Jo-SRC: A Contrastive Approach for Combating Noisy Labels

Yazhou Yao[1,*], Zeren Sun[1,*,†], Chuanyi Zhang[1], Fumin Shen[2], Qi Wu[3], Jian Zhang[4], Zhenmin Tang[1]
[1]Nanjing University of Science and Technology, Nanjing, China
[2]University of Electronic Science and Technology of China, Chengdu, China
[3]The University of Adelaide, Adelaide, Australia
[4]University of Technology Sydney, Sydney, Australia

Abstract

*Due to the memorization effect in Deep Neural Networks (DNNs), training with noisy labels usually results in inferior model performance. Existing state-of-the-art methods primarily adopt a sample selection strategy, which selects small-loss samples for subsequent training. However, prior literature tends to perform sample selection within each mini-batch, neglecting the imbalance of noise ratios in different mini-batches. Moreover, valuable knowledge within high-loss samples is wasted. To this end, we propose a noise-robust approach named Jo-SRC (**J**oint Sample **S**election and Model **R**egularization based on **C**onsistency). Specifically, we train the network in a contrastive learning manner. Predictions from two different views of each sample are used to estimate its "likelihood" of being clean or out-of-distribution. Furthermore, we propose a joint loss to advance the model generalization performance by introducing consistency regularization. Extensive experiments have validated the superiority of our approach over existing state-of-the-art methods. The source code and models have been made available at* https://github.com/NUST-Machine-Intelligence-Laboratory/Jo-SRC.

1. Introduction

DNNs have recently lead to tremendous progress in various computer vision tasks [14, 28, 25, 40, 21]. These successes largely attribute to large-scale datasets with reliable annotations (*e.g*., ImageNet [4]). However, collecting well-annotated datasets is extremely labor-intensive and time-consuming, especially in domains where expert knowledge is required (*e.g*., fine-grained categorization [37, 36]). The high cost of acquiring large-scale well-labeled data poses a bottleneck in employing DNNs in real-world scenarios.

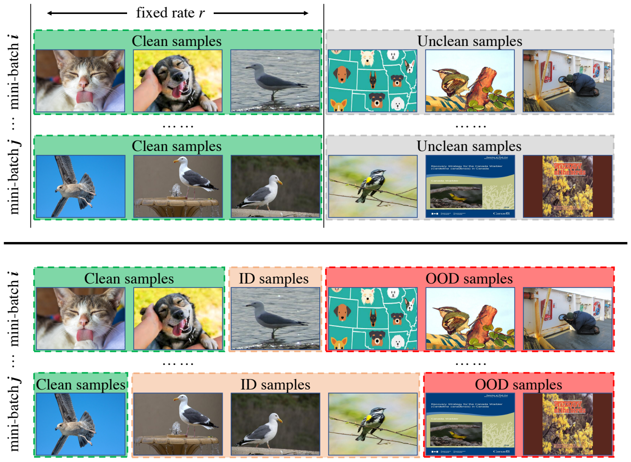


Figure 1. Existing small-loss based sample selection methods (**upper**) tend to regard a human-defined proportion of samples within each mini-batch as clean ones. They ignore the fluctuation of noise ratios in different mini-batches. On the contrary, our proposed method (**bottom**) selects clean samples in a global manner. Moreover, in-distribution (ID) noisy samples and out-of-distribution (OOD) ones are also selected and leveraged for enhancing the model generalization performance.

As an alternative, employing web images to train DNNs has received increasing attention recently [20, 41, 42, 34, 45, 44, 51, 52, 32]. Unfortunately, whereas web images are cheaper and easier to obtain via image search engines [5, 29, 46, 43], they usually yield inevitable noisy labels due to the error-prone automatic tagging system or non-expert annotations [23, 32, 45, 47]. A recent study has suggested that samples with noisy labels would be unavoidably overfitted by DNNs and consequently cause performance degradation [15, 50].

To alleviate this issue, many methods have been proposed for learning with noisy labels. Early approaches primarily attempt to correct losses during training. Some methods correct losses by introducing a noise transition matrix [31, 24, 6, 11]. However, estimating the noise transition

[*] Equal contribution.
[†] Corresponding author.

matrix is challenging, requiring either prior knowledge or a subset of well-labeled data. Some methods design noise-robust loss functions which correct losses according to predictions of DNNs [26, 54, 34]. However, these methods are prone to fail when the noise ratio is high.

Another active research direction in mitigating the negative effect of noisy labels is training DNNs with selected or reweighted training samples [12, 27, 22, 8, 49, 38, 32]. The challenge is to design a proper criterion for identifying clean samples. It has been recently observed that DNNs have a memorization effect and tend to learn clean and simple patterns before overfitting noisy labels [15, 50]. Thus, state-of-the-art methods (*e.g.*, Co-teaching [49], Co-teaching+ [49], and JoCoR [38]) propose to select a human-defined proportion of small-loss samples as clean ones. Although promising performance gains have been witnessed by employing the small-loss sample selection strategy, these methods tend to assume that noise ratios are identical among all mini-batches. Hence, they perform sample selection within each mini-batch based on an estimated noise rate. However, this assumption may not hold true in real-world cases, and the noise rate is also challenging to estimate accurately (*e.g.*, Clothing1M [39]). Furthermore, existing literature mainly focuses on closed-set scenarios, in which only in-distribution (ID) noisy samples are considered. In open-set cases (*i.e.*, real-world cases), both in-distribution (ID) and out-of-distribution (OOD) noisy samples exist. High-loss samples do not necessarily have noisy labels. In fact, hard samples, ID noisy ones, and OOD noisy ones all produce large loss values, but the former two are potentially beneficial for making DNNs more robust [32].

Motivated by the self-supervised contrastive learning [3, 7], we propose a simple yet effective approach named Jo-SRC (**Jo**int Sample **S**election and Model **R**egularization based on **C**onsistency) to address aforementioned issues. Specifically, we first feed two different views of an image into a backbone network and predict two corresponding softmax probabilities accordingly. Then we divide samples based on two likelihood metrics. We measure the likelihood of a sample being clean using the Jensen-Shannon divergence between its predicted probability distribution and its label distribution. We measure the likelihood of a sample being OOD based on the prediction disagreement between its two views. Subsequently, clean samples are trained conventionally to fit their given labels. ID and OOD noisy samples are re-labeled by a mean-teacher model before they are back-propagated for updating network parameters. Finally, we propose a joint loss, including a classification term and a consistency regularization term, to further advance model performance. A comparison between Jo-SRC and existing sample selection methods is provided in Figure 1. The major contributions of this work are:

(1) We propose a simple yet effective contrastive approach named Jo-SRC to alleviate the negative effect of noisy labels. Jo-SRC trains the network with a joint loss, including a cross-entropy term and a consistency term, to obtain higher classification and generalization performance.

