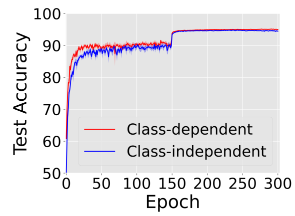
Class-Independent Regularization for Learning with Noisy Labels

Rumeng Yi[1], Dayan Guan[2], Yaping Huang[1*] and Shijian Lu[3]

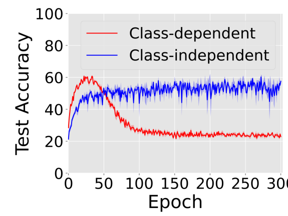
[1]Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing Jiaotong University, China
[2]Mohamed bin Zayed University of Artificial Intelligence, UAE
[3]School of Computer Science and Engineering, Nanyang Technological University, Singapore
rumengyi@bjtu.edu.cn, dayan.guan@mbzuai.ac.ae, yphuang@bjtu.edu.cn, shijian.lu@ntu.edu.sg

Abstract

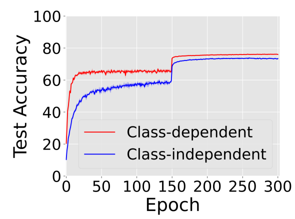
Training deep neural networks (DNNs) with noisy labels often leads to poorly generalized models as DNNs tend to memorize the noisy labels in training. Various strategies have been developed for improving sample selection precision and mitigating the noisy label memorization issue. However, most existing works adopt a class-dependent softmax classifier that is vulnerable to noisy labels by entangling the classification of multi-class features. This paper presents a class-independent regularization (CIR) method that can effectively alleviate the negative impact of noisy labels in DNN training. CIR regularizes the class-dependent softmax classifier by introducing multi-binary classifiers each of which takes care of one class only. Thanks to its class-independent nature, CIR is tolerant to noisy labels as misclassification by one binary classifier does not affect others. For effective training of CIR, we design a heterogeneous adaptive co-teaching strategy that forces the class-independent and class-dependent classifiers to focus on sample selection and image classification, respectively, in a cooperative manner. Extensive experiments show that CIR achieves superior performance consistently across multiple benchmarks with both synthetic and real images. Code is available at https://github.com/RumengYi/CIR.



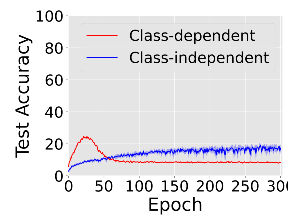
(a) CIFAR-10 with Clean Labels



(b) CIFAR-10 with Noisy Labels



(c) CIFAR-100 with Clean Labels



(d) CIFAR-100 with Noisy Labels

Figure 1: Quantitative comparisons of ***Class-independent*** multi-binary classifier with standard ***Class-dependent*** softmax classifier on CIFAR-10 and CIFAR-100 datasets with ***Clean*** and ***Noisy*** labels. Note that 80% symmetric noise is applied to generate ***Noisy*** labels, which is a challenging setting in existing works.

Introduction

Deep Neural Networks (DNNs) have achieved remarkable success in the computer vision community thanks to the large-scale datasets with precisely human-annotated labels (Chen et al. 2018) (Girshick 2015). However, collecting such high-quality annotations is extremely expensive and time-consuming, which may not be feasible in practice. Two alternative solutions are crowd-sourcing from non-experts and online queries by search engines. Unfortunately, these low-cost approaches inevitably introduce noisy labels. Recent studies have shown that DNNs can easily overfit to noisy labels and result in poor generalization performance (Zhang et al. 2017). Therefore, attention has been concentrated on how to learn with noisy labels.

Recent studies have reached a consensus for learning from noisy labels by jointly minimizing the negative impact of noisy samples and maximizing the exploitation of clean samples. An active research direction is training DNNs with selected or reweighted training samples (Kim et al. 2021a) (Ren et al. 2018) (Malach and Shalev-Shwartz 2017) (Liu et al. 2020) (Wang et al. 2020), where the challenge is to design a proper criterion for identifying clean samples. Existing criteria are mainly divided into two types: loss-based criterion (*i.e.,* small-loss (Han et al. 2018) (Yu et al. 2019) and Gaussian Mixture Model (GMM) (Li, Socher, and Hoi 2020)) and consistency-based criterion (*i.e.,* prediction consistency between two networks (Wei et al. 2020) (Liang et al. 2022) or two views (Yi and Huang 2021)). Although promising performance gains have been witnessed by employing these sample selection strategies, they heavily rely on the predictions from DNNs commonly

* Corresponding Author

trained with softmax cross-entropy loss. However, the standard softmax classifier is sensitive to noisy labels due to its class-dependent property, *i.e.,* the misclassification of one class penalizes the activation on others (Chen et al. 2022) (Chen et al. 2019), which not only outputs misleadingly high confidences on noisy data, but also affects the training of other classes, and eventually degrades the purity of the selected clean samples.

Given this insight, we propose a novel Class-Independent Regularization (CIR) by introducing multi-binary classifier for sample selection, which reformulates the K-way multi-class classification into K binary classification. Specifically, each binary classifier learns to distinguish each individual class versus all the rest of classes together. Due to the nonexclusive activation across different classes, multi-binary classifier are class-independent and more robust for noisy labels. To validate this claim, we design a toy experiment by training DNNs with the standard softmax classifier (SSC) and the multi-binary classifier (MBC), respectively, on the CIFAR-10 and CIFAR-100 datasets with different noise rates. The experimental results are shown in Fig. 1. When the training labels are clean, SSC outperforms MBC slightly as illustrated in Fig. 1 (a) and (c), demonstrating that SSC should be retained when classifying samples with clean labels after the sample selection procedure. When the training labels are noisy, MBC surpasses SSC with large margins under the higher noise rates as illustrated in Fig. 1 (b) and (d), demonstrating the superiority of MBC in learning with noisy labels during the sample selection procedure (more results are shown in the section of experiments). Based on these two empirical demonstrations, we further develop a heterogeneous adaptive co-teaching strategy by coupling MBC in the sample selection procedure and SSC in the image classification procedure. In summary, our contribution is three-fold:

- We propose a class-independent regularization (CIR) method that addresses the negative impact of the standard class-dependent softmax classifier in noisy label learning during sample selection.
- We specially design a heterogeneous adaptive co-teaching strategy to cooperate the class-independent multi-binary classifier with the standard class-dependent softmax classification, which can mutually promote the sample selection and image classification in a cooperative manner.
- We conduct comprehensive experiments on the synthetic and real-world noise benchmarks and the experimental results demonstrate that our method achieves the state-of-the-art performance.

