we adopt different regularization techniques to conquer different degrees of label noise in $\mathcal{C}$, $\mathcal{H}$ and mix set $\mathcal{M}(\mathcal{C} \cup \mathcal{H} \cup \mathcal{P})$, respectively. (b) An intuitive illustration of subset division by our dynamic instance-specific threshold strategy. The subsets could be regarded as the intersection of two views and the given labels.

selection ratio is hard to set, because we couldn't obtain the actual noise rate in advance for the dataset in the wild. Dividemix [29] utilizes Gaussian Mixture Model (GMM) to divide losses, but the loss distribution is not always Gaussian based which may degrade the selection accuracy and GMM also incurs additional computation cost. Since the performance of DivideMix is remarkable, some works such as AugDisc [35], UNICON [24], CC [59] are recently proposed based on DivideMix. However, these methods only divide the dataset into a clean and a noisy set, where the clean set mainly contains easy instances that cannot provide additional discriminative features to improve performance [1]. Compared with the above methods, DISC avoids confirmation bias and fully exploits the entire noisy data via a dynamic instance-specific threshold strategy and a more delicate division of the noisy dataset. Furthermore, since these methods use a co-training framework, the computation cost is much higher than our DISC (see Fig. 6).

Augmentation Techniques for LNL. Fixmatch [39] reached great success in SSL, which utilizes two diverse augmentations, i.e., a weak augmentation and a strong augmentation such as AutoAugment [8] or RandAugment [9], to perform pseudo-labeling. Some works in LNL [30, 35] commence with this technique, while such works are still rare. Similar to Fixmatch, AugDisc [35] decouples two augmentations with one for analyzing loss, the other for back propagation. RRL [30] utilizes two augmentations to perform contrastive learning. Different from these works, DISC treats two augmentations as two equivalent views, and utilizes both two views to select reliable instances and correct noisy labels.

3. Method

The overall framework of DISC is shown in Fig. 3 (a), and the pseudo-code is presented in the Appendix. DISC begins by applying weak and strong augmentations to a noisy labeled image x, resulting in two augmented images x_w, x_s. These two images are then fed into a two-view learning network with shared weight (f_θ) to obtain two prediction confidences $p_w(c; x)$ and $p_s(c; x)$, where c is the predicted class. DISC then utilizes the two-view confidences to select reliable instances and correct noisy labels, and group the noisy data into three delicate subsets, i.e., clean set, hard set, and purified set. Note that these subsets are divided using the confidences of the previous training epoch to better alleviate confirmation bias. Finally, different regularization techniques are utilized to conquer individual subsets. In this section, we first explain the memorization strength of DNNs, and then introduce the dynamic threshold strategy. Finally, we present our dynamic selection and correction in detail.

3.1. Memorization Strength of DNNs

In a single-label image classification task, the memorization of an instance by DNN refers to the maximum prediction confidence of a given class or a confidence value larger than a certain threshold. The confidence value is closely related to DNNs' memorization strength, i.e., as the confidence increases, the strength for DNNs to memorize one instance also increases. Therefore, we can define DNNs' memorization strength for one instance as the confidence value of a given class. We also notice that the memorization strength of DNNs for given labels (even the noisy ones) gets higher with increased training epochs (see Fig. 4).

Before diving into the explanations behind this phenomenon, we need to make some definitions. Assume that the feature extracted from an image by an encoder could be denoted as $f \in \mathbb{R}^D$, where D indicates the dimension of the feature. Let $W \in \mathbb{R}^{D \times K}$ denote the weights of a single-layer linear classifier, where K indicates the total number of classes. We can also regard W as class prototypes, and then the prototype of the c-th class is $W^c \in \mathbb{R}^D$. Therefore, the similarity between feature $f_i = f_\theta(x_i) \in \mathbb{R}^D$ and c-th class prototype W^c can be expressed as $p(c; x_i) = (W^c)^T f_i$,

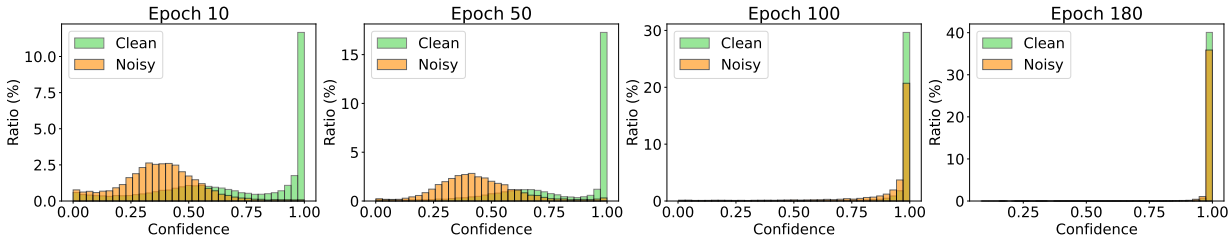


Figure 4. The prediction confidence distributions for given labels during training. The baseline model is PresNet-18 trained on CIFAR-10 with 40% asymmetric noise (actual noise rate 20%) for 200 epochs. It can be seen that the ratio of instances with high confidence increases in the training process, which indicates that the memorization strength of DNNs to each instance gets stronger and stronger.

which also indicates the output logit of a linear classifier. For softmax activation, the confidence of class c can be denoted as:

$$\text{softmax}(p(c;x_i)) = \frac{e^{p(c;x_i)}}{\sum_{c=1}^{K} e^{p(c;x_i)}}. \quad (1)$$

Eq. (1) indicates that the confidence actually reflects the similarity between a feature and a class prototype. As mentioned above, confidence represents the memorization strength of DNNs. Thus, the memorization strength of DNNs towards all instances *de facto* reflects the density of the class cluster. As the training progresses, the memorization strength of DNNs for each instance gets higher and each class cluster gets denser even for the noisy labeled instances. Moreover, different clusters become more dispersed. We visualize the features using t-SNE in Appendix.