(2) Our proposed Jo-SRC selects clean samples globally by adopting the Jensen-Shannon divergence to measure the likelihood of each sample being clean. We also propose to distinguish ID noisy samples and OOD noisy ones based on the prediction consistency between samples' different views. ID and OOD noisy samples are relabeled by a mean-teacher network before being used for network update.

(3) By providing comprehensive experimental results, we show that Jo-SRC significantly outperforms state-of-the-art methods on both synthetic and real-world noisy datasets. Furthermore, extensive ablation studies are conducted to validate the effectiveness of our approach.

2. Related Works

Existing works on learning with noisy labels can be briefly categorized into the following two subsets [32]: 1) Loss Correction and 2) Sample Selection.

Loss correction. A large proportion of existing literature on training with noisy labels focuses on loss correction approaches. Some methods endeavor to estimate the noise transition matrix [31, 2, 24, 6, 11]. For example, Patrini *et al.* [24] provided a loss correction method to estimate the noise transition matrix by using a deep network trained on the noisy dataset. However, these methods are limited in that the noise transition matrix is challenging to estimate accurately and may not be feasible in real-world scenarios. Some methods attempt to design noise-tolerant loss functions [26, 54, 34]. For example, the bootstrapping loss [26] extended the conventional cross-entropy loss with a perceptual term. However, these methods fail to perform well in real-world cases when the noise ratio is high.

Sample Selection. Another idea of dealing with noisy labels is to select and remove corrupted data. The problem is to find proper sample selection criteria. It has been shown that DNNs tend to learn simple patterns first before memorizing noisy data [15, 50]. Resorting to this observation, the small-loss sample selection criterion has been widely adopted: samples with lower loss values are more likely to have clean labels. For example, Co-teaching [8] proposed to maintain two networks simultaneously during training, with one network learning from the other networks selected small-loss samples. JoCoR [38] proposed to use a joint loss, including the conventional cross-entropy loss and the co-regularization loss, to select small-loss samples. However, above methods select samples within each mini-batch based on a human-defined drop rate. In real-world scenarios, noise ratios in different mini-batches are not guaranteed to be identical, and the drop rate is challenging to estimate.

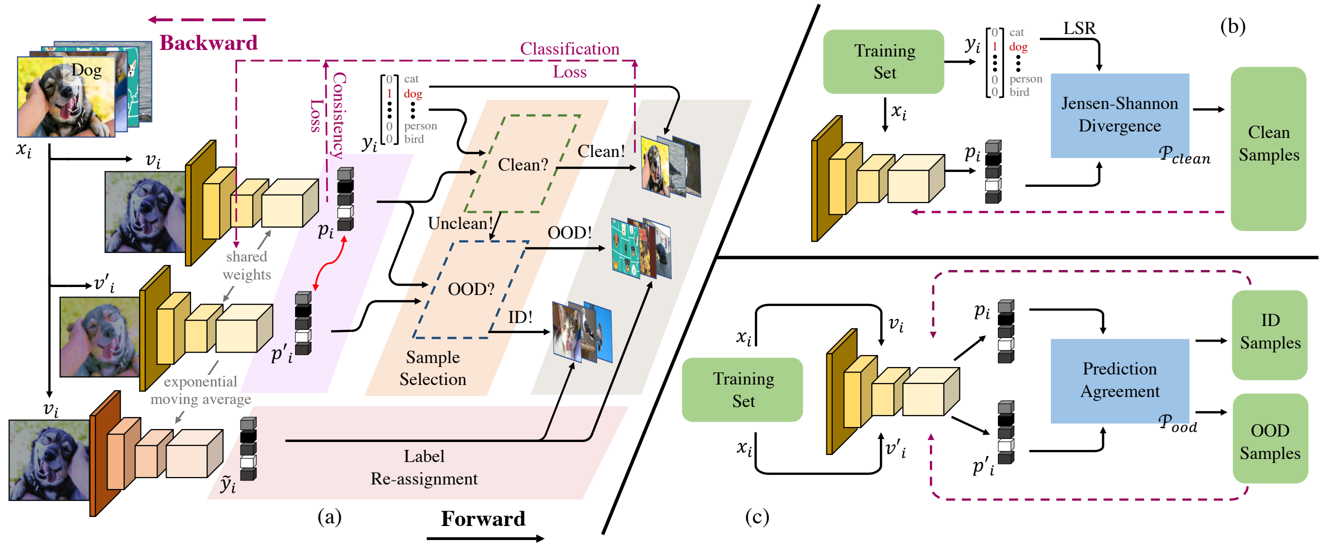


Figure 2. The overall framework of our proposed Jo-SRC approach (a), the clean sample selection module (b), and the ID/OOD sample selection module (c). Each image x_i is augmented into two different views v_i and v'_i before being fed into the backbone network. The network then predicts two probability distributions $\boldsymbol{p_i}$ and $\boldsymbol{p'_i}$ accordingly. Afterwards, we obtain the likelihood of x_i being clean $\mathcal{P}_{clean}$ using the Jensen-Shannon (JS) divergence between its predicted distribution $\boldsymbol{p_i}$ and its label distribution $\boldsymbol{y_i}$. If x_i is judged as "unclean", we obtain its likelihood of being out-of-distribution (OOD) $\mathcal{P}_{ood}$ based on the prediction disagreement between $\boldsymbol{p_i}$ and $\boldsymbol{p'_i}$. Finally, x_i is re-labeled as $\boldsymbol{\tilde{y}_i}$ by a mean-teacher model. The final objective function is a joint loss, including a classification term and a consistency term.