Related Works

Learning from Noisy Labels

Learning from noisy labels can be divided into two categories (Huang et al. 2019): (1) directly training noise-robust models and (2) detecting noisy labels and then reducing their impacts. The former typically focuses on designing noise-robust objective functions (Zhang and Sabuncu 2018) (Wang et al. 2019) or regularizations (Zhang et al. 2020) to reduce the effect of the overfitting on noisy labels, but these methods do not perform well under high noise ratios (Bai et al. 2021). In the solution of the latter, potential noisy labels are first detected, and then removed them from the training set or fed to the model after corrected them. However, the challenge is to find a proper criterion for identifying clean samples. Existing methods roughly fall on two types: loss-based criterion and consistency-based criterion. The representative approaches of the former are small-loss criterion (Han et al. 2018) (Yu et al. 2019), which selects a human-defined proportion of small-loss samples as clean ones, and Gaussian Mixture Model (GMM) criterion (Li, Socher, and Hoi 2020), which fits GMM to the sample losses to model the distribution of clean and noisy samples. The representative approaches of the latter are prediction consistency, which partitions the training data into clean and noisy subsets based on the consistent predictions of two networks (Wei et al. 2020) (Liang et al. 2022) or two views (Yi and Huang 2021).

However, the above methods rely heavily on the predictions from DNNs. We argue that the standard softmax classifier in DNNs is vulnerable to noisy labels due to its class-dependent nature, *i.e.,* the misclassified score of one class suppresses the activation of others, which affects the performance of sample selection. To alleviate the negative impact of noisy labels, we introduce a class-independent multi-binary classifier to regularize the class-dependent standard softmax classifier.

Multi-binary Classifier Training

Multi-binary classifier (MBC) is widely used in open-set recognition (Saito, Kim, and Saenko 2021) and open-set domain adaptation (Zhu and Li 2021) (Saito and Saenko 2021) (Liu et al. 2019) (Bucci, Loghmani, and Tommasi 2020) tasks to identify unknown classes samples. In open-set scenario, there exists outliers that do not belong to the known classes in the training dataset, so the above methods adopt MBC to learn a boundary between inliers and outliers for each class. If all of the binary classifiers regard the input as negative, this sample has a high probability of belonging to an unknown class. In this way, they leverage the MBC to capture the notion of "none of the above", which avoids the closed-world assumption of the standard softmax classifier.

Different from the above methods, we leverage the class-independent property of MBC to regularize the class-dependent standard softmax classifier (SSC). Specifically, the SSC encourages to improve the output of ground truth and penalizes all others simultaneously, when the supervision is noisy, the classification scores of all classes will be affected due to the class-dependent property in SSC, resulting in overfitting to noisy labels. However, the MBC can alleviate this problem. The binary cross-entropy used in MBC is a nonexclusive activation function, which is dedicated to recognizing one class only and misclassification from one class will not affect others, improving the ability of identifying the noisy labels during sample selection.