The overall training objective of DISC is:

$$L = L_{\mathcal{C}} + \lambda_h L_{\mathcal{H}} + L_{\mathcal{M}}, \quad (2)$$

where λ_h is the hyper-parameter of $L_{\mathcal{H}}$ to balance loss. We will give the formation of each loss function in detail in the following subsections.

3.2. Dynamic Instance-Specific Threshold

Given the classification confidences by the two-view network, it is intuitive to separate the reliable data from the noisy ones. As analyzed in Section 1, selection or correction methods based on a global-fixed threshold or class-wise threshold can be sub-optimum. Therefore, we propose a dynamic instance-specific threshold (DIST) strategy, which can not only select reliable labels but also correct noisy labels. Since the memorization strength of DNNs for individual instances towards different labels gets stronger with the increase of learning epochs, just like human beings, we argue the curriculum for testing DNNs' memorization for given labels should increase accordingly. Therefore, we define the DIST as $\tau(t)$ for each instance x_i:

$$\tau(t) = \lambda\tau(t-1) + (1-\lambda)p(t),\ \tau(0) = 0, \quad (3)$$

where t is the epoch index, $p(t) = max(p(c;x)), c = 0, 1, ..., K-1$, and λ is the layback ratio which controls the delaying degree and threshold stability. Note that we calculate $\tau_w(t)$ and $\tau_s(t)$ of both two views according to Eq. (3). The basic rationale of $\tau(t)$ is that with the increasing of historical confidences, the thresholds should also be increased. However, the confidence of a single epoch may be not stable, especially in the early training epochs. Thus, we use the momentum maximum confidence of each instance, computed based on all previous epochs, as the threshold value.

To correct the noisy labels, we choose to use a higher threshold for each instance. So, we add offset to $\tau(t)$:

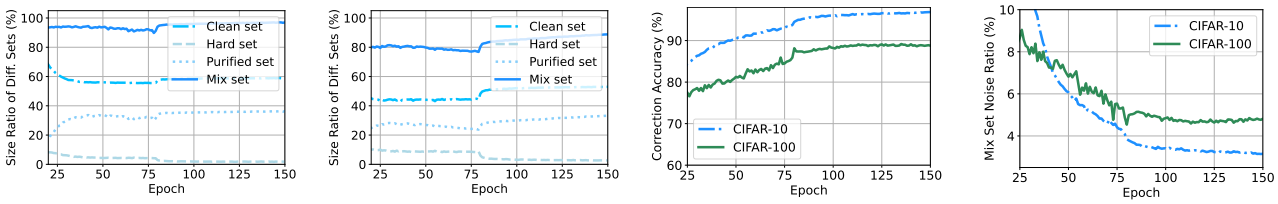
$$\tau'(t) = max(\tau_{ws}(t) + \sigma, 0.99), \quad (4)$$

where σ is a positive offset value. We set an upper bound of 0.99 to limit $\tau'(t)$ value. $\tau_{ws}(t)$ indicates the average value of $\tau_w(t)$ and $\tau_s(t)$.

3.3. Dynamic Selection and Correction

Recent research has found that DNNs have a shortcut learning effect [14, 15], i.e., DNNs may learn biased patterns (such as background, texture and position, etc.) that are not intrinsic cues for classifying images. This effect can cause the model to overfit to the training set, especially to noisy labeled data. To solve this problem, we propose to use weak and strong augmentations to produce two different views, which could provide diverse evidence for the model to memorize data. In this paper, we assume both views have the same importance, and treat them equally. Then we use the confidences of two views and DIST (Sec. 3.2) to finely divide the noisy data into three subsets (see Fig. 3 (b)), i.e., a clean set $\mathcal{C}$ and a hard set $\mathcal{H}$ which are obtained through selecting reliable labels from noisy data, and a purified set $\mathcal{P}$ based on the prediction consistency of two views. The noise degrees in three subsets are different. As a result, we can employ different regularization techniques (detailed below) to conquer these subsets, which can better capitalize on the information in them. This way, the model is expected to memorize the labels with more essential image features. Results show that DISC could suppress the label noise effectively via selection and correction (see Fig. 5).

Dynamic Instance Selection. Assume that the confidence for the i-th instance on its given label could be de-



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

Figure 5. The effect of DISC for noise suppression on CIFAR with 40 % instance noise. The size ratio of different subsets on (a) CIFAR-10 and (b) CIFAR-100. The overall training instances always keep at a high ratio. (c) The correction accuracy of pesudo-labels. DISC can correct a vast majority of the noisy labels. (d) Label noise rate in $\mathcal{M}$. DISC could progressively reduce the noise rate to a low ratio.

noted as $p(y_i; x_i)$, where $y_i \in \mathcal{Y} = \{0, ..., K-1\}$. Thus, the confidences from weak and strong views for image x_i given label y_i could be denoted as $p_w(y_i; x_i)$ and $p_s(y_i; x_i)$, respectively. For an instance, if both confidences in two views are both greater than $\tau(t)$, we can put it into $\mathcal{C}$, i.e.,

$$\mathcal{C} = \{x_i, y_i | p_w(y_i; x_i) > \tau_w(t)\} \cap \{x_i, y_i | p_s(y_i; x_i) > \tau_s(t)\}. \quad (5)$$

It means if an instance can be consistently memorized by two views with high confidence, we regard it as clean.