3. The Proposed Method

Background. Generally, for a multi-class classification task with C classes, we train DNNs using a labeled dataset $\mathbb{D} = \{(x_i, y_i)|1 \leq i \leq N\}$, in which x_i is the i-th training sample and $y_i \in \{0,1\}^C$ is its corresponding one-hot label over C classes. The conventional objective loss function is the cross-entropy between the predicted softmax probability distributions of training samples and their corresponding label distributions:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_i^c \log(p_i^c), \tag{1}$$

in which p_i^c is a simplified form of $p^c(x_i, \theta)$, denoting the predicted probability of sample x_i for class c given a model with parameters θ. However, for datasets with noisy labels (*e.g.*, web image datasets), labels are not guaranteed to be correct. Thus, training DNNs using noisy datasets directly is problematic and usually leads to a dramatic performance drop, given the fact that DNNs have the capability to memorize all training samples, including noisy ones [15].

Terminology. This paper adopts two consistency metrics to reveal how likely each sample could be clean or OOD. We accordingly term them as "likelihood", which is different from the concept of "likelihood" in statistics.

3.1. Global clean sample selection

Regarding samples with small cross-entropy losses as clean ones is one of the most widely-used sample selection criteria. This criterion is justified by the observation, in which DNNs tend to learn clean patterns first and then gradually fit noisy labels [15, 50]. Methods using this criterion (*e.g.*, Co-teaching [8] and Co-teaching+ [49]) typically select a pre-defined proportion of small-loss samples within each mini-batch. Unfortunately, noise ratios in different mini-batches inevitably fluctuate in real-world scenarios. One solution is to record losses for all samples and select samples in the entire training set. However, this becomes impractical when the dataset volume is increasingly huge.

To this end, we propose to reformulate the clean sample selection criterion from another perspective. Specifically, we propose to adopt the Jensen-Shannon (JS) divergence in Eq. (2) to quantify the difference d_i between the predicted probability distribution $\boldsymbol{p_i} = [p_i^1, p_i^2, ..., p_i^C]$ and the given ground truth label distribution $\boldsymbol{y_i} = [y_i^1, y_i^2, ..., y_i^C]$ of the sample x_i as follows:

$$\begin{aligned} d_i &= D_{JS}(\boldsymbol{p_i}\|\boldsymbol{y_i}) \\ &= \frac{1}{2}D_{KL}(\boldsymbol{p_i}\|\frac{\boldsymbol{p_i}+\boldsymbol{y_i}}{2}) + \frac{1}{2}D_{KL}(\boldsymbol{y_i}\|\frac{\boldsymbol{p_i}+\boldsymbol{y_i}}{2}), \end{aligned} \tag{2}$$

in which $D_{KL}(\cdot\|\cdot)$ is the Kullback-Leibler (KL) divergence function. The JS divergence is a measure of differences between two probability distributions. It is known to be bounded in $[0, 1]$, given a base 2 logarithm is used [19]. Therefore, intuitively, we can leverage d_i to measure the "likelihood" of x_i being clean as follows:

$$\mathcal{P}_{clean}(x_i) = 1 - d_i \in [0, 1]. \tag{3}$$

In fact, $\mathcal{P}_{clean}(x_i)$ reveals the consistency between $\boldsymbol{p_i}$ and $\boldsymbol{y_i}$. Here, we adopt smoothed label distributions [33] in calculating Eq. (2) to avoid the issue of $\log(0)$. We finally define our clean sample selection criterion as follows:

Criterion 3.1. The sample x is a clean one if its likelihood of being clean $\mathcal{P}_{clean}(x) > \tau_{clean}$.

Why can we select clean samples globally based on $\mathcal{P}_{clean}$? Similar to the cross-entropy, the JS divergence is a measurement depicting differences between two probability distributions. Since the $\boldsymbol{y_i}$ in Eq. (2) is not updated in the back-propagation process, the JS divergence between $\boldsymbol{p_i}$ and $\boldsymbol{y_i}$ is equivalent to the cross-entropy between them. Accordingly, our proposed Criterion 3.1 is consistent with the small-loss sample selection criterion. However, whereas the value of cross-entropy is not constrained, the JS divergence is bounded in $[0, 1]$, making it a natural global selection metric to describe how likely a sample could be clean. By directly modeling the likelihood of a sample being clean using Eq. (3), clean samples are selected more efficiently in a global manner, alleviating the issue caused by the imbalance of noise ratios within different mini-batches.

3.2. Out-of-distribution detection

Real-world scenarios contain both in-distribution (ID) noisy samples and out-of-distribution (OOD) ones. Despite their noisy labels, they can contribute to the model if their labels are re-assigned properly, especially for ID samples. Therefore, dropping all "unclean" samples directly is not data-efficient.

DNNs are usually uncertain about OOD samples when making predictions since their correct labels are outside the task scope. Conversely, while ID noisy samples have corrupted labels, they usually lead to consistent model predictions. Therefore, inspired by the self-supervised contrastive learning [3] and agreement maximization principle [30], we propose to use the prediction consistency to distinguish OOD and ID samples. Specifically, we first generate two augmented views $v_i = T(x_i)$ and $v'_i = T'(x_i)$ from a sample x_i by applying two different image transformations $T(\cdot)$ and $T'(\cdot)$. These two views are subsequently fed into a DNN to produce their corresponding predictions $\boldsymbol{p_i}$ and $\boldsymbol{p'_i}$, respectively. Finally, we adopt the consistency between these two predictions to determine if this sample is out-of-distribution or not. More explicitly, we define the "likelihood" of a sample being out-of-distribution (OOD) as:

$$\mathcal{P}_{ood}(x_i) = \min(1, |\underset{c}{\operatorname{argmax}} \boldsymbol{p_i} - \underset{c}{\operatorname{argmax}} \boldsymbol{p'_i}|). \tag{4}$$

Consequently, given $\tau_{ood} \in (0, 1)$, our OOD/ID sample selection criterion is defined as follows:

Criterion 3.2. Given a sample x that is selected as a "unclean" one by Criterion 3.1, it is judged as an OOD noisy one if $\mathcal{P}_{ood}(x_i) > \tau_{ood}$ (*i.e.*, its predictions of two differently augmented views disagree). If $\mathcal{P}_{ood}(x_i) \leq \tau_{ood}$ (*i.e.*, its predictions of two differently augmented views is consistent), it is deemed as an ID noisy sample.

3.3. Label re-assignment

The proposed Criterion 3.1 and 3.2 jointly divide training data into three subsets: a clean subset $\mathbb{S}_{clean}$, an ID subset $\mathbb{S}_{id}$, and an OOD subset $\mathbb{S}_{ood}$. To leverage all training data efficiently, we treat their labels differently before feeding them into the network.