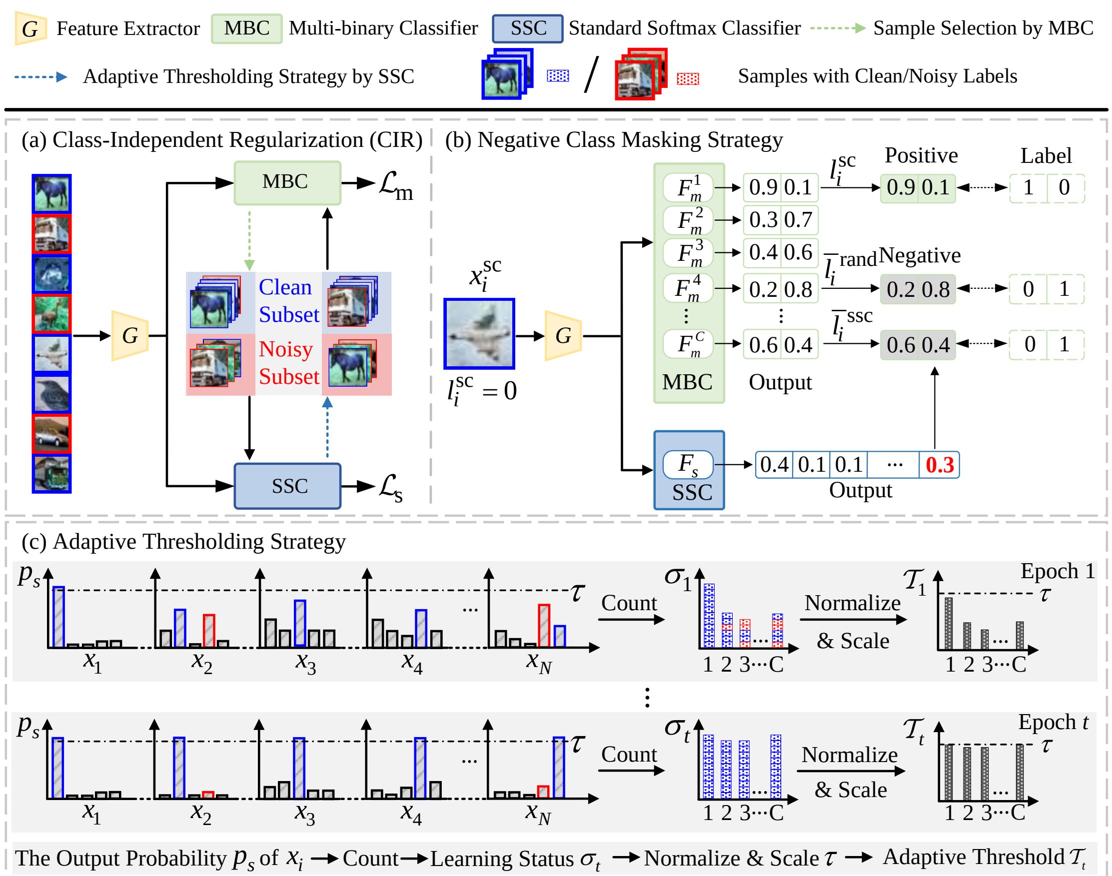


Figure 2: (a) Overview of the proposed **Class-Independent Regularization (CIR)**, which develops a heterogeneous adaptive co-teaching strategy to cooperate multi-binary classifier (MBC) and standard softmax classifier (SSC) to select data with possibly clean labels for each other. For the training of MBC, the clean samples are collected according to the prediction confidence of the SSC by using the (c) **Adaptive Thresholding Strategy**, which dynamically set the threshold for each class according to their learning status, and then the MBC is trained by (b) **Negative Class Masking Strategy**, which makes MBC learn an effective boundary among the positive and the nearest negative classes by masking and only remaining two kinds of negative classes, *i.e.,* the most difficult class $\bar{l}^{ssc}$ for SSC, and the randomly selected class $\bar{l}^{rand}$. Subsequently, the clean samples are identified according to whether the predictions of the binary classifier with maximum confidence are consistent with their given labels. Finally, SSC utilizes clean subset as labeled data and noisy subset as unlabeled data to perform semi-supervised learning.

Class-Independent Regularization

Problem Definition

We consider a classification problem with a training set $\mathcal{D} = \{(x_1, y_1), ..., (x_N, y_N)\}$, where x_i is an image and $y_i \in \{0,1\}^C$ is a one-hot label over C classes which may contains noise. Let G and F_s denote the feature extractor and standard softmax classifier (SSC) of DNNs, respectively. Therefore, the model's output softmax probability of x_i is $p_s(x_i) = F_s(G(x_i))$. In general, the objective function is empirical risk of cross-entropy loss, which is formulated by:

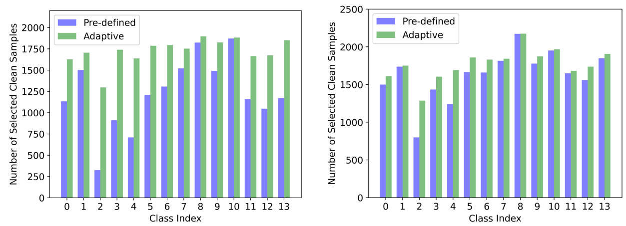
$$\mathcal{L}_c = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log p_s(x_i), \tag{1}$$

where N is the total number of samples and $\cdot$ denotes dot product. Since y_i contains noise, the model will overfit the noisy labels and result a poor classification performance.

Existing methods try to divide training data into clean and noisy subsets by designing a criterion for identifying clean samples, but they rely heavily on the predictions from SSC. We argue that the class-dependent nature of SSC might enlarge the effects of noisy labels. Therefore, we propose a class-independent regularization equipped with a heterogeneous adaptive co-teaching strategy to mitigate the negative impact of class-dependent property in SSC.

Heterogeneous Adaptive Co-teaching

The illustration of the proposed CIR is given in Fig. 2 (a). Different from the traditional co-teaching strategy that deploys two networks with the same architecture to find possibly clean samples for each other (Han et al. 2018) (Yu et al. 2019), the proposed CIR employs a heterogeneous adap-



(a) Clean set after warm-up (b) Clean set after 10 epochs

Figure 3: Class distribution in the selected clean set after (a) warm-up and (b) 10 epochs on Clothing1M dataset.

tive co-teaching strategy to learn with noisy labels. Specifically, for the training of MBC, the clean samples are collected according to the prediction confidence of the SSC by an adaptive thresholding strategy. To make MBC learn an effective boundary among the positive and the nearest negative classes, a negative class masking strategy is applied to keep an appropriate number of negative classes for training. Subsequently, the clean samples are identified according to whether the predictions of the binary classifier with maximum confidence are consistent with their given labels. Finally, the SSC utilizes the clean subset as labeled data and the noisy subset as unlabeled data to perform semi-supervised learning. In this cooperative manner, MBC and SSC share the same feature extractor, the network therefore can learn better feature representations by using SSC to perform semi-supervised learning, which in turn promotes the discriminative ability of MBC to distinguish clean samples from noisy ones.

Multi-binary Classifier Training To alleviate the negative effect of class-dependent SSC during sample selection, we introduce MBC to regularize the SSC. Let F_m represent MBC with C classes as $F_m = \{F_m^1, ..., F_m^C\}$, where F_m^k is the k-th binary classifier with output $p_m^k(x_i) = F_m^k(G(x_i))$. $p_m^k(z = 0|x_i)$ and $p_m^k(z = 1|x_i)$ denote the output probability that the instance x_i belongs to the k-th class or not, respectively, where $p_m^k(z = 0|x_i) + p_m^k(z = 1|x_i) = 1$.