Similarly, if only one of the two views' confidences for given classes is greater than $\tau(t)$, we can group the instances into the hard set $\mathcal{H}$:

$$\mathcal{H} = \{x_i, y_i | p_w(y_i; x_i) > \tau_w(t)\} \cup \{x_i, y_i | p_s(y_i; x_i) > \tau_s(t)\} - \mathcal{C}. \quad (6)$$

We think the instances in $\mathcal{H}$ may be close to the decision boundary, thus have the potential to improve DNN's generalization ability.

After obtaining $\mathcal{C}$ and $\mathcal{H}$, we adopt different regularization techniques when calculating the loss. For clean set $\mathcal{C}$, we utilize the conventional cross-entropy (CE) loss L_c for both views:

$$L_{\mathcal{C}} = -\frac{1}{N}\sum_{i=1}^{N_c}(\log p_w(y_i; x_i) + \log p_s(y_i; x_i)), \quad (7)$$

where N_c denotes the size of $\mathcal{C}$ and N denotes the size of the entire dataset. We use a scalar $\frac{1}{N}$ to resize the learning rate, and give the basis of such a setting in the Appendix.

While mean absolute error (MAE) has been theoretically proved as a noise-robust loss under certain assumptions [16]. However, using MAE loss can cause hard optimization. CE loss is easy to optimize, but it is sensitive to label noise. Theretofore, for $\mathcal{H}$ (size N_h), we use a more robust regularization technique, i.e., generalized cross-entropy (GCE) loss [58] to handle its label noise. GCE could be regarded as a general loss combining the advantages of MAE and CE:

$$L_{\mathcal{H}} = \frac{1}{N}\sum_{i=1}^{N_h}\left(\frac{1 - p_w(y_i; x_i)^q}{q} + \frac{1 - p_s(y_i; x_i)^q}{q}\right), \quad (8)$$

where $q \in (0, 1]$. Following [58], we also set $q = 0.7$. It can be proved that GCE is equivalent to CE using L'Hôpital's rule when $q \to 0$, and becomes MAE when $q = 1$ [58].

Dynamic Instance Correction. In order to exploit useful information from the remaining noisy data after the selection of $\mathcal{C}$ and $\mathcal{H}$, we perform dynamic instance correction, which can also be regarded as pseudo-labeling. We fuse the confidences from two views as follows:

$$p_{ws}(c; x_i) = \gamma \cdot p_w(c; x_i) + (1-\gamma) \cdot p_s(c; x_i), \forall c \in \mathcal{Y}, \quad (9)$$

where γ is fusion coefficient, and we set $\gamma = 0.5$. Then, we use the maximum value of $p_{ws}(c; x_i)$ rather than the confidence of given classes to obtain a purified set $\mathcal{P}$,

$$\mathcal{P} = \{x_i, \hat{y}_c = \arg\max_c p_{ws}(c; x_i) | \max_c p_{ws}(c; x_i) > \tau'(t), \forall c \in \mathcal{Y}\} - \{\mathcal{C} \cup \mathcal{H}\}, \quad (10)$$

where $\tau'(t)$, which is computed with Eq. 4, gives the dynamic threshold used for correction.

It is prone to induce confirmation bias if we directly use loss using purified set $\mathcal{P}$ for supervision. Inspired by Mixup [56] which is an effective regularization technique dealing with label noise, we perform Mixup on the mix set $\mathcal{M}$ consisting of $\mathcal{C}$, $\mathcal{H}$, $\mathcal{P}$:

$$\mathcal{M} = \{\mathcal{C} \cup \mathcal{H} \cup \mathcal{P}\}. \quad (11)$$

Then, the Mixup for image and label in $\mathcal{M}$ can be denoted as $\tilde{x}_i = \lambda x_i + (1-\lambda)x_{m(i)}$, $\tilde{\boldsymbol{y}}_i = \lambda \boldsymbol{y}_i + (1-\lambda)\boldsymbol{y}_{m(i)}$, $\lambda \sim \text{Beta}(\alpha, \alpha)$, where $\tilde{x}_i$ and $\tilde{\boldsymbol{y}}_i$ are linearly interpolated image and label (one-hot encoding) between index i and another random index $m(i)$ in $\mathcal{M}$ (size N_m). Then, $(x_i, y_i) \in \mathcal{M}$ are included in the network training with a binary cross-entropy loss (BCE):

$$L_{bce}(p(c; x_i), \boldsymbol{y}_i) = -\sum_{c=1}^{C}[\boldsymbol{y}_{ic} \log p(c; x_i) + (1 - \boldsymbol{y}_{ic}) \log(1 - p(c; x_i))]. \quad (12)$$

Then, we can perform Mixup on the mix set $\mathcal{M}$:

$$L_{\mathcal{M}} = \frac{1}{N}\sum_{i=1}^{N_m}[L_{bce}(p_w(c; \tilde{x}_i^w), \tilde{\boldsymbol{y}}_i^w) + L_{bce}(p_s(c; \tilde{x}_i^s), \tilde{\boldsymbol{y}}_i^s)]. \quad (13)$$

Table 1. Comparison with the SOTA methods on CIFAR-10 and CIFAR-100 with IDN. The results with * are implemented by us, and all the other results are directly from [59].

