DEEP DUPLEX LEARNING FOR WEAK SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Weak supervision widely exists in practice and shows various forms such as noisy labels, partial labels, or pseudo labels. As a weak supervisor might provide false training signals, most existing works focus on correcting the supervisor or ignoring certain constraints. While they tackle each type separately, we propose a deep duplex learning (DDL) method to deal with all kinds of weak supervision from a unified perspective of supervision utilization. We exploit both the supervision and counter-supervision signals for training and allow the network to implicitly and adaptively balance the two signals. We describe each image using a duplex representation composed of a superficial representation (SR) and a hypocritical representation (HR). We then impose the supervision signal and counter-supervision signal on SR and HR, respectively. The SR and HR collaborate to interact with the weak supervisor to adaptively confine the effect of false supervisions on the network. Our DDL sets new state-of-the-arts for noisy label learning, partial label learning, and semi-supervised learning on standard benchmarks. [1]

1 INTRODUCTION

The vast quantity of labeled data enables us to train high-performing deep models in various tasks, such as image classification (He et al., 2016a; Dosovitskiy et al., 2020), object detection (Carion et al., 2020; Zhu et al., 2020), and semantic segmentation (Li et al., 2017b; Strudel et al., 2021). With the development of computing hardware, we can scale deep models to an enormous size (Riquelme et al., 2021; Radford et al., 2021), which demands larger-scale data for training (Zhai et al., 2021). However, annotating clean labels is expensive and time-consuming, rendering the use of automated crawled noisy labels (Xiao et al., 2015; Li et al., 2017a; Radford et al., 2021), partial labels (Wang et al., 2022), and machine-annotated labels (Li et al., 2021) a more practical choice. Weakly supervised learning is thus considered a promising direction and attracts increasing attention, where researchers have delved into specific fields to tackle different types of weak supervision, including learning with noisy labels (Li et al., 2020; Tan et al., 2021), partial label learning (Feng et al., 2020b; Wang et al., 2022), and semi-supervised learning (Berthelot et al., 2019b; Li et al., 2021).

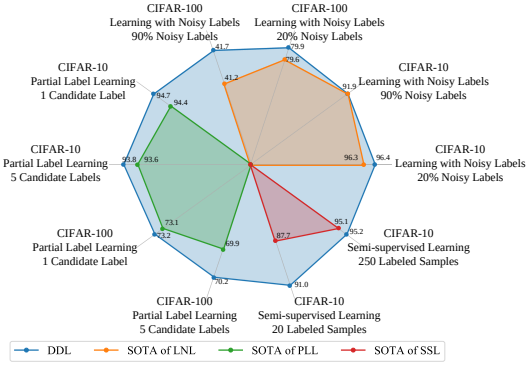

Figure 1: DDL improves the state of the art on three main types of weakly supervised learning.

As a weak supervisor may provide false information, most existing works focus on how to modulate the weak supervision to produce a more accurate training signal. For example, some works explore ways to identify false supervision (Han et al., 2018; Li et al., 2020; Yu et al., 2019; Wei et al., 2020). They use loss distribution to differentiate clean or false supervision with a small loss criterion (Han et al., 2018) or a gaussian mixture model criterion (Li et al., 2020). Other works focus on correcting the instructed relations provided by the weak supervisor (Patrini et al., 2017; Tanaka et al., 2018; Li et al., 2021; Wang et al., 2022). The state-of-the-art methods employ an exponential mean averaged model to generate more accurate labels for training (Li et al., 2021; Wang et al., 2022). However, they can usually deal with only one type of weak supervision. Also, it is difficult

[1] Code is provided in the supplementary material.