For samples in $\mathbb{S}_{clean}$, we keep their labels unaltered. To enhance the generalization performance, we adopt the label smoothing regularization (LSR) [33] when calculating their losses. Therefore, the label distribution of a clean sample x_i is provided as Eq. (5), given its label $l_i \in \{1, 2, 3, ..., C\}$:

$$\tilde{y}_i^c = \begin{cases} 1 - \epsilon, & c = l_i \\ \frac{\epsilon}{C-1}, & c \neq l_i \end{cases}, \tag{5}$$

in which ϵ is a hyper-parameter controlling the smoothness of the label distribution.

For samples in ID subset $\mathbb{S}_{id}$, inspired by the mean-teacher model [35], we use the temporally averaged model (*i.e.* mean-teacher model) to generate reliable pseudo label distributions for providing supervision. Therefore, given an ID sample x_i, its pseudo label distribution is provided as:

$$\tilde{y}_i^c = p^c(x_i, \theta_{mt}), \tag{6}$$

where θ_{mt} denotes parameters of the mean-teacher model.

Finally, for samples in $\mathbb{S}_{ood}$, we also use the mean-teacher model to create their corresponding pseudo label distributions. However, since OOD samples' true labels are outside the task scope, the DNN should be highly confused when predicting their label assignments. Therefore, we propose to enforce predictions of OOD samples to fit an approximately uniform distribution for boosting generalization performance. In practice, given an OOD sample x_i, we relabel it with the following pseudo label distribution:

$$\tilde{y}_i^c = \frac{e^{p^c(x_i, \theta_{mt})/s}}{\sum_{j=1}^{C} e^{p^j(x_i, \theta_{mt})/s}}, \tag{7}$$

in which s is a large scaling constant. In our experiments, we empirically, set $s = 10$ to make this label distribution smooth enough (*i.e.*, $\forall c \in \{1, 2, 3, ..., C\}, \tilde{y}_i^c \approx 1/C$).

It should be noted that the mean-teacher model is not updated via the loss back-propagation. Instead, its parameters θ_{mt} is an exponential moving average of θ. Specifically, given a decay rate $\omega \in [0, 1]$, θ_{mt} is updated in each training step as follows:

$$\theta_{mt} \leftarrow \omega\theta_{mt} + (1 - \omega)\theta. \tag{8}$$

Algorithm 1: Jo-SRC

```
Input: Network $\theta$, mean-teacher $\theta_{mt}$, learning rate $\eta$,
       iteration $I_{\max}$, epoch $t_w$ and $t_{\max}$.
for $t = 1, 2, ..., t_{\max}$ do
    for $iter = 1, 2, 3, ..., I_{\max}$ do
        Sample a mini-batch $\mathbb{B}$ randomly.
        Predict $\boldsymbol{p}(x, \theta)$ and $\boldsymbol{p'}(x, \theta)$.
        Divide samples into $\mathbb{B}_{clean}$, $\mathbb{B}_{id}$, and $\mathbb{B}_{ood}$ based
          on Criterion 3.1 and 3.2.
        Re-label samples by Eq. (5), (6), and (7).
        if $t_w \leq t \leq t_{\max}$ then
            Obtain $\mathcal{L}$ using entire $\mathbb{B}$ by Eq. (10).
            Update $\theta \leftarrow \theta - \eta \nabla \mathcal{L}$.
        else
            Obtain $\mathcal{L}_c$ using only $\mathbb{B}_{clean}$ by Eq. (11).
            Update $\theta \leftarrow \theta - \eta \nabla \mathcal{L}_c$.
        end
        Update $\theta_{mt}$ by Eq. (8).
    end
end
Output: Updated network $\theta$.
```

3.4. Consistency regularization

As stated above, we use each sample's prediction consistency to measure its likelihood of being OOD. We follow the intuition that in-distribution samples (including clean ones and noisy ones) tend to produce consistent predictions while out-of-distribution samples do not. Thus, we propose to use an auxiliary consistency loss as Eq. (9) to provide joint supervision for enhancing the separability between ID and OOD samples.

$$\mathcal{L}_o = \frac{1}{N} \sum_{i=1}^{N} \rho_i (D_{KL}(\boldsymbol{p_i} \| \boldsymbol{p'_i}) + D_{KL}(\boldsymbol{p'_i} \| \boldsymbol{p_i})), \tag{9}$$

in which $\rho_i = 1$ if $x_i \in \mathbb{S}_{clean} \cup \mathbb{S}_{id}$; otherwise, $\rho_i = -1$.

On the one hand, resorting to this additional regularization term, clean samples and ID ones are encouraged to make consistent predictions. Meanwhile, this consistency term also enhances the prediction divergence of OOD noisy samples. Our approach is accordingly able to select clean/ID/OOD samples more effectively. On the other hand, this auxiliary consistency loss also implicitly promotes representation learning in a self-supervised fashion.

3.5. The overall framework

Combining all submodules together, our final objective loss function is

$$\mathcal{L} = (1 - \alpha)\mathcal{L}_c + \alpha \mathcal{L}_o, \tag{10}$$

in which α is a hyper-parameter, and

$$\mathcal{L}_c = \frac{1}{N} \sum_{i=1}^{N} (- \sum_{c=1}^{C} \tilde{y}_i^c \log(p_i^c) - \sum_{c=1}^{C} \tilde{y}_i^c \log(p_i'^c)). \tag{11}$$

Details of Jo-SRC are shown in Figure 2 and Algorithm 1.

In practice, the model gets increasingly stronger during training and will eventually overfit noisy labels. Thus, we proposed to dynamically adjust the selection threshold τ_{clean} as Eq. (12):

$$\tau_{clean} = \begin{cases} \frac{t}{t_w} \tau_c, & 1 \leq t \leq t_w \\ \frac{(t - t_w)\Delta\tau}{t_{\max} - t_w} + \tau_c, & t_w < t \leq t_{\max} \end{cases}, \tag{12}$$

in which $\Delta\tau = \tau_m - \tau_c$. τ_c is a hyper-parameter and τ_m is a large constant (τ_m is empirically set to 0.95 in our experiments). Accordingly, more samples will be treated as clean ones in initial epochs so that the model can learn simple and easy patterns from as much samples as possible. As the training proceeds, fewer samples are fed into the model as clean ones for ensuring the quality of learned data.