Dataset	CIFAR-10			CIFAR-100		
Noise type	Inst. 20%	Inst. 40%	Inst. 60%	Inst. 20%	Inst. 40%	Inst. 60%
CE*	83.93 ± 0.15	67.64 ± 0.26	43.83 ± 0.33	57.35 ± 0.08	43.17 ± 0.15	24.42 ± 0.16
Forward T [36]	87.22 ± 1.60	79.37 ± 2.72	66.56 ± 4.90	58.19 ± 1.37	42.80 ± 1.01	27.91 ± 3.35
DMI [36]	88.57 ± 0.60	82.82 ± 1.49	69.94 ± 1.34	57.90 ± 1.21	42.70 ± 0.92	26.96 ± 2.08
Mixup* [56]	87.71 ± 0.66	82.65 ± 0.38	58.59 ± 0.58	46.31 ± 0.25	45.14 ± 0.31	23.77 ± 0.26
GCE* [58]	89.80 ± 0.12	78.95 ± 0.15	60.76 ± 3.08	58.01 ± 0.26	45.69 ± 0.14	35.08 ± 0.23
Co-teaching [17]	88.87 ± 0.24	73.00 ± 1.24	62.51 ± 1.98	43.30 ± 0.39	23.21 ± 0.57	12.58 ± 0.58
Co-teaching+ [53]	89.80 ± 0.28	73.78 ± 1.39	59.22 ± 6.34	41.71 ± 0.78	24.45 ± 0.71	12.58 ± 0.58
JoCoR [47]	88.78 ± 0.15	71.64 ± 3.09	63.46 ± 1.58	43.66 ± 1.32	23.95 ± 0.44	13.16 ± 0.91
Reweight-R [49]	90.04 ± 0.46	84.11 ± 2.47	72.18 ± 2.47	58.00 ± 0.36	43.83 ± 8.42	36.07 ± 9.73
Peer Loss [33]	89.12 ± 0.76	83.26 ± 0.42	74.53 ± 1.22	61.16 ± 0.64	47.23 ± 1.23	31.71 ± 2.06
DivideMix [29]	93.33 ± 0.14	95.07 ± 0.11	85.50 ± 0.71	79.04 ± 0.21	76.08 ± 0.35	46.72 ± 1.32
CORSES^2 [7]	91.14 ± 0.46	83.67 ± 1.29	77.68 ± 2.24	66.47 ± 0.45	58.99 ± 1.49	38.55 ± 3.25
CAL [62]	92.01 ± 0.12	84.96 ± 1.25	79.82 ± 2.56	69.11 ± 0.46	63.17 ± 1.40	43.58 ± 3.30
CC [59]	93.68 ± 0.12	94.97 ± 0.09	94.95 ± 0.11	79.61 ± 0.19	76.58 ± 0.25	59.40 ± 0.46
DISC (ours)	**96.48 ± 0.04**	**95.94 ± 0.04**	**95.05 ± 0.05**	**80.12 ± 0.13**	**78.44 ± 0.19**	**69.57 ± 0.14**

Table 2. Comparison with the SOTA methods on Tiny ImageNet with sym. and asym. noise with the highest (Best) and the average (Avg.) test accuracy (%) over the last 10 epochs. The results of all SOTA methods are directly from [37].

Noise	Sym. 0%		Sym. 20%		Sym. 50%		Asym. 45%	
Methods	Avg.	Best	Avg.	Best	Avg.	Best	Avg.	Best
Standard	56.7	57.4	35.6	35.8	19.6	19.8	26.2	26.3
Decoupling [34]	-	-	36.3	37.0	22.6	22.8	26.1	26.6
F-Correction [36]	-	-	-	-	32.8	33.1	0.6	0.67
MentorNet [23]	-	-	-	-	35.5	35.8	26.2	26.6
Co-teaching+ [53]	52.1	52.4	47.7	48.2	41.2	41.8	26.5	26.9
M-Correction [1]	57.2	57.7	56.6	57.2	51.3	51.6	24.1	24.8
NCT [37]	61.5	62.4	57.2	58.2	47.4	47.8	42.4	43.0
UNICON [24]	62.7	63.1	58.4	59.2	52.4	52.7	-	-
DISC (Ours)	**68.2**	**68.5**	**67.5**	**67.9**	**63.9**	**64.3**	**52.8**	**53.6**

4. Experiments

In this section, we evaluate the effectiveness of DISC on benchmarks with controllable label noise, including different levels of noises. Then, we perform experiments on benchmarks with real-world label noise. Finally, we delineate ablation studies to verify each component. All the experiments are implemented on one GeForce RTX3090 GPU and PyTorch 1.8.0.