The clean subset used to train MBC is selected by SSC. A naive way is to set a pre-defined threshold for all classes to cut off high-confidence samples according to the SSC's prediction. However, in the challenging dataset such as Clothing 1M (Xiao et al. 2015), since some classes may be inherently more difficult to learn than others, it is not suitable to set a fixed threshold for all classes, which may lead to few clean samples selected in some hard classes. As shown in Fig. 3(a), the pre-defined threshold is set to 0.4 for all classes to select clean samples with high confidence after warm-up stage, but we can observe that 81.8% samples are selected from class-10, whereas only 14% samples from class-2 are selected. It verifies that this simple strategy opts to only select high-quality clean data for training, but ignores a considerable amount of clean data with lower confidence, especially at the early training stage. To address this issue, we design an adaptive thresholding strategy to dynamically determine the threshold for each class according to their learning status. In this way, SSC can provide more clean samples for MBC during training as shown in Fig. 3(b).

As shown in Fig. 2 (c), the learning status of each class can be reflected by counting the number of samples whose predictions of SSC fall into this class, and meanwhile their confidences are above a fixed pre-defined threshold τ:

$$\sigma_t(c) = \sum_{i=1}^{N} \mathbb{1}(\arg\max(p_{s,t}(x_i)) = c) \cdot \mathbb{1}(\max(p_{s,t}(x_i)) \geq \tau), \tag{2}$$

where $\sigma_t(c)$ represents the learning status of class c at training epoch t, and $p_{s,t}(x_i)$ is the prediction of SSC for sample x_i at training epoch t. The larger $\sigma_t(c)$ means the better learning status of class c. Then we normalize the $\sigma_t(c)$ to $[0, 1]$ and use the normalized $\sigma_t(c)$ to scale the fixed pre-defined threshold τ, which is formulated by:

$$\mathcal{T}_t(c) = \frac{\sigma_t(c)}{\max\limits_c \sigma_t} \cdot \tau, \tag{3}$$

where $\mathcal{T}_t(c)$ is a threshold of class c at training epoch t, and can be adaptively adjusted during the training process according to the learning status. A smaller $\sigma_t(c)$ means the class is hard to learn, therefore we set a lower threshold $\mathcal{T}_t(c)$ to select clean samples for this class. As the number of training epochs increases, all classes are well trained and their thresholds will all approach the fixed threshold τ. At training epoch t, given the image x_i and its label $l_i \in \{1, ..., C\}$, we can obtain the clean subset as follow:

$$\begin{aligned}\mathcal{D}_c^s =& \{(x_i^{sc}, y_i^{sc}) | \arg\max(p_{s,t}(x_i)) = l_i \text{ and} \\ & \max(p_{s,t}(x_i)) \geq \mathcal{T}_t(\arg\max(p_{s,t}(x_i)))\}.\end{aligned} \tag{4}$$

Given clean subset $\mathcal{D}_c^s = \{(x_1^{sc}, y_1^{sc}), ..., (x_{N_{sc}}^{sc}, y_{N_{sc}}^{sc})\}$, where N_{sc} is the total number of selected clean samples, we apply a negative class masking strategy for MBC to learn an effective boundary among positive and the nearest negative classes. As shown in Fig. 2 (b), for each training sample, the corresponding negative classes are remained in two manners: (1) The class $\bar{l}_i^{\text{ssc}}$ that the SSC is most difficult to distinguish, *i.e.,* the class is different from the ground-truth but having the largest prediction score in SSC. (2) The class $\bar{l}_i^{\text{rand}}$ that randomly selected from the category set excluding the ground-truth label l_i^{sc} and $\bar{l}_i^{\text{ssc}}$. Therefore, the loss function used for training the MBC can be formulated as:

$$\begin{aligned}\mathcal{L}_{\text{m}} =& \frac{1}{N_{sc}} \Big[\sum_{i=1}^{N_{sc}} - \log(p_m^{l_i^{\text{sc}}}(z = 0|x_i^{sc})) \\ & - \sum_{k \in \bar{l}_i^{\text{sc}}} \log(p_m^k(z = 1|x_i^{sc})) \Big],\end{aligned} \tag{5}$$

where $\bar{l}_i^{\text{sc}} = \{\bar{l}_i^{\text{ssc}}, \bar{l}_i^{\text{rand}}\}$.

Standard Softmax Classifier Training After each training of MBC, clean samples are selected according to whether the predictions of the binary classifier with maximum confidence are consistent with their given labels. Given the image x_i and its label $l_i \in \{1, ..., C\}$, we can obtain the clean subset as follow:

$$\mathcal{D}_c^m = \{(x_i^{mc}, y_i^{mc}) | \arg\max (p_m^k(z=0|x_i)) = l_i\}, \quad (6)$$

and the noisy subset is $\mathcal{D}_n^m = \mathcal{D} \backslash \mathcal{D}_c^m$. Then SSC utilizes the clean subset $\mathcal{D}_c^m$ as labeled dataset and the noisy subset $\mathcal{D}_n^m$ as unlabeled dataset to perform semi-supervised learning.

Following FixMatch (Sohn et al. 2020), the supervised loss $\mathcal{L}_{\text{sup}}$ is the standard cross-entropy loss on $\mathcal{D}_c^m = \{(x_1^{mc}, y_1^{mc}), ..., (x_{N_{mc}}^{mc}, y_{N_{mc}}^{mc})\}$, which is formulated by:

$$\mathcal{L}_{\text{sup}} = -\frac{1}{N_{mc}} \sum_{i=1}^{N_{mc}} y_i^{mc} \cdot \log(p_s(x_i^{mc})). \quad (7)$$

The unsupervised loss $\mathcal{L}_{\text{unsup}}$ on $\mathcal{D}_n^m = \{x_1^{mn}, ..., x_{N_{mn}}^{mn}\}$ is formulated by:

$$\mathcal{L}_{\text{unsup}} = \frac{1}{N_{mn}} \sum_{i=1}^{N_{mn}} \mathbb{1}(\max(q) \geq \tau_{\text{unsup}})(\hat{q} \cdot \log(p_s(\beta(x_i^{mn})))), \quad (8)$$

where $q = p_s(\alpha(x_i^{mn}))$, $\hat{q} = \arg\max(q)$ and τ_{unsup} is a threshold to determine whether a pseudo label is retained or not. α and β are the weakly-augmented and strong-augmented functions, respectively. In our experiments, the weak augmentation is a simple random cropping and horizontal flipping, while the strong augmentation is RandAugment (Cubuk et al. 2020) which contains color inversion, translation, contrast adjustment, etc.