4. Experiments

4.1. Experiment setup

Datasets. We evaluate Jo-SRC in four benchmark datasets: CIFAR100N-C, CIFAR80N-O, Clothing1M [39], and Food101N [16]. CIFAR100N-C and CIFAR80N-O are two synthetic datasets created from CIFAR100 [13]. Specifically, we follow JoCoR [38] to create the closed-set synthetic dataset CIFAR100N-C with a noise ratio $\mathfrak{n}_c \in (0, 1)$. The noise type $\mathfrak{T}$ could be either "Symmetry" or "Asymmetry". To create the open-set synthetic dataset CIFAR80N-O, we first regard the last 20 categories in CIFAR100 as out-of-distribution ones. Then we create in-distribution noisy samples by randomly corrupting $\mathfrak{n}_c$ percentage of remaining samples' labels in a $\mathfrak{T}$ fashion. This finally leads to an overall noise ratio $\mathfrak{n}_{all} = 0.2 + 0.8\mathfrak{n}_c$. Clothing1M and Food101N are two large-scale real-world datasets with noisy labels. Details are in supplementary materials.

Evaluation Metrics. For evaluating the model classification performance, we take the test accuracy as the evaluation metric. Besides, we also adopt the label precision as the metric to evaluate our sample selection criteria.

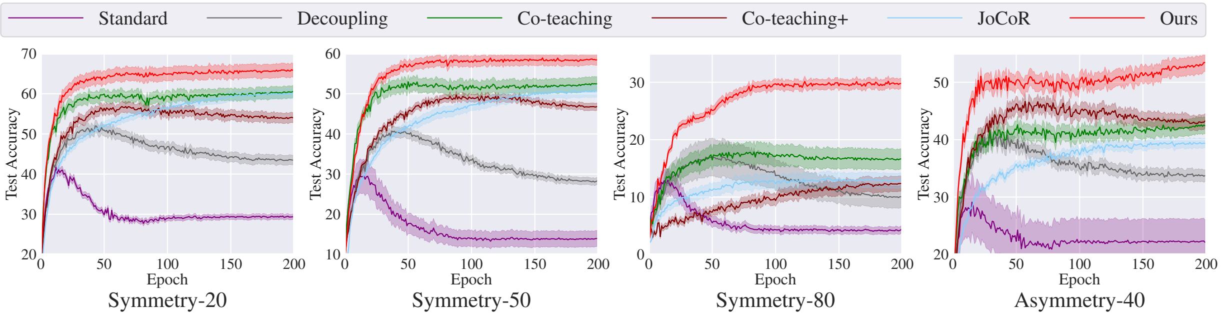
Implementation Details. Following JoCoR [38], we adopt a 7-layer DNN for CIFAR100N-C and CIFAR80N-O. During training, we use Adam optimizer with a momentum of 0.9. The initial learning rate is 0.001, and the batch size is 128. We train the network for 200 epochs and start to decay the learning rate linearly after 80 epochs. The decay rate in updating the mean-teacher network is set to $\omega = 0.99$. The τ_m and the LSR parameter ϵ is empirically set to 0.95 and 0.6, respectively. For Clothing1M, we follow settings in JoCoR [38] and use ResNet-18 [10] with ImageNet pre-trained weights to take a fair comparison with results presented in JoCoR. We also conduct experiments using ResNet-50 [10] and follow experimental settings used in DivideMix [17] for fair comparison. For Food101N, we

$\mathfrak{T} - \mathfrak{n}_c$	Standard	Decoupling	Co-teaching	Co-teaching+	JoCoR	Jo-SRC
Symmetry – 20%	35.14 ± 0.44	33.10 ± 0.12	43.73 ± 0.16	49.27 ± 0.03	53.01 ± 0.04	**58.15** ± 0.14
Symmetry – 50%	16.97 ± 0.40	15.25 ± 0.20	34.96 ± 0.50	40.04 ± 0.70	43.49 ± 0.46	**51.26** ± 0.11
Symmetry – 80%	4.41 ± 0.14	3.89 ± 0.16	15.15 ± 0.46	13.44 ± 0.37	15.49 ± 0.98	**23.80** ± 0.05
Asymmetry – 40%	27.29 ± 0.25	26.11 ± 0.39	28.35 ± 0.25	33.62 ± 0.39	32.70 ± 0.35	**38.52** ± 0.20

Table 1. Average test accuracy (%) on CIFAR100N-C over the last 10 epochs.

$\mathfrak{T} - \mathfrak{n}_c$	Standard	Decoupling	Co-teaching	Co-teaching+	JoCoR	Jo-SRC
Symmetry – 20%	29.37 ± 0.09	43.49 ± 0.39	60.38 ± 0.22	53.97 ± 0.26	59.99 ± 0.13	**65.83** ± 0.13
Symmetry – 50%	13.87 ± 0.08	28.22 ± 0.19	52.42 ± 0.51	46.75 ± 0.14	50.61 ± 0.12	**58.51** ± 0.08
Symmetry – 80%	4.20 ± 0.07	10.01 ± 0.29	16.59 ± 0.27	12.29 ± 0.09	12.85 ± 0.05	**29.76** ± 0.09
Asymmetry – 40%	22.25 ± 0.08	33.74 ± 0.26	42.42 ± 0.30	43.01 ± 0.59	39.37 ± 0.16	**53.03** ± 0.25

Table 2. Average test accuracy (%) on CIFAR80N-O over the last 10 epochs.



Symmetry-20 Symmetry-50 Symmetry-80 Asymmetry-40

Figure 3. Comparison on CIFAR80N-O: test accuracy (%) *vs.* epochs.

use ResNet-50 [10] pre-trained on ImageNet and follow experimental settings used in DeepSelf [9]. All experiments are repeated five times and averaged results are reported accordingly. Our code implementation is based on PyTorch.

Baselines. To evaluate Jo-SRC on CIFAR100N-C and CIFAR80N-O, we follow JoCoR [38] and compare Jo-SRC with the following state-of-the-art sample selection methods: Decoupling [22], Co-teaching [8], Co-teaching+ [49], and JoCoR [38]. To evaluate our approach on Clothing1M, besides the above methods, other state-of-the-art methods like F-correction [24], M-correction [1], Joint-Optim [34], Meta-Cleaner [53], Meta-Learning [18], P-correction [48], and DivideMix [17] are also compared. To perform evaluation on Food101N, CleanNet [16] and DeepSelf [9] are compared with our approach. Finally, training directly on noisy datasets is also adopted into comparison as a simple baseline (named as Standard).