4.1. Controllable Noise Benchmarks

Dataset. We validate the proposed DISC on CIFAR-10/100 [25] with instance-dependent noise (IDN) [48], and Tiny-ImageNet [26] with two commonly used label noise: symmetric (sym.) noise and asymmetric (asym.) noise. Existing controllable label noise contain two types according to the dependency of data features and class labels [13, 41], i.e., instance-independent noise (including sym. and asym. noise) and IDN. IDN [48] is obtained by setting a random noise rate for each instance following truncated Gaussian distribution, and the noise rate of each class is set randomly (see the Appendix for more details). Following [59, 62], we also adopt noise rates between 20% and 60% on CIFAR-10 and CIFAR-100. Sym. (or uniform noise) noise is generated by uniformly flipping a percentage of the original labels into all possible labels [29]. Asym. noise is generated by only flipping a pair of neighbor classes or confusing classes with a fixed probability. Following previous methods [24, 37], we perform experiments on three different noise rates $\rho \in \{0\%, 20\%, 50\%\}$ for sym. noise and $\rho = 45\%$ for asym. noise on Tiny-ImageNet. We also provide experiments on CIFAR-10/100 with sym. and asym. noise in Appendix.

Experimental setup. We use PresNet-34 [19] and PresNet-18 as a backbone, and train the model for 200 epochs on CIFAR and Tiny-ImageNet, respectively. We use an SGD optimizer with a momentum of 0.9, a weight decay of 0.001, and a batch size of 128. The initial learning rate is set as 0.1 with decaying by a factor of 0.1 in epochs 80 and 160, respectively. We use random cropping and horizontal flipping as weak augmentation, and adopt RandAugment [9] ($n = 2, m = 10$) as strong augmentation. Other training details about hyper-parameters settings of DISC are detailed in Appendix.

Comparison with SOTA methods. Tables 1 and 2 show the results on CIFAR and Tiny-ImageNet, respectively, in which DISC is compared with several baselines.

From Table 1, we can see that DISC outperforms other methods across all the label noise settings on CIFAR. Results show that DISC can handle more challenging noise rates. For instance, DISC improves over 10% on CIFAR-100 with 60% instance noises compared with CC. Moreover, since regularization-based methods (DMI, GCE, Peer loss, Mixup) utilize all the instances during training, they are more susceptible to heavy label noise.

Results of Table 2 show that DISC outperforms all the other methods by a large margin. Although some methods show some robustness to label noise, they may harm the performance without label noise (0%). By contrast, our method

Table 3. Comparison with the SOTA methods on Animals-10N. The results with * are implemented by us, and other results are directly from their original papers.

Method	Accuracy (%)
CE [11]	79.4 ± 0.14
GCE* (2018) [58]	81.5± 0.08
SELFIE (2019) [40]	81.8 ± 0.09
Mixup* (2017) [56]	82.7 ± 0.03
Co-learning (2021) [43]	83.0
PLC (2021) [57]	83.4 ± 0.43
Nested Co-teaching (2021) [6]	84.1 ± 0.1
GJS (2021) [11]	84.2 ± 0.07
DISC (ours)	**87.1 ± 0.15**

Table 4. Comparison with the SOTA methods on Food-101. The results with * are implemented by us, and the other results are directly from their original papers.

Method	Accuracy (%)
CE [11]	81.67
CleanNet (2018) [28]	83.95
GCE* (2018) [58]	85.83
PLC (2021) [57]	83.4
GJS (2021) [11]	86.56
Mixup* (2017) [56]	87.34
Co-learning (2021) [43]	87.57
DISC (ours)	**89.02**

still performs well when the noise rate is 0. In addition, DISC is more robust to label noise compared with other methods such as UNICON. DISC's test accuracy drops less than 5% when adding 50% sym. noise, while UNICON drops more than 10% in the same situation.

4.2. Real-world Noise Benchmarks

Dataset. WebVision1.0 [31] is an in-the-wild benchmark with more than 2.4 million instances where images and annotations are obtained through Google and Flickr using 1000 classes from ImageNet ILSVRC2012 as query words [10]. Following [5], we also use a subset of WebVision which contains the first 50 classes as the training set, and test on both WebVision and ILSVRC2012 validation set. Food-101 [4] is a benchmark containing 101 food categories. It consists of 75,000 noisy labeled training images and 25,000 manually annotated testing images. Animal-10N [40] is a web-crawled benchmark with 5 pairs of confusing animals, containing 50,000 training images and 5,000 testing images. Clothing1M [50] is a large-scale crawled clothing noisy dataset with 1 million training images and 10,000 testing images, which are obtained from several online shopping websites. There are 14 classes in Clothing1M and the noise rate is about 38.5%.

Experimental setup. Following [5, 40], for Animals-10N and WebVision, we use VGG-19 [38] (not pretrained using ImageNet) and Inception-ResNetV2 [42] (not pretrained using ImageNet) as the backbone. Following [29, 57], for Food101N and Clothing1M, we use ResNet-50 [18] (pretrained on ImageNet) as the backbone. The training epochs for Animal-10N are 120, and the epochs are 100 for the others. SGD optimizer is used with initial learning rates of 0.05, 0.01, 0.2, 0.1 for Animals-10N (weight decay 5e-4), Food-101 (weight decay 5e-4), WebVision (weight decay 1e-4) and Clothing1M (weight decay 1e-3), respectively. The batch size is set to 64 for Animals-10N, and 32 for the others. More training details are given in Appendix.

Table 5. Comparison with the SOTA methods on WebVision. The results of all SOTA methods are directly from their original papers.