Training and Inference In summary, combining the training of MBC and SSC together, our final objective loss function is:

$$\mathcal{L} = \mathcal{L}_{\text{m}} + \mathcal{L}_{\text{s}}, \quad (9)$$

where $\mathcal{L}_{\text{s}} = \mathcal{L}_{\text{sup}} + \mathcal{L}_{\text{unsup}}$. In the test stage, we only utilize SSC for getting the final classification score.

Experiments

Datasets and Implementation Details

Datasets and Noise Setting We extensively evaluate our approach on CIFAR-10, CIFAR-100 (Krizhevsky, Hinton et al. 2009), Clothing1M (Xiao et al. 2015) and Food101N (Lee et al. 2018) datasets. Both CIFAR-10 and CIFAR-100 contain 50K training images and 10K test images of size 32 × 32, which involve 10 classes and 100 classes, respectively. Clothing1M contains 1 million images of clothes with 14 categories. Food101N contains 310k images of food with 101 categories. For CIFAR-10 and CIFAR-100 datasets, following previous works (Li, Socher, and Hoi 2020) (Liu et al. 2020) (Bai et al. 2021), we inject two types of label noise: *symmetric* and *asymmetric* into the dataset in a specified noise rate. The symmetric label noise is generated by using a random one-hot vector to replace the ground-truth label of one sample. The asymmetric label noise is designed to mimic the structure of real-world label noise, such as CAT↔DOG, BIRD↔AIRPLANE. For real-world noisy datasets Clothing1M and Food101N, the overall label accuracy are 61.54% and 80%, respectively.

Method	Symmetric			Asymmetric				Avg.
	20%	50%	80%	10%	20%	30%	40%	
SSC	**82.7**	57.9	25.5	88.8	86.1	81.7	76.0	71.2
MBC	81.6	**73.7**	**53.5**	**90.4**	**87.0**	**86.2**	**82.6**	**79.3**
Forward (CVPR' 17)	83.1	59.4	26.2	90.4	86.7	81.9	76.7	72.1
GCE (NeurIPS' 18)	86.6	81.9	54.6	89.5	85.6	80.6	76.0	79.3
PCIL (CVPR' 19)	92.0	88.7	76.5	93.1	92.9	92.6	91.6	89.6
ELR (NeurIPS' 20)	93.8	92.6	88.0	94.4	93.3	91.5	85.3	91.3
CIR (Ours)	**95.3**	**94.0**	**88.5**	**95.4**	**94.7**	**94.2**	**93.2**	**93.6**
MixUp (ICLR' 18)	92.3	77.6	46.7	93.3	88.0	83.3	77.7	79.8
M-correct (ICML' 19)	93.8	91.9	86.6	89.6	91.8	92.2	91.2	91.0
DivideMix (ICLR' 20)	95.0	93.7	92.4	93.8	93.2	92.5	91.4	93.1
MOIT+ (CVPR' 21)	94.1	91.8	81.1	94.2	94.3	94.3	93.3	91.9
Sel-CL+ (CVPR' 22)	95.5	93.9	89.2	**95.6**	**95.2**	94.5	93.4	93.9
CIR+ (Ours)	**95.6**	**95.5**	**92.6**	95.5	95.0	**94.6**	**93.5**	**94.6**

Table 1: Comparison with state-of-the-art methods in the test accuracy (%) on CIFAR-10 dataset. The best results are in bold. CIR+ denotes the CIR with mixup data augmentation and rotation prediction.

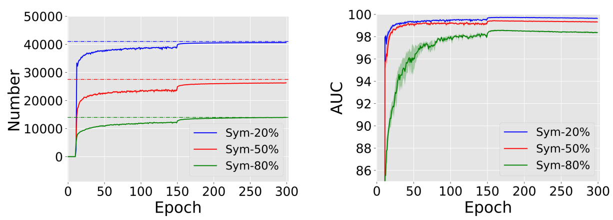
Implementation Details For experiments on CIFAR-10 and CIFAR-100 datasets, following previous work (Li, Socher, and Hoi 2020), we use an 18-layer PreAct ResNet architecture (He et al. 2016) and train it using SGD with a momentum of 0.9, a weight decay of 0.0005, and a batch size of 128. The network is trained for 300 epochs. We set the initial learning rate as 0.02, and reduce it by a factor of 10 after 150 epoch. The warm-up epochs are set to 10 for CIFAR-10 and 30 for CIFAR-100. For real-world datasets, following previous works (Li, Socher, and Hoi 2020) (Yao et al. 2021), we use ResNet-50 with ImageNet pretrained weight and train the network for 100 epochs. We set the initial learning rate as 0.002 and reduce it by a factor of 10 after 30 and 60 epochs. The warm-up epochs are set to 5, and other experiment settings are the same as CIFAR datasets. The hyperparameters τ and τ_{unsup} are selected from {0, 0.1, ..., 0.9}.