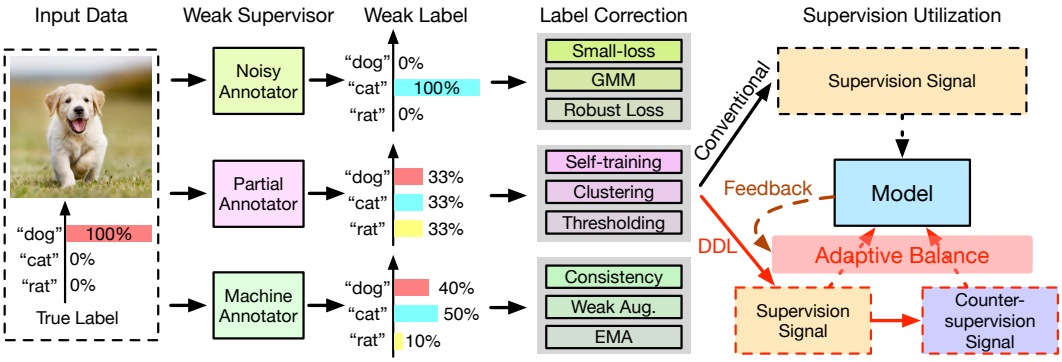

Figure 2: An overview of the proposed deep duplex learning for weak supervision. While existing methods mainly focus on developing different techniques to correct a specific type of weak supervision, we tackle weakly supervised learning from a unified perspective of label utilization. We allow the network itself to adaptively balance the effect of supervision and counter-supervision signals.

to fully identify all the false supervisions or provide completely authentic labels for training. The learned representation still suffers from inaccurate training signals, causing degraded performance.

To further confine the effect of false supervisions on the image representation, we propose a deep duplex learning (DDL) method to tack weakly supervised learning from a unified perspective of supervision utilization, as shown in Figure 2. We assume that a carefully-designed network can more easily learn from true supervisions, enabling the network itself to implicitly emphasize the beneficial signals. We employ a superficial representation (SR) and a hypocritical representation (HR) to represent each image and class prototype and compute a superficial similarity (SS) and a hypocritical similarity (HS) accordingly. We use the SR and the HR to obey and resist the supervision, respectively. Still, we constrain the overall effect on the SR and HR to be consistent with the provided supervision. To facilitate the adaptive balance of the two training signals, we further require the learning of SR and HR to be entangled with each other. The two representations collaborate with each other to adaptively learn from the weak supervisor and allow the network to implicitly identify the true supervisions. We further propose a simple duplex similarity function for efficient instantiation of deep duplex learning, which can be easily implemented and readily applied to existing methods. To demonstrate the effectiveness and generality of DDL, we conduct experiments on three types of weakly-supervised learning: learning with noisy labels (LNL), partial label learning (PLL), and semi-supervised learning (SSL). We apply DDL to the state-of-the-art method in the respective fields (DivideMix (Li et al., 2020) and AugDesc (Nishi et al., 2021) for LNL, PiCO (Wang et al., 2022) for PLL, and CoMatch (Li et al., 2021) for SSL). Our DDL shows consistent improvement and attains the best performance in all three tasks on various datasets, as shown in Figure 1.

2 Related Work

Learning with Noisy Labels. Web crawling (Xiao et al., 2015; Radford et al., 2021) and automatic annotation (Chen et al., 2017) facilitate the collection of large-scale labeled data for supervised training, but they inevitably introduce non-negligible noise to the labels. The ability to learn from noisy labels (LNL) thus becomes a valuable characteristic for a machine learning system. One category of works attempt to identify the noisy data and employs a different training strategy on them (Han et al., 2018; Li et al., 2020; Jiang et al., 2018; Yu et al., 2019; Liang et al., 2022). The widely used small-loss criterion (Han et al., 2018) deems samples with a high loss as noisy data. The Gaussian Mixture Model (GMM) criterion (Li et al., 2020) employs a GMM to identify noisy samples, assuming that the loss distributions of clean and noisy data are statistically separable. Other works seek to correct the loss imposed by the weak supervisor. Label correction methods employ a prediction network to infer the true labels to rectify noisy ones (Tanaka et al., 2018; Yi & Wu, 2019; Liu et al., 2020). Noise transition methods estimate a label transition matrix to infer the underlying true labels (Menon et al., 2015; Natarajan et al., 2013; Patrini et al., 2017; Xia et al., 2019). Furthermore, some methods designed more noise-robust loss functions, such as mean absolute error (MAE) (Ghosh et al., 2017), weighted MAE (Wang et al., 2019a), generalized cross-entropy (Zhang & Sabuncu, 2018), and symmetric cross-entropy (Wang et al., 2019b).