Dataset	WebVision		ILSVRC12	
Accuracy (%)	top1	top5	top1	top5
F-correction (2017) [36]	61.12	82.68	57.36	82.36
Decoupling (2017) [34]	62.54	84.74	58.26	82.26
D2L (2019) [27]	62.68	84.00	57.80	81.36
MentorNet [23]	63.00	81.40	57.80	79.92
Co-teaching (2018) [17]	63.58	85.20	61.48	84.70
INCV (2019) [5]	65.24	85.34	61.60	84.98
MentorMix (2020) [22]	76.0	90.2	72.9	91.1
ELR (2020) [32]	76.26	91.26	68.7	87.8
DivideMix (2020) [29]	77.32	91.64	75.20	90.84
ELR+ (2020) [32]	77.78	91.68	70.29	89.76
RRL (2021) [30]	77.8	91.3	74.4	90.9
GJS (2021) [11]	77.99	90.62	74.33	90.33
CC (2022) [59]	79.36	**93.64**	76.08	**93.86**
DISC (ours)	**80.28**	92.28	**77.44**	92.28

Comparison with SOTA methods. The results on Animals-10N, Food-101, WebVision and Clothing1M are presented on Tables 3, 4, 5 and 6, respectively. For Clothing1M, we run the open-sourced codes, i.e., DivideMix [29], ELR+ [32], AugDesc [35] and CC [59], using the same random seed as our method. Since all these methods are based on DivideMix, we also use our DIST on the standard DivideMix to replace the GMM division for fair comparison.

The results in Tables 3, 4, 5 show that DISC outperforms the baseline methods on Animals-10N, Food-101N and WebVision. Note that DISC only requires a single network, while some of these methods need two networks, e.g., Decoupling, Co-teaching, nested Co-teaching, ELR+, DivideMix and CC, which indicates that DISC is superior in computation cost (the time comparison with several SOTA methods could be found in Fig. 6). From Table 5, we can see that the performance of DISC on ILSVRC12 is similar to that on WebVision, because the two views from weak and strong augmentations could reduce the shortcut learning effect that leads to overfitting. These results also reflect the strong robustness and generalization ability of DISC.

The results in Table 6 show that DIST can improve performance when combined with DivideMix. However, we notice there exist label noise in the test set of Clothing1M which means the results of individual methods on the test sets of Clothing1M may be not reliable enough, and we provide a detailed analysis of the failed cases on Clothing1M test set, which can support our argument to some extent (see the details in Appendix).

Table 6. Comparison with the SOTA methods on Clothing1M. The results with * are reimplemented by us, and the others are directly from the original paper.

Method	Accuracy (%)
CE	68.94
Co-teaching (2018) [17]	69.21
JoCoR (2018) [47]	70.30
DMI (2019) [51]	72.46
DivideMix* (2019) [29]	74.45
ELR+* (2020) [32]	74.39
GJS (2021) [11]	71.64
CAL (2021) [62]	74.17
AugDesc* (2021) [35]	74.33
CC* (2022) [59]	74.54
DISC (ours)	73.72
DIST+DivideMix	**74.79**

Table 7. Ablation study on CIFAR-10/100 under IDN 20%, 40% and 60%.

Modules				CIFAR-10		CIFAR-100	
Two views	DIST	$\mathcal{H}$	$\mathcal{M}$	Inst. 20%	Inst. 40%	Inst. 20%	Inst. 40%
				83.93	67.63	53.35	43.16
✓				85.62	70.09	66.87	52.42
	✓			92.81	88.85	74.11	70.11
✓	✓			94.44	92.80	76.39	72.41
✓	✓	✓		94.52	92.82	76.45	72.51
✓	✓		✓	96.31	95.74	79.88	78.29
✓	✓	✓	✓	**96.48**	**95.94**	**80.12**	**78.44**

Table 8. Classification accuracies (%) on CIFAR-10 and CIFAR-100 using different selection methods over the last 10 epochs.

Dataset	CIFAR-10			CIFAR-100		
Noise type	Inst. 20%	Inst. 40%	Inst. 60%	Inst. 20%	Inst. 40%	Inst. 60%
Small-losses [17]	90.83	84.81	21.47	71.82	63.89	22.56
GMM [29]	92.78	85.12	48.81	72.91	30.73	11.19
Fixed thres. 0.5 [30]	84.25	60.53	20.85	61.37	45.40	14.78
DIST	**92.81**	**88.85**	**80.66**	**74.11**	**70.11**	**60.07**

Table 9. Classification accuracies (%) on CIFAR-10 and CIFAR-100 using one or two augmented views over the last 10 epochs. W and S denote weak and strong augmentation, respectively.

Dataset	CIFAR-10			CIFAR-100		
Noise type	Inst. 20%	Inst. 40%	Inst. 60%	Inst. 20%	Inst. 40%	Inst. 60%
DISC-W	95.73	94.32	88.37	78.18	75.21	66.88
DISC-WW	96.20	94.47	93.65	79.24	77.67	68.85
DISC-WS	**96.48**	**95.94**	**94.86**	**80.12**	**78.44**	**69.57**

4.3. Ablation Study

We add the key modules of our DISC to the baseline model one by one to investigate their effectiveness. The results are shown in Tables 7, 8 and 9. Additional comparisons can be found in Appendix.

