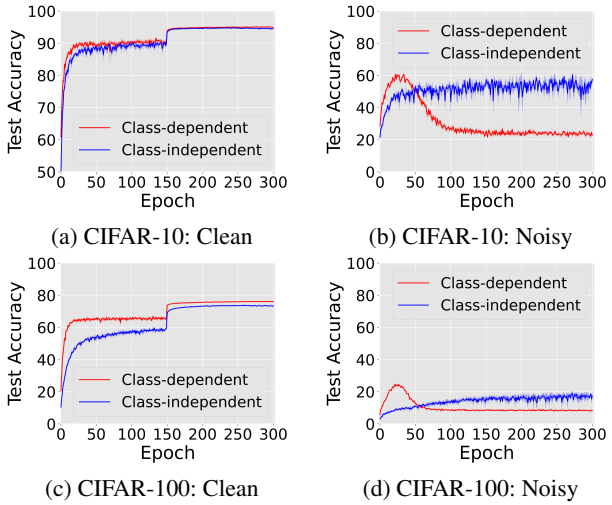
Class-Independent Regularization for Learning with Noisy Labels

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: Clean (b) CIFAR-10: Noisy

(c) CIFAR-100: Clean (d) CIFAR-100: Noisy

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 80% symmetric noisy labels.

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, 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 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

*Corresponding Author

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 CIFAR 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) 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 (SSC).

Different from the above methods, we leverage the class-independent property of MBC to regularize the 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.

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

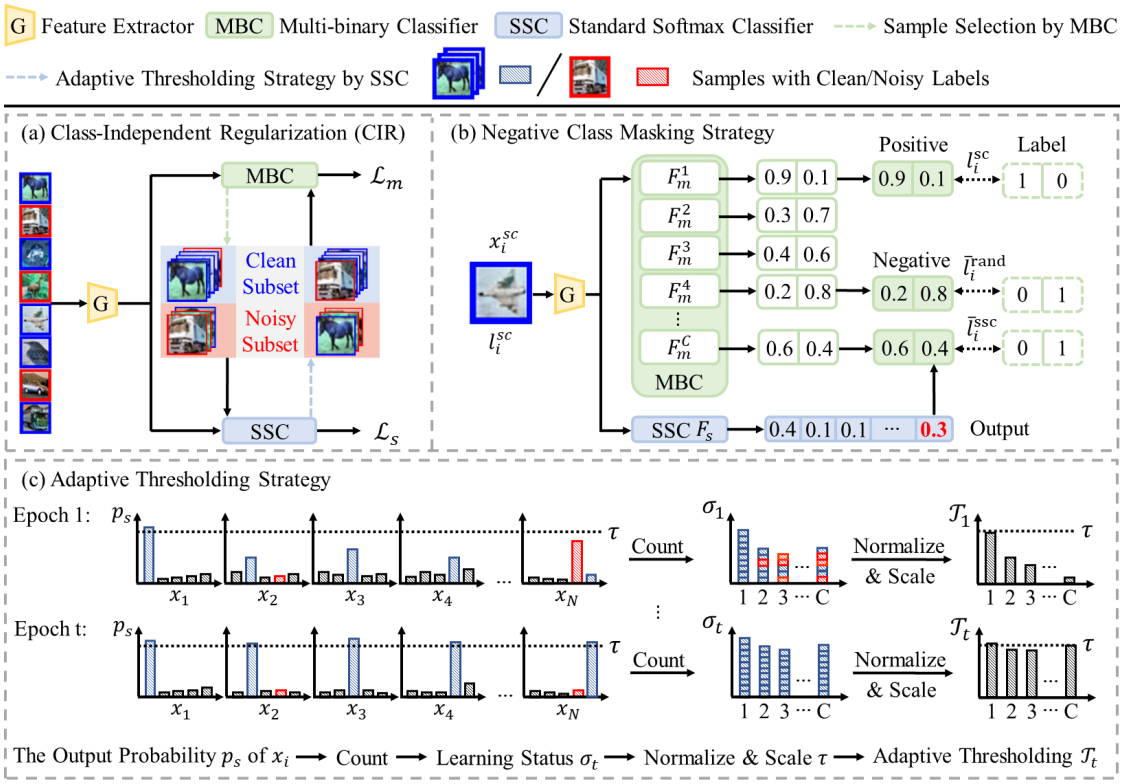


Figure 2: (a) Overview of the proposed Class-Independent Regularization (CIR), which develops a heterogeneous adaptive co-teaching strategy to cooperate multi-binary classifier (MBC) and standard softmax classifier (SSC) to select data with possibly clean labels for each other. For the training of MBC, 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 $\bar{l}^{\text{ssc}}$ for SSC, and the randomly selected class $\bar{l}^{\text{rand}}$. 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.