Comparison with State-of-the-art Methods

Results on CIFAR-10 and CIFAR-100 Datasets We use the conventional training with the softmax cross-entropy loss (SSC) and binary cross-entropy loss (MBC) on noisy datasets as our baselines, and compare the proposed CIR with recent state-of-the-art methods, including Forward (Patrini et al. 2017), GCE (Zhang and Sabuncu 2018), PCIL (Yi and Wu 2019), ELR (Liu et al. 2020), MixUp (Zhang et al. 2018), M-correct (Arazo et al. 2019), DivideMix (Li, Socher, and Hoi 2020), MOIT+ (Ortego et al. 2021) and Sel-CL+ (Li et al. 2022). Since the last five approaches apply MixUp data augmentation (Zhang et al. 2018) or contrastive learning (Chen et al. 2020) to reduce the risk of noise memorization, we incorporate the similar techniques to further facilitate our CIR, which called CIR+. In CIR+, we empirically find that the contrastive learning performs worse than the rotation recognition (Komodakis and Gidaris 2018), so in this paper, we introduce the rotation recognition

Method	Symmetric			Asymmetric				Avg.
	20%	50%	80%	10%	20%	30%	40%	
SSC	**61.8**	37.3	8.2	68.1	63.6	53.3	44.5	48.1
MBC	60.3	**38.2**	**17.1**	**71.5**	**69.6**	**64.9**	**53.8**	**53.6**
Forward (CVPR' 17)	61.4	37.3	9.0	68.7	63.2	54.4	45.3	48.5
GCE (NeurIPS' 18)	59.2	47.8	15.8	68.0	58.6	51.4	42.9	49.1
PCIL (CVPR' 19)	68.1	56.4	20.7	76.1	68.9	59.3	48.3	56.8
ELR (NeurIPS' 20)	74.5	70.2	45.2	75.8	74.8	73.6	70.0	69.2
CIR (Ours)	**76.8**	**73.4**	**63.0**	**77.2**	**77.1**	**76.3**	**74.0**	**74.0**
MixUp (ICLR' 18)	66.0	46.6	17.6	72.4	65.1	57.6	48.1	53.3
M-correct (ICML' 19)	73.4	65.4	47.6	67.1	64.5	58.6	47.4	60.6
DivideMix (ICLR' 20)	74.8	72.1	57.6	69.5	69.2	68.3	51.0	66.1
MOIT+ (CVPR' 21)	75.9	70.6	47.6	77.4	76.4	75.1	74.0	71.0
Sel-CL+ (CVPR' 22)	76.5	72.4	59.6	78.7	77.5	76.4	74.2	73.6
CIR+ (Ours)	**77.0**	**74.5**	**67.2**	**78.8**	**78.1**	**76.6**	**74.5**	**75.2**

Table 2: Comparison with state-of-the-art methods in the test accuracy (%) on CIFAR-100 dataset.



(a) *Num.* of Samples *vs.* epochs (b) AUC scores *vs.* epochs

Figure 4: The performance of sample selection during training on CIFAR-10 with 20%, 50% and 80% symmetric noises. (a) The number of selected samples *vs.* epochs; (b) The Area Under a Curve (AUC) scores *vs.* epochs.

as an auxiliary task for feature learning enhancement. For fair comparisons, the reported results are all obtained with one single model. We report the average test accuracy over the last 10 epochs. As shown in Table 1 and Table 2, we first observe that the test accuracy of MBC outperforms SSC in most cases. Especially on symmetric 80% and asymmetric 40%, the margin is clearer, demonstrating that MBC is more robust to noisy label than SSC.

For CIFAR-10, from moderate to severe label noise, CIR performs better than the compared methods in all cases, which exceeds the second-best method ELR by 2.3% on average accuracy. And the performance can be further boosted by CIR+, which exceeds the second-best method Sel-CL+ by 0.7% on average accuracy. For the more difficult CIFAR-100, CIR achieves a significant improvement over the second-best method ELR by 4.8% on average accuracy. Moreover, CIR+ exceeds the second-best method Sel-CL+ by 1.6% on average accuracy.

In addition, we also evaluate the performance of sample selection during training on CIFAR-10 with 20%, 50% and 80% symmetric noises, and the results are shown in Fig. 4. Fig. 4 (a) and (b) show the number of selected clean samples and Area Under a Curve (AUC) scores of our sample selection mechanism, respectively, where solid and dashed lines in (a) indicate the number of selected and real (oracle) clean samples under different noise rates. It can be observed from (a) that CIR approaches the oracle at a fast rate irrespective of the noise level, and the corresponding AUC curves in (b) also prove that CIR can distinguish clean and noisy samples accurately and comprehensively.

Clothing1M	
Methods	Acc.
SSC	69.21
MBC	**71.55**
MetaL (CVPR' 19)	73.47
PCIL (CVPR' 19)	73.49
DMix (ICLR' 20)[1]	74.30
ELR (NeurIPS' 20)	72.87
FINE (NeurIPS' 21)	74.37
UPM (AAAI' 21)	74.02
JNPL (CVPR' 21)	74.15
CAL (CVPR' 21)	74.17
CIR (Ours)	74.41
CIR+ (Ours)	**74.66**

Table 3: Comparison with state-of-the-art methods in the test accuracy (%) on Clothing1M dataset.

Food101N	
Methods	Acc.
SSC	84.51
MBC	**84.74**
CNet-hard (CVPR' 18)	83.47
CNet-soft (CVPR' 18)	83.95
DeepSelf (ICCV' 19)	85.11
MCleaner (CVPR' 19)	85.05
GJS (NeurIPS' 21)	86.56
Co-L (ACM MM' 21)	87.57
PNP-hard (CVPR' 22)	87.31
PNP-soft (CVPR' 22)	87.50
CIR (Ours)	87.65
CIR+ (Ours)	**87.81**

Table 4: Comparison with state-of-the-art methods in the test accuracy (%) on Food101N dataset.

Results on Real-world Datasets We compare CIR with two baselines (SSC and MBC) and the state-of-the-art methods, including MetaL (Li et al. 2019), PCIL (Yi and Wu 2019), DMix (Li, Socher, and Hoi 2020), ELR (Liu et al. 2020), FINE (Kim et al. 2021a), UPM (Wang et al. 2021), JNPL (Kim et al. 2021b), CAL (Zhu, Liu, and Liu 2021) CNet (Lee et al. 2018), DeepSelf (Han, Luo, and Wang 2019), MCleaner (Zhang, Wang, and Qiao 2019), GJS (Englesson and Azizpour 2021), Co-L (Tan et al. 2021) and PNP (Sun et al. 2022) on Clothing1M and Food101N datasets. For fair comparisons, the reported results are all obtained with one single model. The results are shown in Table 3 and Table 4, respectively. Similar to CIFAR, the test accuracy of MBC also outperforms SSC, especially on the more challenge dataset Clothing1M, the performance of MBC exceeds the SSC by 2.34%. Meanwhile, CIR consistently outperforms competing methods across all datasets, and the performance can be further boosted by CIR+, which exceeds the second-best methods by 0.29% and 0.24% on Clothing1M and Food101N, respectively.