Partial Label Learning. Annotating each sample with an accurate label is highly expensive while associating it with multiple label candidates can be much cheaper, avoiding the cherry-picking between several ambiguous labels. Promoted by this, partial label learning (PLL) serves as a promising approach for a more efficient learning paradigm, where each sample is assigned a set of labels. A straightforward way is to treat all the assigned label candidates equally, which is susceptible to the misleading effects of false labels (Hüllermeier & Beringer, 2006; Cour et al., 2011; Zhang & Yu, 2015). This motivates subsequent methods to learn to identify the correct label from the label candidates. They perform the label disambiguation using various strategies, such as distance thresholding (Nguyen & Caruana, 2008; Wang et al., 2020), graph inference (Zhang et al., 2016; Xu et al., 2019; Wang et al., 2021), and feature clustering (Liu & Dietterich, 2012). The most successful way is through self-training, which employs the learning model itself to produce disambiguated labels (Feng et al., 2020b; Lv et al., 2020; Wen et al., 2021; Wang et al., 2022). For example, PiCO (Wang et al., 2022) employs an auxiliary contrastive objective with a momentum encoder to produce more accurate disambiguated labels.

Semi-supervised Learning. An evident way for efficient data collection is to only annotate a small portion of samples and leave the rest large amount of data unlabeled. Semi-supervised learning (SSL) is a long-standing problem in machine learning to target this situation, where the key is how to produce accurate and informative training signals from the vast unlabeled data. Pseudo-label-based methods use the prediction model being learned to generate a pseudo label for each unlabeled sample (Iscen et al., 2019; Berthelot et al., 2019b;a; Sohn et al., 2020). The working mechanism is the reciprocation of the model and the pseudo labels, where the quality of the generated pseudo labels improves as training and the prediction model can further benefit from more accurate pseudo labels (Lee, 2013; Arazo et al., 2020). Consistency-based methods require the different augmentations of the same sample to share similar representations, which produce more accurate yet less informative training signals for unlabeled data (Chen et al., 2020; Tang et al., 2021; Jeong et al., 2019; Tarvainen & Valpola, 2017). Recent methods (Li et al., 2021) combined the training signals from pseudo labels and representation consistency and have achieved improved performance. They generate pseudo labels based on weakly-augmented samples to improve accuracy and impose the representation consistency on strongly-augmented samples to improve informativeness.

A core issue for all LNL, PLL, and SSL is the misleading false supervisions. While existing methods proposed various ways to reduce the harmful effect of false supervisions, we do not explicitly differentiate between true and false supervision. Instead, we use both the supervision signals and the counter-supervision signals to train the model. We assume that a neural network with certain inductive biases can be more easily trained with the true supervision. We rely on the model itself to implicitly and adaptively decide the training direction.

3 Proposed Approach

In this section, we first present a general framework of weakly-supervised learning to cover learning with noisy labels, partial label learning, and semi-supervised learning. We then detail the proposed deep duplex learning (DDL) and present an efficient instantiation of DDL using the duplex similarity.

3.1 Problem Formulation

Given a set of N training samples $\{\mathbf{x}_1, ..., \mathbf{x}_N\}$, we define a supervisor a which assigns a label $l_i = a(\mathbf{x}) \in 1, 2, ..., L$ to each image $\mathbf{x}$. We assume that each image can be described with a single correct label and focus on image classification as it is the most basic task in computer vision. The generalization to learning with multiple labels or dense labels is beyond the scope of this paper.

Deep learning employs deep neural networks to obtain a vector $\mathbf{y}$ to represent each image $\mathbf{x}$. They usually employ a softmax classifier $\mathbf{c}$ to predict the probability that $\mathbf{y}$ belongs to the l-th class:

$$c(\mathbf{y}, l) = \frac{e^{\mathbf{w}_l \cdot \mathbf{y}}}{\sum_{i=1}^{L} e^{\mathbf{w}_i \cdot \mathbf{y}}}, \tag{1}$$

where $\mathbf{w}_i$ is a learnable vector with the same dimension of $\mathbf{y}$. We omit the bias terms for brevity.