The effect of DIST. Experiments in Table 7 demonstrate that the performance of the baseline model improves by a large margin after adding DIST $\tau(t)$, especially for CIFAR-100. We also provide the comparison of DIST and other instance selection methods in Table 8, where all the results are reimplemented using original selection methods without other techniques. The small-losses method [17] is not robust enough when facing extreme label noise. GMM [29] performs well on CIFAR-10, but not on CIFAR-100 with a large noise ratio. Fixed threshold method is susceptible to label noise. The success of DIST attributes to three folds. First, it sets a dynamic threshold according to the memorization strength of DNN. Second, it sets a reasonable threshold for each instance. Third, it uses the confidence from the previous epoch rather than the current one, which could better reduce confirmation bias.

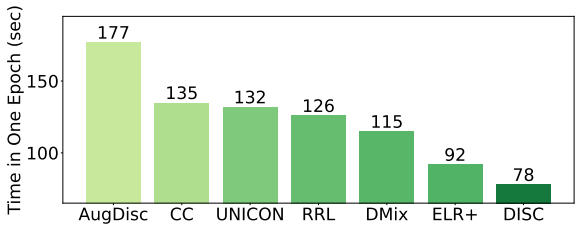


Figure 6. Training and testing time profiling with PresNet-34 backbone and RTX 3090 GPU on CIFAR-10 with 20% inst. noise in one epoch.

The effect of two views. Two views could provide different clues for dividing the noisy set, and prevent overfitting (see Table 7). We also probe into the effect of different views in Table 9. Results show that using two-view is necessary especially when label noise is heavy, and the results will be improved further if two views are diverse.

The effect of different subsets. To better exploit the useful information in noisy labeled data, we select and correct noisy data to obtain clean set $\mathcal{C}$, hard set $\mathcal{H}$ and purified set $\mathcal{P}$. From Table 7, we can see that the hard set $\mathcal{H}$ is beneficial since the hard instances mainly lie in the decision boundary which could enhance the representative ability of DNNs. Since the mix set is much cleaner compared with the original noise set (see Fig. 5 (c)), the mixup regularization could handle these instances better.

4.4. Limitations

DISC corrects the noisy labels based on the confidences of both two views, which may also induce confirmation bias, since high-confidence instances may be the easy ones with noisy labels rather than the clean ones. Moreover, although DIST could lessen the class imbalanced problem to some extent, extremely imbalanced data distribution remains a big challenge.

5. Conclusion

In this paper, we reveal that the memorization strength of DNNs towards individual instances (no matter clean or noisy one) could be reflected by confidences, which usually become higher along with training. Based on this, we propose a Dynamic Instance-specific Selection and Correction method (DISC) for LNL. DISC is able to set a reasonable threshold for each instance and delicately divide the noisy data into different subsets, which can more effectively suppress the label noise during classification model learning. Experiments verify the effectiveness of our method.

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *Proc. ICML*, pages 312–321, 2019. 1, 3, 6

[2] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *Proc. ICML*, pages 233–242, 2017. 1, 2

[3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Proc. NeurIPS*, volume 32, pages 5049–5059, 2019. 2

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *Proc. ECCV*, 2014. 7

[5] Pengfei Chen, Ben Ben Liao, Guangyong Chen, and Shengyu Zhang. Understanding and utilizing deep neural networks trained with noisy labels. In *Proc. ICML*, pages 1062–1070. PMLR, 2019. 7

[6] Yingyi Chen, Xi Shen, Shell Xu Hu, and Johan AK Suykens. Boosting co-teaching with compression regularization for label noise. In *Proc. CVPRW*, pages 2688–2692, 2021. 7

[7] Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. In *Proc. ICLR*, 2021. 1, 6

[8] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proc. CVPR*, pages 113–123, 2019. 3

[9] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proc. CVPRW*, pages 702–703, 2020. 3, 6

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. CVPR*, pages 248–255, 2009. 7

[11] Erik Englesson and Hossein Azizpour. Generalized jensen-shannon divergence loss for learning with noisy labels. In *Proc. NeurIPS*, volume 34, 2021. 1, 2, 7, 8

[12] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proc. CVPR*, pages 5562–5570, 2016. 1

[13] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE TNNLS*, 25(5):845–869, 2013. 6

[14] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 4

[15] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *Proc. ICLR*, 2018. 4

[16] Aritra Ghosh, Himanshu Kumar, and PS Sastry. Robust loss functions under label noise for deep neural networks. In *Proc. AAAI*, volume 31, 2017. 5

[17] 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. In *Proc. NeurIPS*, volume 31, 2018. 1, 2, 6, 7, 8

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pages 770–778, 2016. 7

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Proc. ECCV*, pages 630–645, 2016. 6

[20] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. In *Proc. NeurIPS*, volume 31, 2018. 2

[21] Mengying Hu, Hu Han, Shiguang Shan, and Xilin Chen. Weakly supervised image classification through noise regularization. In *Proc. CVPR*, pages 11517–11525, 2019. 2

[22] Lu Jiang, Di Huang, Mason Liu, and Weilong Yang. Beyond synthetic noise: Deep learning on controlled noisy labels. In *Proc. ICML*, pages 4804–4815, 2020. 7

[23] 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 *Proc. ICML*, pages 2304–2313, 2018. 1, 6, 7