Further Analysis

Ablation Study To verify the effectiveness of the CIR, the ablation study is conducted on CIFAR-100 with 50% symmetric and 40% asymmetric noise, Clothing1M and Food101N datasets. The results are shown in Table 5.

[1] The reported result of DMix (Li, Socher, and Hoi 2020) is taken from FINE (Kim et al. 2021a), which retrained the algorithm using one single network.

Dataset	CIFAR-100	
Method/Noise rate	Sym-50%	Asym-40%
(1) CIR w/o NCM	72.7	71.1
(2) CIR w/o $\bar{l}^{\text{ssc}}$	72.9	71.9
(3) CIR w/o $\bar{l}^{\text{rand}}$	72.8	72.9
(4) CIR w/o AT	73.4	73.9
(5) CIR w/o HCo-T	72.5	71.7
(6) CIR	**73.4**	**74.0**
Method/Dataset	Clothing1M	Food101N
(7) CIR w/o AT	74.23	87.54
(8) CIR	**74.41**	**87.65**

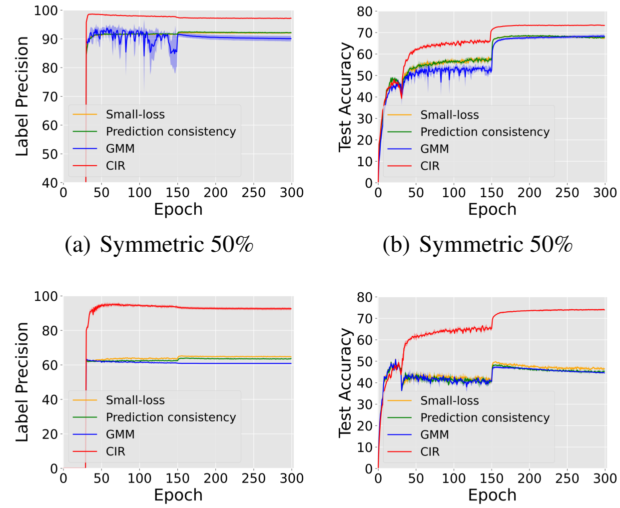
Table 5: Ablation study on each components of CIR on CIFAR-100 with 50% symmetric and 40% asymmetric noise rates, Clothing1M and Food101N datasets. Here, NCM, AT and HCo-T are the abbreviations of negative class masking strategy, adaptive thresholding strategy and heterogeneous co-teaching strategy, respectively.

Effect of Negative Class Masking (NCM) Strategy. To study the effect of NCM, we treat all classes except ground-truth label as negative classes to train the MBC, and the results are corresponding to (1). Compared with (6), the performances degrade in all two cases. We further explore the effect of two kinds of negative class $\bar{l}^{\text{ssc}}$ and $\bar{l}^{\text{rand}}$, the results are corresponding to (2) and (3). The performances also degrade when excluding any of them, demonstrating that our proposed NCM can enforce the binary classifiers learn an effective boundary among the positive and negative classes.

Effect of Adaptive Thresholding (AT) Strategy. To study the effect of AT, we select clean samples according to the fixed thresholds 0.5 and 0.9 for Sym-50% and Asym-40%, instead of an adaptive threshold in Eq.(3). The results are corresponding to (4). we can see that the performances of (4) are almost equal to (6). The reason is that the learning status of all classes on CIFAR datasets are balanced during training, therefore the AT strategy degenerates into the fixed thresholding among all classes. However, on the more challenging Clothing1M and Food101N, this strategy can improve 0.18% and 0.11%, respectively, evidencing the superiority of the AT strategy in handling real biased data.

Effect of Heterogeneous Co-teaching (HCo-T) Strategy. To study the effect of HCo-T, the clean samples used to train MBC are obtained directly by itself according to Eq.(6), and the results are corresponding to (5), the performances of two cases degrade by 0.9% and 2.3%, respectively, which demonstrates that the HCo-T strategy can not only mutually promote cooperation of sample selection and image classification, but also filter different types of errors and avoid confirmation bias in self-training.

Robustness Analysis of Different Criteria To evaluate the performance of sample selection, we compare the proposed CIR with class-dependent based methods, *i.e.,* small-loss criterion (Han et al. 2018), GMM criterion (Li, Socher, and Hoi 2020) and prediction consistency criterion (Yi and Huang 2021) using one single model on two difficult cases, *i.e.,* symmetric 50% and asymmetric 40% noise rates on CIFAR-100 dataset. We report the label precision of the selected clean samples and the corresponding test accuracy during training, and the results are shown in Fig. 5. It can be seen from (a) and (c) that CIR achieves a higher label precision only after a few epochs. Especially on symmetric 50%, the label precision is close to 1. It is worth noting that in asymmetric noise cases, the label precision of the class-dependent based methods are all below 65%, but CIR exceeds them by a large margin, which demonstrates that CIR is tolerant to noisy labels. Meanwhile, the accurate separation also provides the accurate supervisions for the subsequent training process. The corresponding test accuracy curve also verifies the effectiveness of CIR.



(a) Symmetric 50% (b) Symmetric 50%

(c) Asymmetric 40% (d) Asymmetric 40%

Figure 5: Comparison of four criteria on CIFAR-100 with symmetric 50% and asymmetric 40% noise rates. (a) and (c) are the label precision of the selected clean samples; (b) and (d) are the test accuracy corresponding to (a) and (c).