4.2. Comparison on synthetic noisy datasets

Results on CIFAR100N-C. Whereas our proposed Jo-SRC method is designed for open-set scenarios, it is also applicable and useful in closed-set cases. Comparison in test accuracy with state-of-the-art approaches on CIFAR100N-C is shown in Table 1. For simplicity, the results of existing methods are drawn directly from JoCoR [38], and our method is evaluated using the same experimental settings. From Table 1, we can observe that our proposed Jo-SRC method consistently outperforms state-of-the-art methods. Although performance of all methods drops dramatically in the most inferior case (*i.e.*, Symmetry-80%), our methods still obtain the highest test accuracy.

Results on CIFAR80N-O. CIFAR80N-O is created to simulate the real-world scenario (*i.e.*, open-set problem). We present the comparison in test accuracy with state-of-the-art methods on CIFAR80N-O in Table 2. We implement all these methods with default parameters. Results in Table 2 come from experiments under the same experiment settings. From this table, we can observe that our Jo-SRC method performs consistently better than other methods. In the simplest case (*i.e.*, Symmetry-20%), while all methods work effectively and robustly (except Standard), our method achieves the best test accuracy. When the noise scenario becomes harder (*i.e.*, Symmetry-50%, and Asymmetry-40%), model performance inevitably starts to drop, especially Decoupling. However, our method is still effective and outperforms other methods. Finally, when it goes to the most challenging case (*i.e.*, Symmetry-80%), all approaches fail to combat the massive noisy labels. However, Jo-SRC once again achieves significantly higher performance than other methods, demonstrating the superiority of our method in

Method	Backbone	Test accuracy
Stardard	ResNet-18	67.22
Decoupling [22]	ResNet-18	68.48
Co-teaching [8]	ResNet-18	69.21
Co-teaching+ [49]	ResNet-18	59.32
JoCoR [38]	ResNet-18	70.30
Stardard	ResNet-50	69.21
F-correction [24]	ResNet-50	69.84
M-correction [1]	ResNet-50	71.00
Joint-Optim [34]	ResNet-50	72.16
Meta-Cleaner [53]	ResNet-50	72.50
Meta-Learning [18]	ResNet-50	73.47
P-correction [48]	ResNet-50	73.49
DivideMix [17]	ResNet-50	74.76
Jo-SRC	ResNet-18	**71.78**
Jo-SRC	ResNet-50	**75.93**

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

Method	Backbone	Test accuracy
Stardard	ResNet-50	84.51
CleanNet ω_{hard} [16]	ResNet-50	83.47
CleanNet ω_{soft} [16]	ResNet-50	83.95
DeepSelf [9]	ResNet-50	85.11
Jo-SRC	ResNet-50	**86.66**

Table 4. Comparison with state-of-the-art methods in test accuracy (%) on Food101N using ResNet-50.

coping with extremely noisy scenarios. Figure 3 shows the test accuracy *vs*. epochs. From this figure, we can observe that Jo-SRC consistently outperforms other methods by a large margin. Moreover, the superiority in the robustness of our method is demonstrated clearly in these curves.

4.3. Comparison on real-world noisy datasets

Results on Clothing1M. To verify the effectiveness of our Jo-SRC, we provide experimental results on real-world scenarios. Clothing1M is a large-scale real-world dataset. It contains one million training images and yield a 61.54% accuracy of noisy labels [39]. Table 3 shows comparison with state-of-the-art methods using ResNet-18 and ResNet-50 as the backbone network. From this table, we can observe that our proposed Jo-SRC approach achieves the best scores on both backbones. Using ResNet-18 as the backbone, our method achieves an improvement of 1.48% over the existing state-of-the-art. When ResNet-50 is adopted, Jo-SRC boosts the test accuracy from 74.76% to 75.93%.

Results on Food101N. Food101N is another real-world noisy dataset. It contains 310k training images in 101 food categories and also has a large proportion of noisy labels.

Noise	ID Sample		OOD Sample	
	best	*last*	*best*	*last*
Symmetry – 20%	60.91	42.62	59.54	54.38
Symmetry – 50%	83.86	65.92	40.70	38.75
Symmetry – 80%	96.31	72.84	26.67	24.60
Asymmetry – 40%	45.86	45.52	63.97	45.37

Table 5. The precision of ID/OOD sample selection on CIFAR80N-O at the *best* and *last* epoch.

Model	Test accuracy
Standard	29.37 ± 0.09
Jo-SRC-C	57.12 ± 0.33
Jo-SRC-CI	61.32 ± 0.18
Jo-SRC-CIO	63.10 ± 0.07
Jo-SRC	65.83 ± 0.13

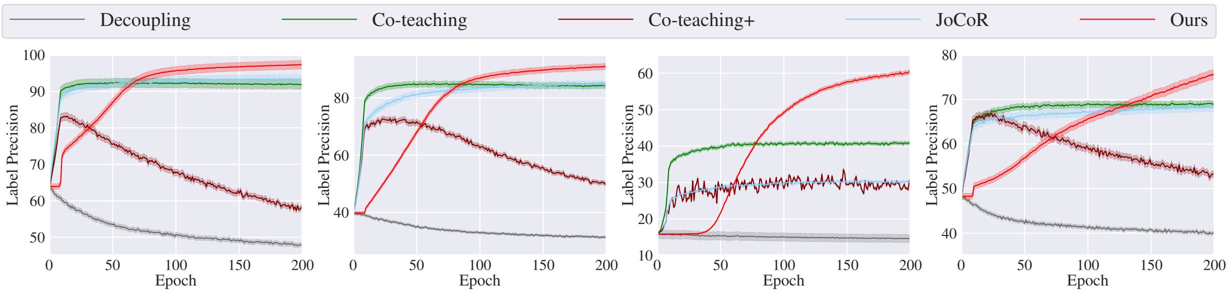
Table 6. Effect of different steps in test accuracy (%) on CIFAR80N-O (Symmetry-20%) over the last 10 epochs.

Table 4 presents the performance comparison with state-of-the-arts. As shown in Table 4, Jo-SRC achieves the best score and outperforms the state-of-the-art DeepSelf [9] by 1.55%, validating the effectiveness of our approach in dealing with real-world noisy cases.