[24] Nazmul Karim, Mamshad Nayeem Rizve, Nazanin Rahnavard, Ajmal Mian, and Mubarak Shah. Unicon: Combating label noise through uniform selection and contrastive learning. In *Proc. CVPR*, pages 9676–9686, 2022. 1, 2, 3, 6

[25] A Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*, 2009. 6

[26] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 6

[27] Kimin Lee, Sukmin Yun, Kibok Lee, Honglak Lee, Bo Li, and Jinwoo Shin. Robust inference via generative classifiers for handling noisy labels. In *Proc. ICML*, pages 3763–3772, 2019. 7

[28] Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *Proc. CVPR*, pages 5447–5456, 2018. 2, 7

[29] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *Proc. ICLR*, 2019. 1, 2, 3, 6, 7, 8

[30] Junnan Li, Caiming Xiong, and Steven CH Hoi. Learning from noisy data with robust representation learning. In *Proc. ICCV*, pages 9485–9494, 2021. 1, 2, 3, 7, 8

[31] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. In *arXiv preprint arXiv:1708.02862*, 2017. 1, 7

[32] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. In *Proc. NeurIPS*, volume 33, pages 20331–20342, 2020. 1, 2, 7, 8

[33] Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. In *Proc. ICML*, pages 6226–6236, 2020. 6

[34] Eran Malach and Shai Shalev-Shwartz. Decoupling" when to update" from" how to update". In *Proc. NeurIPS*, volume 30, 2017. 2, 6, 7

[35] Kento Nishi, Yi Ding, Alex Rich, and Tobias Hollerer. Augmentation strategies for learning with noisy labels. In *Proc. CVPR*, pages 8022–8031, 2021. 2, 3, 7, 8

[36] 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 *Proc. CVPR*, pages 1944–1952, 2017. 6, 7

[37] Fahad Sarfraz, Elahe Arani, and Bahram Zonooz. Noisy concurrent training for efficient learning under label noise. In *Proc. WACV*, pages 3159–3168, 2021. 1, 2, 6

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *arXiv preprint arXiv:1409.1556*, 2014. 7

[39] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Proc. NeurIPS*, volume 33, pages 596–608, 2020. 2, 3

[40] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. Selfie: Refurbishing unclean samples for robust deep learning. In *Proc. ICML*, pages 5907–5915, 2019. 1, 7

[41] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE TNNLS*, 2022. 6

[42] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proc. AAAI*, 2017. 7

[43] Cheng Tan, Jun Xia, Lirong Wu, and Stan Z Li. Co-learning: Learning from noisy labels with self-supervision. In *Proc. ACM MM*, pages 1405–1413, 2021. 7

[44] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *Proc. CVPR*, pages 5552–5560, 2018. 1, 2

[45] Arash Vahdat. Toward robustness against label noise in training deep discriminative neural networks. In *Proc. NeurIPS*, volume 30, 2017. 2

[46] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In *Proc. CVPR*, pages 839–847, 2017. 2

[47] Hongxin. Wei, Lei. Feng, Xiangyu Chen, and Bo. An. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proc. CVPR*, 2020. 2, 6, 8

[48] Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. In *Proc. NeurIPS*, volume 33, pages 7597–7610, 2020. 6

[49] Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? In *Proc. NeurIPS*, volume 32, 2019. 6

[50] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proc. CVPR*, pages 2691–2699, 2015. 2, 7

[51] Yilun Xu, Peng Cao, Yuqing Kong, and Yizhou Wang. L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise. In *Proc. NeurIPS*, volume 32, 2019. 8

[52] Yan Yan, Rómer Rosales, Glenn Fung, Ramanathan Subramanian, and Jennifer Dy. Learning from multiple annotators with varying expertise. *Machine learning*, 95(3):291–327, 2014. 1

[53] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption. In *Proc. ICML*, pages 7164–7173, 2019. 1, 2, 6

[54] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *Proc. NeurIPS*, volume 34, pages 18408–18419, 2021. 2

[55] 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, 2

[56] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. In *Proc. ICLR*, 2018. 1, 2, 5, 6, 7

[57] Yikai Zhang, Songzhu Zheng, Pengxiang Wu, Mayank Goswami, and Chao Chen. Learning with feature-dependent label noise: A progressive approach. In *Proc. ICLR*, volume 9, 2021. 1, 2, 7

[58] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Proc. NeurIPS*, volume 31, 2018. 1, 2, 5, 6, 7

[59] Ganlong Zhao, Guanbin Li, Yipeng Qin, Feng Liu, and Yizhou Yu. Centrality and consistency: two-stage clean samples identification for learning with instance-dependent noisy labels. In *arXiv preprint arXiv:2207.14476*, 2022. 3, 6, 7, 8

[60] Xiong Zhou, Xianming Liu, Chenyang Wang, Deming Zhai, Junjun Jiang, and Xiangyang Ji. Learning with noisy labels via sparse regularization. In *Proc. ICCV*, pages 72–81, 2021. 1, 2

[61] Xiong Zhou, Xianming Liu, Deming Zhai, Junjun Jiang, and Xiangyang Ji. Asymmetric loss functions for noise-tolerant learning: Theory and applications. *IEEE TPAMI*, 2023. 2

[62] Zhaowei Zhu, Tongliang Liu, and Yang Liu. A second-order approach to learning with instance-dependent label noise. In *Proc. CVPR*, pages 10113–10123, 2021. 6, 8