Generally, the learning objective is to enlarge the predicted probability for the labeled class and reduce the probabilities for the other classes. The most commonly used loss function to train a

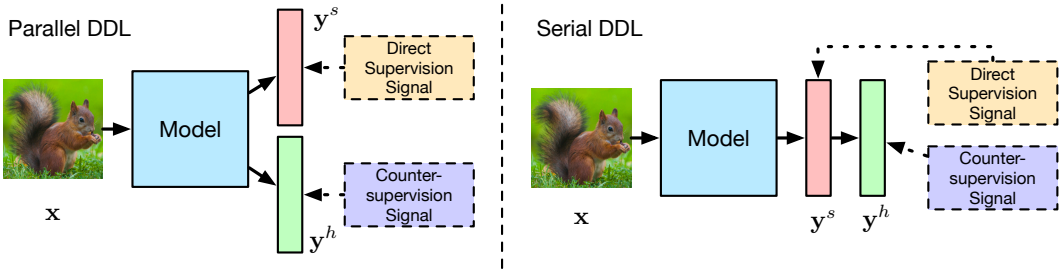

Figure 3: Illustration of Parallel DDL and Serial DDL. Parallel DDL obtains SR and HR using a parallel structure. Serial DDL first obtains SR and applies a transformation on SR to obtain HR.

network with the supervisor a can be formulated as:

$$L(\mathbf{y}, \mathbf{W}, a) = -\log(c(\mathbf{y}, a(\mathbf{y}))) = -\log(\frac{e^{\mathbf{w}_a(\mathbf{y})\cdot\mathbf{y}}}{\sum_{i=1}^{L} e^{\mathbf{w}_i\cdot\mathbf{y}}}), \tag{2}$$

where $\mathbf{w}_i$ is the i-th row vector of $\mathbf{W}$.

We generalize the learning objective Eq. (2) and formulate the learning process as the enforcement of certain relations between image representations and class prototypes. We employ a class prototype $\mathbf{p}_i$ to represent the i-th class and then compute a similarity score $s(\mathbf{y}, \mathbf{p})$ between an image representation $\mathbf{y}$ and a class prototype $\mathbf{p}$. The similarity score can be computed in a various ways, such as $s(\mathbf{y}, \mathbf{p}) = \mathbf{y} \cdot \mathbf{p}$ or $s(\mathbf{y}, \mathbf{p}) = \frac{\mathbf{y}\cdot\mathbf{p}}{||\mathbf{y}||_2||\mathbf{p}||_2}$. The enforced deviations are determined by the supervisor a, where only the similarities between images and their corresponding class prototypes increase and the other similarities decrease. The loss function can be then formulated as $L(s, a) = L(\mathbf{y}, \mathbf{W}, a)$.

The learning process of s with a loss function $L(s, a)$ using gradient descent can be formulated as:

$$s = s - \lambda\frac{\partial L(s, a)}{\partial s} = s + D(s, a), \tag{3}$$

where λ is the learning rate and $D(s, a) = -\lambda\frac{\partial L(s,a)}{\partial s}$ is the enforced deviation.

For a weak supervisor a^w, the enforced deviation $D(s, a^w)$ might be wrong and thus mislead the training process. Therefore, weakly-supervised learning attempts to correct the enforced deviation $D(s, a^w)$ with an amending factor $A(s)$ to obtain the amended relation $\hat{D}(s, a^w)$:

$$\hat{D}(s, a^w) = -\lambda\frac{\partial L(s, a)}{\partial s} \cdot A(s). \tag{4}$$

Note that $A(s)$ can be negative, indicating that the learning algorithm completely overturns the weak supervisor. For example, the LNL methods with the small-loss criterion (Han et al., 2018) employs $A(s) = I(T - L(s))$, where $I(x)$ denotes the indicator function which equals 1 for $x > 0$. and outputs 0 otherwise, and T is a pre-defined loss threshold. Other weakly-supervised learning methods design various $A(s)$ with a sharing goal to reduce the effect of false supervisions (Yi & Wu, 2019; Tanaka et al., 2018; Wang et al., 2019a; Zhang & Sabuncu, 2018).

3.2 Deep Duplex Learning

While most existing methods focus on seeking more accurate supervision signals for training, it is still impossible to completely eliminate the false supervisions. Therefore, we do not explicitly differentiate between true and false supervisions and propose to employ both the supervision and counter-supervision signals (the opposite of the supervision) to train the network, allowing the network itself to adaptively balance the effects of the two signals. We assume that a carefully designed deep neural network with certain inductive biases can more easily fit samples with true labels. So by training the network with both the supervision and counter-supervision signals, the network tends to emphasize the effect of true supervisions and undermine that of false supervisions.