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 adaptive 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 (Fig. 2 (c)). To make MBC learn an effective boundary among the positive and the nearest negative classes, a negative class masking strategy (Fig. 2 (b)) 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, this strategy can only make sure that high-quality clean data contribute to the model training, while it ignores a considerable amount of other clean data with low-confidence, especially at the early stage of the training process, where only a few clean data have their prediction confidence above the threshold. To address this issue, we design an adaptive thresholding strategy to dynamically determine the threshold for each class according to their learning status. 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:

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

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:

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

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 is 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:

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

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.

Similar to DivideMix (Li, Socher, and Hoi 2020), we improve MixMatch (Berthelot et al. 2019) by label refinement and label guessing on clean and noisy samples to perform semi-supervised learning. Specifically, we first generate two copies of each sample in $\mathcal{D}_c^m$ and $\mathcal{D}_n^m$ with weak augmentation: $\hat{\mathcal{D}}_{c,d}^m = \{(\hat{x}_{1,d}^{mc}, y_1^{mc}), ..., (\hat{x}_{N_{mc},d}^{mc}, y_{N_{mc}}^{mc}); d \in (1, 2)\}$ and $\hat{\mathcal{D}}_{n,d}^m = \{\hat{x}_{1,d}^{mn}, ..., \hat{x}_{N_{mn},d}^{mn}; d \in (1, 2)\}$.

Second, we perform label refinement for the labeled sample x_i^{mc} by linearly combining the ground-truth label y_i^{mc} with the soft label $p_{s_{\text{soft}}}$ generated by SSC's prediction $p_s(\hat{x}_{i,d}^{mc})$ (averaged across two weak augmentations of x_i^{mc}), which is guided by ω_i (the prediction confidence of the binary classifier corresponding to its ground-truth label l_i^{mc}):

$$\tilde{y}_i^{mc} = \omega_i y_i^{mc} + (1 - \omega_i) p_{s_{\text{soft}}}, \tag{7}$$

where

$$p_{s_{\text{soft}}} = \frac{1}{2} \sum_{d=1}^{2} p_s(\hat{x}_{i,d}^{mc}), \tag{8}$$

$$\omega_i = \frac{1}{2} \sum_{d=1}^{2} p_m^{l_i^{mc}}(z = 0|\hat{x}_{i,d}^{mc}). \tag{9}$$

Third, we perform label guessing for the unlabeled sample x_i^{mn} by averaging the SSC's predictions of two weak augmentations to produce more reliable guessed label:

$$\tilde{y}_i^{mn} = \frac{1}{2} \sum_{d=1}^{2} p_s(\hat{x}_{i,d}^{mn}). \tag{10}$$

Besides, we also apply temperature sharpening on $\tilde{y}_i^{mc}$ and $\tilde{y}_i^{mn}$ to get $\hat{y}_i^{mc}$ and $\hat{y}_i^{mn}$.

Then we aggregate the labeled and unlabeled images with their refined and guessed labels respectively to form $\hat{\mathcal{X}}$ and