4.4. Ablation Study

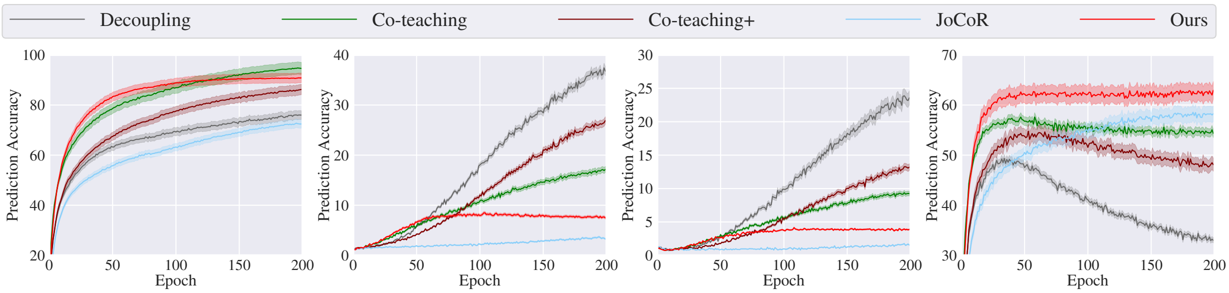
Precision of sample selection. The key reason for our approach in obtaining state-of-the-art performance is accurate and reliable sample selection. To study and verify the superiority of our proposed sample selection strategy, we show the precision of sample selection in Figure 4 and Table 5. Figure 4 presents the precision of clean sample selection *vs*. epochs. From this figure, Jo-SRC is shown to be highly effective in selecting clean samples accurately and reliably. In all cases, our proposed Jo-SRC achieves the best performance in selecting clean data, compared with state-of-the-art sample selection methods. Furthermore, in the most demanding scenario (*i.e*., Symmetry-80%), while all other methods suffer in finding clean samples, the selection precision of our Jo-SRC increases steadily as the training proceeds. These results validate the effectiveness of our clean sample selection strategy. Table 5 presents the precision in selecting ID/OOD samples. In this table, the *best* and *last* denote the selection precision at the best and last epochs, respectively. Results shown in this table verify the effectiveness of our Jo-SRC in selecting ID/OOD samples.

Prediction accuracy of different training samples. The memorization effect argues that DNNs would eventually memorize all samples (including noisy ones). Therefore, it is critical to prevent networks from overfitting noisy labels when training with noisy datasets. To further prove the effectiveness of our proposed Jo-SRC, we show the pre-



Symmetry-20 Symmetry-50 Symmetry-80 Asymmetry-40

Figure 4. Comparison on CIFAR80N-O: precision of clean sample selection (%) *vs.* epochs.



Clean Samples *w.r.t.* Given Labels

OOD Noisy Samples *w.r.t.* Given Labels

ID Noisy Samples *w.r.t.* Given Labels

ID Noisy Samples *w.r.t.* **True** Labels

Figure 5. The prediction accuracy (%) on different groups of CIFAR80N-O (Symmetry-20%) training data during the training process.

diction accuracy of different training samples in Figure 5. As shown in this figure, all methods achieve increasing prediction accuracy on clean samples. JoCoR and our Jo-SRC achieve the lowest prediction accuracy on noisy samples (including both ID ones and OOD ones) w.r.t. given labels (*i.e.*, noisy labels). This indicates that JoCoR and Jo-SRC perform best in preventing networks from memorizing noisy labels. Although JoCoR obtains lower prediction accuracy on ID and OOD training samples, it yields an underfitting issue in clean samples, leading to sub-optimal final test accuracy. While Co-teaching fits clean samples slightly better than our Jo-SRC, it suffers from overfitting on noisy labels. This causes its final performance decrease in test samples. Moreover, by observing the last sub-figure, we can find that our Jo-SRC achieves the best prediction accuracy on ID noisy samples w.r.t. their true labels. This further demonstrates the effectiveness of our sample selection and model regularization, given the fact that ID noisy samples are not supervised by their true labels during training.

Influence of different steps. Table 6 reveals the effect of different steps in our method. The Jo-SRC-C denotes the case in which only selected clean samples are adopted in training. The Jo-SRC-CI denotes the case where clean samples and ID noisy samples are adopted in training. The Jo-SRC-CIO denotes the case when all samples are adopted in training. The mean-teacher-based re-labeling is performed accordingly when noisy samples are leveraged in training. Lastly, the Jo-SRC denotes the final proposed method. From this table, we can observe that the proposed clean sample selection plays the most crucial role in addressing the label noise issue. Moreover, appropriately treated noisy samples (including ID and OOD ones) can contribute to the model generalization performance. Finally, the consistency loss promotes model performance by further regularization.

5. Conclusion

In this paper, we proposed a simple yet effective approach named Jo-SRC to address the performance degradation caused by noisy labels. Jo-SRC trained DNNs in a contrastive manner. Clean samples were identified globally based on JS divergence, while ID and OOD noisy samples were distinguished based on consistency. Samples were selected and divided accordingly for subsequent network learning. Finally, a joint loss, including a classification term and a consistency regularization term, was proposed to further advance the performance and robustness. Comprehensive experiments on both synthetic and real-world noisy datasets validated the superiority of the proposed method.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61976116) and Fundamental Research Funds for the Central Universities (No. 30920021135).