A naive way is to impose both the loss and counter-loss on the image representation $\mathbf{y}$:

$$\hat{L}(\mathbf{y}, a) = L(\mathbf{y}, a) - \gamma L(\mathbf{y}, a) = (1 - \gamma)L(\mathbf{y}, a), \tag{5}$$

where $\gamma < 1$ controls the intensity of the counter-supervision signal. We see that the two signals neutralize each other, and Eq. (5) simply equals to using a smaller learning rate or $A(s) = 1 - \gamma$.

To address this, we propose to use a duplex representation $\mathbf{y}^d = \{\mathbf{y}^s, \mathbf{y}^h\}$ to describe each image, which is composed of a superficial representation (SR) $\mathbf{y}^s$ and a hypocritical representation (HR) $\mathbf{y}^h$. The two representations should be entangled to facilitate the automatic balance of the two supervision signals. We thus propose two ways to obtain them as shown in Figure 3:

- Parallel: $\mathbf{y}^s = g(f(\mathbf{x}))$ and $\mathbf{y}^h = h(f(\mathbf{x}))$.
- Serial: $\mathbf{y}^s = f(\mathbf{x})$ and $\mathbf{y}^h = h(f(\mathbf{x}))$.

We similarly use a duplex prototype $\mathbf{p} = \{\mathbf{p}^s, \mathbf{p}^h\}$ to represent each class. We compute a superficial similarity (SS) $s^s(\mathbf{y}^s, \mathbf{p}^s)$ and a hypocritical similarity (HS) $s^h(\mathbf{y}^h, \mathbf{p}^h)$ for each image-class pair. We impose the supervision signal on the SR $\mathbf{y}^s$ and the counter-supervision signal on the HR $\mathbf{y}^h$:

$$\hat{L}(\mathbf{y}^d, a) = L(\mathbf{y}^s, a) - \gamma L(\mathbf{y}^h, a), \tag{6}$$

which equals to $A^s(s^s) = 1$ and $A^h(s^h) = -\gamma$. The A^s and A^s are constants and are not aware of the current estimation of the similarity. However, we argue that the enforcements should be adapted to different similarities. A small HS indicates a negative judgment of the actual similarity between the image-class pair, so we should be cautious and update s_s in a smaller rate, i.e., $\frac{\partial |A^s|}{s^h} > 0$. On the other hand, a small SS indicates a confident estimation of the current similarity, so the intensity of the counter-supervision signal A^h should be large as a hedge, i.e., $\frac{\partial |A^h|}{s^s} < 0$..

We summarize five conditions for the proposed deep duplex learning as follows:

$$(1)\ A^s > 0;\ \ (2)\ A^h < 0;\ \ (3)\ A^s + A^h > 0;\ \ (4)\ \frac{\partial |A^s|}{s^h} > 0;\ \ (5)\ \frac{\partial |A^h|}{s^s} < 0. \tag{7}$$

Condition (1) requires the SS to always follow the supervisor, and condition (2) requires the HS to always overturn the supervisor. Condition (3) constrains the sum of the two amending factors to be positive so that the overall effect on the model is still consistent with the supervisor. Conditions (4)(5) require the learning of both similarities to be adaptive, so that the network can better learn to balance the two training signals. We see that using Eq. (6) as the loss function follows conditions (1)(2)(3) but not conditions (4)(5).

3.3 Duplex Similarity for Efficient DDL

Our DDL is based on a simple motivation to impose both the supervision and counter-supervision signals on the network and allow the network to choose the training direction. However, the instantiation of DDL is not straightforward and trivial. We provide a simple yet effective way to achieve Eq. (7), which can act as a simple plug-and-play module to be readily applied to existing methods.

We introduce a duplex similarity function $s^d(s^s, s^h)$ as:

$$s^d(s^s, s^h) = \alpha - (\alpha - s^s)e^{-\frac{\beta - s^h}{\alpha - s^s}}, \tag{8}$$

where α and β are set to be the upper bond of s^s and s^h, respectively, i.e., $s^s \leq \alpha$ and $s^h \leq \beta$. For cosine similarities, we set $\alpha = \beta = 1$. We can directly impose the loss function L directly on s^d. The learning process of the SS s^s is then:

$$s^s = s^s - \lambda \frac{\partial L(s^d, a)}{\partial s^d} \frac{\partial s^d(s^s, s^h)}{\partial s^s} = s + \hat{R}^s(s^s, s^h), \tag{9}$$

where the amending factor for s^s is:

$$A^s(s^s, s^h) = \frac{\partial s^d(s^s, s^h)}{\partial s^s} = (1 + \frac{\beta - s^h}{\alpha - s^s})e^{-\frac{\beta - s^h}{\alpha - s^s}}. \tag{10}$$

As $s^s \leq \alpha$ and $s^h \leq \beta$, we can see that $A^s(s^s, s^h) > 0$ and easily prove that $\frac{\partial |A^s|}{s^h} > 0$.