Dataset	CIFAR-10									CIFAR-100								
Methods	Symmetric				Asymmetric				Avg.	Symmetric				Asymmetric				Avg.
	20%	50%	80%	90%	10%	20%	30%	40%		20%	50%	80%	90%	10%	20%	30%	40%	
SSC	**82.7**	57.9	25.5	16.8	88.8	86.1	81.7	76.0	64.4	**61.8**	37.3	8.2	3.5	68.1	63.6	53.3	44.5	42.5
MBC	81.6	**73.7**	**53.5**	**20.4**	**90.4**	**87.0**	**86.2**	**82.6**	**71.9**	60.3	**38.2**	**17.1**	**4.3**	**71.5**	**69.6**	**64.9**	**53.8**	**47.5**
MixUp	92.3	77.6	46.7	43.9	93.3	88.0	83.3	77.7	75.4	66.0	46.6	17.6	8.1	72.4	65.1	57.6	48.1	47.7
Forward	83.1	59.4	26.2	18.8	90.4	86.7	81.9	76.7	65.4	61.4	37.3	9.0	3.4	68.7	63.2	54.4	45.3	42.8
GCE	86.6	81.9	54.6	21.2	89.5	85.6	80.6	76.0	72.0	59.2	47.8	15.8	7.2	68.0	58.6	51.4	42.9	43.9
P-correct	92.0	88.7	76.5	58.2	93.1	92.9	92.6	91.6	85.7	68.1	56.4	20.7	8.8	76.1	68.9	59.3	48.3	50.8
M-correct	93.8	91.9	86.6	68.7	89.6	91.8	92.2	91.2	88.2	73.4	65.4	47.6	20.5	67.1	64.5	58.6	47.4	55.6
DivideMix	95.0	93.7	92.4	74.2	93.8	93.2	92.5	91.4	90.8	74.8	72.1	57.6	29.2	69.5	69.2	68.3	51.0	61.5
ELR	93.8	92.6	88.0	63.3	94.4	93.3	91.5	85.3	87.8	74.5	70.2	45.2	20.5	75.8	74.8	73.6	70.0	63.1
CIR	**95.6**	**95.1**	**93.0**	**83.5**	**95.9**	**94.7**	**94.0**	**91.7**	**92.9**	**76.5**	**73.2**	**59.3**	**35.3**	**78.2**	**77.6**	**76.4**	**73.5**	**68.8**
GCE+	90.0	89.3	73.9	36.5	91.1	87.3	82.2	78.1	78.6	68.1	53.3	22.1	8.9	70.2	60.2	52.6	44.1	47.4
ELR+	94.4	93.0	88.3	86.2	95.0	94.7	94.4	93.3	92.4	76.2	71.9	57.9	40.8	77.2	75.5	74.3	70.4	68.0
MOIT+	94.1	91.8	81.1	74.7	94.2	94.3	94.3	93.3	89.7	75.9	70.6	47.6	41.8	77.4	76.4	75.1	74.0	67.4
Sel-CL+	95.5	93.9	89.2	81.9	95.6	95.2	94.5	**93.4**	92.4	76.5	72.4	59.6	48.8	78.7	77.5	76.4	74.2	70.5
CIR+	**96.0**	**95.7**	**94.4**	**92.6**	**96.0**	**95.3**	**95.0**	92.5	**94.7**	**77.4**	**75.0**	**66.8**	**53.8**	**78.8**	**78.4**	**77.6**	**74.3**	**72.8**

Table 1: Comparison with state-of-the-art methods in the test accuracy (%) on CIFAR dataset. The best results are in bold.

$\hat{\mathcal{U}}$, and use MixMatch to generate $\mathcal{X}^{'}$ and $\mathcal{U}^{'}$. The semi-supervised losses are formulated as:

$$\mathcal{L}_{\text{sup}} = \frac{1}{|\mathcal{X}'|} \sum_{x,y \in \mathcal{X}'} y \cdot \log p_s(x), \tag{11}$$

$$\mathcal{L}_{\text{unsup}} = \frac{1}{|\mathcal{U}'|} \sum_{x,y \in \mathcal{U}'} \|y - p_s(x)\|_2^2. \tag{12}$$

In summary, the total loss for training SSC can be computed as follows:

$$\mathcal{L}_{\text{s}} = \mathcal{L}_{\text{sup}} + \mathcal{L}_{\text{unsup}} + \mathcal{L}_{\text{reg}}, \tag{13}$$

where $\mathcal{L}_{\text{reg}}$ is a regularization term to regularize the network's output across all samples similar to DivideMix.

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

$$\mathcal{L} = \mathcal{L}_{\text{m}} + \mathcal{L}_{\text{s}}. \tag{14}$$

In the test stage, we utilize the ensemble of SSC and MBC for getting the final classification score.

Experiments

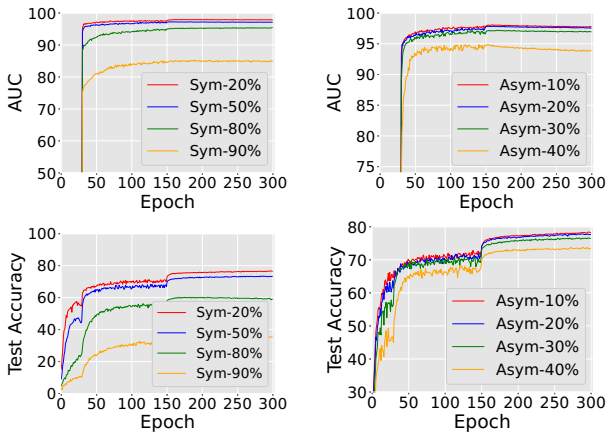
Datasets and Implementation Details

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.