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *ICML*, pages 312–321, 2019. 6, 7

[2] Haw-Shiuan Chang, Erik Learned-Miller, and Andrew McCallum. Active bias: Training more accurate neural networks by emphasizing high variance samples. In *NeurIPS*, pages 1002–1012, 2017. 2

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 2, 4

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 1

[5] Rob Fergus, Li Fei-Fei, Pietro Perona, and Andrew Zisserman. Learning object categories from internet image searches. *Proceedings of the IEEE*, 98(8):1453–1466, 2010. 1

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

[7] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 2

[8] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. pages 8527–8537, 2018. 2, 3, 6, 7

[9] Jiangfan Han, Ping Luo, and Xiaogang Wang. Deep self-learning from noisy labels. In *ICCV*, pages 5138–5147, 2019. 6, 7

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5, 6

[11] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. In *NeurIPS*, pages 10456–10465, 2018. 1, 2

[12] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*, pages 2309–2318, 2017. 2

[13] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Technical report, University of Toronto*, 1(4):7, 2009. 5

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1097–1105, 2012. 1

[15] David Krueger, Nicolas Ballas, Stanislaw Jastrzebski, Devansh Arpit, Maxinder S Kanwal, Tegan Maharaj, Emmanuel Bengio, Asja Fischer, Aaron Courville, Simon Lacoste-Julien, et al. A closer look at memorization in deep networks. In *ICML*, 2017. 1, 2, 3

[16] Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *CVPR*, pages 5447–5456, 2018. 5, 6, 7

[17] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *ICLR*, 2020. 5, 6, 7

[18] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Learning to learn from noisy labeled data. In *CVPR*, pages 5051–5059, 2019. 6, 7

[19] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991. 3

[20] Huafeng Liu, Chuanyi Zhang, Yazhou Yao, Xiushen Wei, Fumin Shen, Jian Zhang, and Zhenmin Tang. Exploiting web images for fine-grained visual recognition by eliminating open-set noise and utilizing hard examples. *IEEE TMM*, 2021. 1

[21] Haonan Luo, Guosheng Lin, Zichuan Liu, Fayao Liu, Zhenmin Tang, and Yazhou Yao. Segeqa: Video segmentation based visual attention for embodied question answering. In *ICCV*, pages 9667–9676, 2019. 1

[22] Eran Malach and Shai Shalev-Shwartz. Decoupling "when to update" from "how to update". pages 960–970, 2017. 2, 6, 7

[23] Li Niu, Ashok Veeraraghavan, and Ashutosh Sabharwal. Webly supervised learning meets zero-shot learning: A hybrid approach for fine-grained classification. In *CVPR*, pages 7171–7180, 2018. 1

[24] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*, pages 1944–1952, 2017. 1, 2, 6, 7

[25] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, pages 7263–7271, 2017. 1

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

[27] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *ICML*, pages 4331–4340, 2018. 2

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015. 1

[29] Florian Schroff, Antonio Criminisi, and Andrew Zisserman. Harvesting image databases from the web. *IEEE TPAMI*, 33(4):754–766, 2010. 1

[30] Vikas Sindhwani, Partha Niyogi, and Mikhail Belkin. A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of ICML workshop on learning with multiple views*, volume 2005, pages 74–79. Citeseer, 2005. 4

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

[32] Zeren Sun, Xian-Sheng Hua, Yazhou Yao, Xiu-Shen Wei, Guosheng Hu, and Jian Zhang. Crssc: salvage reusable samples from noisy data for robust learning. In *ACM MM*, pages 92–101, 2020. 1, 2

[33] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016. 4

[34] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *CVPR*, pages 5552–5560, 2018. 1, 2, 6, 7

[35] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, pages 1195–1204, 2017. 4

[36] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, pages 8769–8778, 2018. 1

[37] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *CNS-TR-2011-001*, 2011. 1

[38] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *CVPR*, pages 13726–13735, 2020. 2, 5, 6, 7

[39] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *CVPR*, pages 2691–2699, 2015. 2, 5, 7

[40] Guo-Sen Xie, Li Liu, Fan Zhu, Fang Zhao, Zheng Zhang, Yazhou Yao, Jie Qin, and Ling Shao. Region graph embedding network for zero-shot learning. In *ECCV*, pages 562–580. Springer, 2020. 1

[41] Jufeng Yang, Xiaoxiao Sun, Yu-Kun Lai, Liang Zheng, and Ming-Ming Cheng. Recognition from web data: A progressive filtering approach. *IEEE TIP*, 27(11):5303–5315, 2018. 1

[42] Yazhou Yao, Xiansheng Hua, Guanyu Gao, Zeren Sun, Zhibin Li, and Jian Zhang. Bridging the web data and fine-grained visual recognition via alleviating label noise and domain mismatch. In *ACM MM*, pages 1735–1744, 2020. 1

[43] Yazhou Yao, Xian-sheng Hua, Fumin Shen, Jian Zhang, and Zhenmin Tang. A domain robust approach for image dataset construction. In *ACM MM*, pages 212–216, 2016. 1

[44] Yazhou Yao, Fumin Shen, Guosen Xie, Li Liu, Fan Zhu, Jian Zhang, and Heng Tao Shen. Exploiting web images for multi-output classification: From category to subcategories. *IEEE TNNLS*, (7):2348–2360, 2020. 1

[45] Yazhou Yao, Fumin Shen, Jian Zhang, Li Liu, Zhenmin Tang, and Ling Shao. Extracting multiple visual senses for web learning. *IEEE TMM*, 21(1):184–196, 2019. 1

[46] Yazhou Yao, Jian Zhang, Fumin Shen, Xiansheng Hua, Jingsong Xu, and Zhenmin Tang. Exploiting web images for dataset construction: A domain robust approach. *IEEE TMM*, 19(8):1771–1784, 2017. 1

[47] Yazhou Yao, Jian Zhang, Fumin Shen, Li Liu, Fan Zhu, Dongxiang Zhang, and Heng Tao Shen. Towards automatic construction of diverse, high-quality image datasets. *IEEE TKDE*, 32(6):1199–1211, 2020. 1

[48] Kun Yi and Jianxin Wu. Probabilistic end-to-end noise correction for learning with noisy labels. In *CVPR*, pages 7017–7025, 2019. 6, 7

[49] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *ICML*, pages 7164–7173, 2019. 2, 3, 6, 7

[50] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017. 1, 2, 3

[51] Chuanyi Zhang, Yazhou Yao, Huafeng Liu, Guo-Sen Xie, Xiangbo Shu, Tianfei Zhou, Zheng Zhang, Fumin Shen, and Zhenmin Tang. Web-supervised network with softly update-drop training for fine-grained visual classification. In *AAAI*, pages 12781–12788, 2020. 1

[52] Chuanyi Zhang, Yazhou Yao, Xiangbo Shu, Zechao Li, Zhenmin Tang, and Qi Wu. Data-driven meta-set based fine-grained visual recognition. In *ACM MM*, pages 2372–2381, 2020. 1

[53] Weihe Zhang, Yali Wang, and Yu Qiao. Metacleaner: Learning to hallucinate clean representations for noisy-labeled visual recognition. In *CVPR*, pages 7373–7382, 2019. 6, 7

[54] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*, pages 8778–8788, 2018. 2