Conclusion

In this paper, we propose the Class-Independent Regularization (CIR) to alleviate the negative impact of noisy label learning. Specifically, CIR regularizes the standard class-dependent softmax classifier by introducing a class-independent multi-binary classifier, where each binary classifier is dedicated to recognizing one class only. For training CIR effectively, we design a heterogeneous adaptive co-teaching strategy that forces the class-independent and class-dependent classifiers to focus on sample selection and image classification, respectively, in a cooperative manner. Experiments on synthetic and real-world noise benchmarks demonstrate the effectiveness of CIR. Moving forward, we will explore class-independent regularization in natural language processing tasks with noisy labels.

Acknowledgements

This work is supported by National Natural Science Foundation of China (62271042, 62106017, 61906013), Beijing Natural Science Foundation (M22022, L211015), Hebei Natural Science Foundation (F2022105018), Fundamental Research Funds for the Central Universities (2019JBZ104).

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 (ICML)*, 312–321.

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

Bucci, S.; Loghmani, M. R.; and Tommasi, T. 2020. On the effectiveness of image rotation for open set domain adaptation. In *European Conference on Computer Vision (ECCV)*, 422–438. Springer.

Chen, C.; Li, O.; Tao, D.; Barnett, A.; Rudin, C.; and Su, J. K. 2019. This looks like that: deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1–12.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, 801–818.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 1597–1607.

Chen, Z.; Wang, T.; Wu, X.; Hua, X.-S.; Zhang, H.; and Sun, Q. 2022. Class re-activation maps for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 969–978.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 1–10.

Englesson, E.; and Azizpour, H. 2021. Generalized jensen-shannon divergence loss for learning with noisy labels. *Advances in Neural Information Processing Systems (NeurIPS)*, 34: 30284–30297.

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1440–1448.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 8536–8546.

Han, J.; Luo, P.; and Wang, X. 2019. Deep self-learning from noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 5138–5147.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Identity mappings in deep residual networks. In *European conference on computer vision (ECCV)*, 630–645.

Huang, J.; Qu, L.; Jia, R.; and Zhao, B. 2019. O2U-Net: A simple noisy label detection approach for deep neural networks. In *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, 3326–3334.

Kim, T.; Ko, J.; Choi, J.; Yun, S.-Y.; et al. 2021a. FINE samples for learning with noisy labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, 24137–24149.

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

Komodakis, N.; and Gidaris, S. 2018. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 1–16.

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

Lee, K.-H.; He, X.; Zhang, L.; and Yang, L. 2018. Cleannet: Transfer learning for scalable image classifier training with label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5447–5456.

Li, J.; Socher, R.; and Hoi, S. C. 2020. DivideMix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations (ICLR)*, 1–14.

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 (CVPR)*, 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 (CVPR)*, 316–325.

Liang, X.; Yao, L.; Liu, X.; and Zhou, Y. 2022. Tripartite: Tackle noisy labels by a more precise partition. In *arXiv preprint arXiv:2202.09579*, 1–16.

Liu, H.; Cao, Z.; Long, M.; Wang, J.; and Yang, Q. 2019. Separate to adapt: Open set domain adaptation via progressive separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2927–2936.

Liu, S.; Niles-Weed, J.; Razavian, N.; and Fernandez-Granda, C. 2020. Early-learning regularization prevents memorization of noisy labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, 20331–20342.

Malach, E.; and Shalev-Shwartz, S. 2017. Decoupling" when to update" from" how to update". In *Advances in Neural Information Processing Systems (NeurIPS)*, 960–970.

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 (CVPR)*, 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1944–1952.

Ren, M.; Zeng, W.; Yang, B.; and Urtasun, R. 2018. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning (ICML)*, 4334–4343.

Saito, K.; Kim, D.; and Saenko, K. 2021. OpenMatch: Open-set consistency regularization for semi-supervised learning with outliers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 25956–25967.

Saito, K.; and Saenko, K. 2021. Ovanet: One-vs-all network for universal domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9000–9009.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 596–608.

Sun, Z.; Shen, F.; Huang, D.; Wang, Q.; Shu, X.; Yao, Y.; and Tang, J. 2022. PNP: Robust learning from noisy labels by probabilistic noise prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5311–5320.

Tan, C.; Xia, J.; Wu, L.; and Li, S. Z. 2021. Co-learning: Learning from noisy labels with self-supervision. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 1405–1413.

Wang, K.; Peng, X.; Yang, J.; Lu, S.; and Qiao, Y. 2020. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6897–6906.

Wang, Q.; Han, B.; Liu, T.; Niu, G.; Yang, J.; and Gong, C. 2021. Tackling instance-dependent label noise via a universal probabilistic model. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, 10183–10191.

Wang, Y.; Ma, X.; Chen, Z.; Luo, Y.; Yi, J.; and Bailey, J. 2019. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 322–330.

Wei, H.; Feng, L.; Chen, X.; and An, B. 2020. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13726–13735.

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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2691–2699.

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 (CVPR)*, 5192–5201.

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

Yi, R.; and Huang, Y. 2021. TC-Net: Detecting noisy labels via transform consistency. *IEEE Transactions on Multimedia*, 1–14.

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 (ICML)*, 7164–7173.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2017. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 1–15.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. Mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 1–13.

Zhang, W.; Wang, Y.; and Qiao, Y. 2019. Metacleaner: Learning to hallucinate clean representations for noisy-labeled visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7373–7382.

Zhang, Z.; and Sabuncu, M. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. 1–11.

Zhang, Z.; Zhang, H.; Arik, S. O.; Lee, H.; and Pfister, T. 2020. Distilling effective supervision from severe label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9294–9303.

Zhu, R.; and Li, S. 2021. CrossMatch: Cross-classifier consistency regularization for open-set single domain generalization. In *International Conference on Learning Representations (ICLR)*, 1–17.

Zhu, Z.; Liu, T.; and Liu, Y. 2021. A second-order approach to learning with instance-dependent label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10113–10123.