Implementation Details For experiments on CIFAR 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 100 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 80 epochs. We set the initial learning rate as 0.002 and reduce it by a factor of 10 after 30 epochs. The warm-up epochs are set to 5, and other experiment settings are the same as CIFAR datasets. The hyperparameter τ used in CIFAR is selected from $\{0.5, ..., 0.9\}$, and used in Clothing1M and Food101N are 0.4 and 0.2, respectively.

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, includ-



(a) Symmetric noises (b) Asymmetric noises

Figure 3: The performance of sample selection during training on CIFAR-100 with (a) symmetric and (b) asymmetric noises. The first row are the Area Under a Curve (AUC) scores *vs.* epochs, and the second row are the corresponding test accuracy *vs.* epochs.

Clothing1M		Food101N	
Methods	Acc.	Methods	Acc.
SSC	69.21	SSC	84.51
MBC	**71.55**	MBC	**84.74**
MetaL	73.47	CNet-hard	83.47
P-correct	73.49	CNet-soft	83.95
DivideMix	74.30	DeepSelf	85.11
ELR	72.87	MCleaner	85.05
FINE	74.37	AFM	87.23
UPM	74.02	GJS	86.56
JNPL	74.15	PNP-hard	87.31
CAL	74.17	PNP-soft	87.50
CIR (Ours)	**74.53**	CIR (Ours)	**87.71**

Table 2: Comparison with state-of-the-art methods in the test accuracy (%) on Clothing1M and Food101N datasets.

ing MixUp (Zhang et al. 2018), Forward (Patrini et al. 2017), GCE (Zhang and Sabuncu 2018), P-correct (Yi and Wu 2019), M-correct (Arazo et al. 2019), DivideMix (Li, Socher, and Hoi 2020), ELR (Liu et al. 2020), GCE + (Zhang and Sabuncu 2018), ELR+ (Liu et al. 2020), MOIT+ (Ortego et al. 2021) and Sel-CL+ (Li et al. 2022). Since the last four approaches apply 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 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, we first observe that the test accuracy of MBC outperforms SSC in most cases. The margin is clearer, especially on symmetric 80% and asymmetric 40%, demonstrating that MBC is more robust to noisy labels.

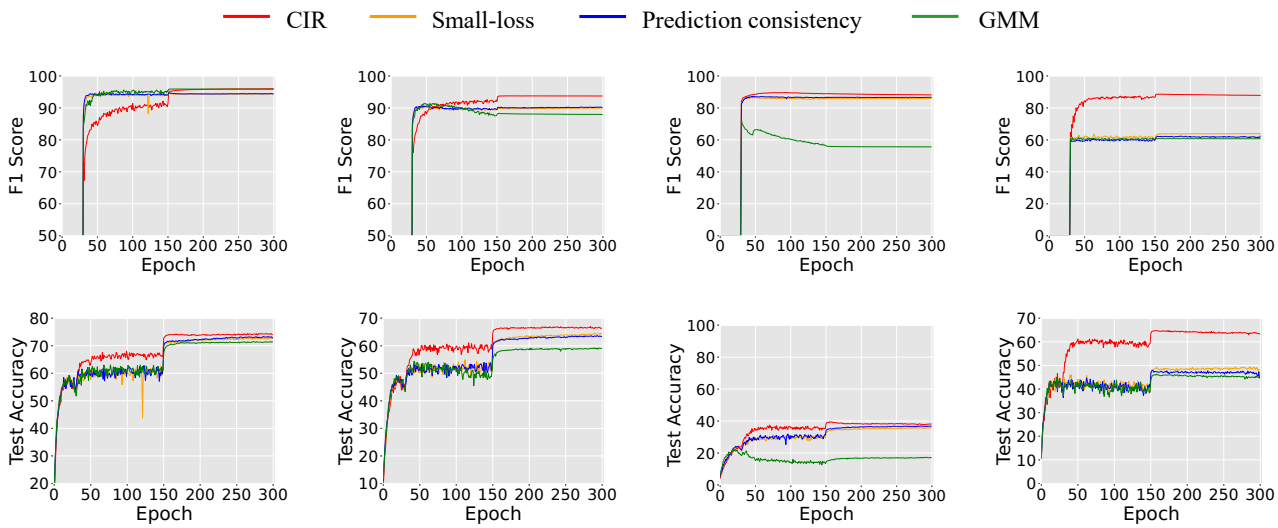
For CIFAR-10, from moderate to severe label noise, CIR performs better than the compared methods in most cases, which exceeds the second-best method DivideMix by 2.1% on average accuracy. And the performance can be further boosted by CIR+, which exceeds the second-best method Sel-CL+ by 2.3% on average accuracy. For the more difficult CIFAR-100, CIR achieves a significant improvement over the second-best method ELR by 5.7% on average accuracy. Moreover, CIR+ exceeds the second-best method Sel-CL+ by 2.3% on average accuracy. In addition, we also evaluate the performance of sample selection during training on CIFAR-100 with all noise rates, and the results are shown in Fig. 3. We show the Area Under a Curve (AUC) for clean/noisy classification from MBC during training (the first row), and those curves prove that CIR can distinguish clean and noisy samples accurately and comprehensively as training proceeds, even for high noise ratio, and the corresponding test accuracy curve (the second row) also verifies the effectiveness of CIR.

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), P-correct (Yi and Wu 2019), DivideMix (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), AFM (Peng et al. 2020), GJS (Englesson and Azizpour 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 2. 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 exceeds the second-best methods by 0.16% and 0.21% on Clothing1M and Food101N, respectively.

Further Analysis

Ablation Study To verify the effectiveness of the CIR, the ablation studies are conducted on CIFAR-10 with symmetric 20% (Sym-20%) and 50% (Sym-50%), asymmetric 20% (Asym-20%) and 30% (Asym-30%) noise rates, respectively. The results are shown in Table 3.

As the baseline of CIR, we use the clean samples selected by MBC to train SSC, where the samples used to train MBC are selected by a pre-defined threshold, and the MBC are trained only using l^{rand} as negative class. As shown in (1) of Table 3, the test accuracies are 93.5% (Sym-20%), 90.2% (Sym-50%), 93.2% (Asym-20%) and 91.8% (Asym-30%), respectively, which exceed several approaches in Table 1, demonstrating that MBC can select clean samples



(a) Symmetry-20% (b) Symmetry-50% (c) Symmetry-80% (d) Asymmetry-40%

Figure 4: Comparison of four criteria on CIFAR-100 with symmetric 20%, 50%, 80% and asymmetric 40% noise rates.

accurately. Since the MBC provides accurate supervision for SSC training, after applying semi-supervised learning (SSL), the performance can be further boosted (the results are shown in (2)). Especially on symmetric noise, the network regards more noisy samples as unlabeled data for training and the performance is improved by 1.9% (Sym-20%) and 4.0% (Sym-50%), respectively.

To study the effect of adaptive thresholding (AT) strategy, we select clean samples according to the adaptive threshold instead of a pre-defined threshold on the basis of (2), and the results are shown in (3). Compared with (2), the performance is further improved, especially on asymmetric noise. A possible reason is that the network is more difficult to identify clean samples from the asymmetric noisy dataset because this noise type is designed to mimic the structure of real-world label noise, resulting a large number of clean samples have low-confidence. In this case, the proposed AT strategy provides more effective supervision at the early stage, and lays a solid foundation for the subsequent training.

To study the effect of negative class masking (NCM) strategy, we treat $\bar{l}^{\text{ssc}}$ as negative class to train the MBC on the basis of (3), and the results are shown in (4). Compared with (3), the performance is further improved by 0.1%-0.7%, demonstrating that the proposed NCM can enforce the binary classifiers learn an effective boundary among the positive and negative classes.

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 four cases, *i.e.,* symmetric 20%, 50%, 80% and asymmetric 40% noise rates on CIFAR-100 dataset. We only use the selected clean samples to train the network, and report the F1 score (the first row) and the corresponding test accuracy (the second row) during training. The results are shown in Fig. 4. It can be seen from the first row that CIR can select clean samples from noisy ones accurately irrespective of the noise level. It is worth noting that in asymmetric noise case, the F1 score 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.

Methods/Noise	Sym-20%	Sym-50%	Asym-20%	Asym-30%
(1). Baseline	93.5	90.2	93.2	91.8
(2). (1)+SSL	95.4	94.2	93.9	92.1
(3). (2)+AT	95.5	94.4	94.3	93.8
(4). (3)+NCM	**95.6**	**95.1**	**94.7**	**94.0**

Table 3: Ablation study of CIR on CIFAR-10 with symmetric 20% and 50%, asymmetric 20% and 30% noise rates.

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.

Acknowledgments

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.

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

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.

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. *Master's thesis University of Toronto*.

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.

Peng, X.; Wang, K.; Zeng, Z.; Li, Q.; Yang, J.; and Qiao, Y. 2020. Suppressing mislabeled data via grouping and self-attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, 786–802. Springer.

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.

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.

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