Similarly, the amending factor for the HS s^h can be computed as:

$$A^h(s^s, s^h) = \frac{\partial s^d(s^s, s^h)}{\partial s^h} = -e^{-\frac{\beta - s^h}{\alpha - s^s}}. \tag{11}$$

We see that $A^h(s^s, s^h) < 0$ and thus $\frac{\partial |A^h|}{s^s} = \frac{\partial (-A^h)}{s^s} < 0$.

Finally, the sum of the two amending factors is:

$$A^s(s^s, s^h) + A^h(s^s, s^h) = \frac{\beta - s^h}{\alpha - s^s} e^{-\frac{\beta - s^h}{\alpha - s^s}} > 0, \tag{12}$$

which indicates that the overall similarity always follows the supervisor.

Our DDL with duplex similarity can be readily applied to most existing weakly-supervised learning methods, including learning with noisy labels, partial label learning, and semi-supervised learning methods. For a baseline method, we regard the original representation as the SR and simply add a fully connected layer to obtain the HR. We then compute the duplex similarity between images and class prototypes and substitute the original similarity in the loss function. DDL only yields very little additional computation cost compared to the original method for training. During inference, we discard HR and use the same network as the baseline method, resulting in no additional workload.

4 Experiments

In this section, we conducted various experiments on three types of weakly supervised learning tasks including learning with noisy labels, partial label learning, and semi-supervised learning. We show that the proposed deep duplex learning improves the state-of-the-art method for all three tasks.

4.1 Datasets

We followed existing weakly supervised learning methods (Liang et al., 2022; Wang et al., 2022; Li et al., 2020) to conduct experiments on the CIFAR-10 and CIFAR-100 datasets. The CIFAR-10 and CIFAR-100 datasets contain the same 60,000 images classified into 10 and 100 categories, respectively, resulting in 6,000 images per class for CIFAR-10 and 600 images per class for CIFAR-100. Among them, we used 50,000 images for training and the rest 10,000 images for evaluation.

4.2 Experimental Settings

Learning with Noisy Labels. We strictly followed the evaluation protocol of existing methods (Li et al., 2020; Tan et al., 2021; Nishi et al., 2021; Liang et al., 2022; Li et al., 2021) for fair comparisons. We generate two types of label noise to simulate the learning process with noisy labels. In the symmetric noisy setting, we randomly modify the labels of a certain percentage (20%, 50%, 80%, and 90%) of training samples with uniform possibilities to all other labels. In the asymmetric noisy setting, we modify the labels of 40% training samples only to other similar classes (e.g., automobile to truck) to simulate the distribution of real-world noise. We use a ratio of 40% since some classes turn theoretically indistinguishable when using ratios higher than 50% (Li et al., 2020).

We adopted an 18-layer PreAct ResNet (He et al., 2016b) as the backbone network. We used the SGD-M optimizer with a momentum of 0.9 and a weight decay of 0.0005. We set the batch size to 128 and trained the network for a total of 300 epochs. We used an initial learning rate of 0.02 and reduced it to 0.002 at the 150-th epoch. We also performed warm-up for 10 epochs on CIFAR-10 and 30 epochs for CIFAR-100. We reported the mean accuracy of 5 runs using random seeds.

Partial Label Learning. We strictly followed the evaluation protocol of existing methods (Feng et al., 2020b; Lv et al., 2020; Wen et al., 2021; Wang et al., 2022) for fair comparisons. We generate a label candidate set for each training sample by augmenting the ground-truth label with C additional labels. The additional labels are randomly selected with a uniform probability among negative categories. We set C to $\{1, 3, 5\}$ for CIFAR-10 and $\{1, 5, 10\}$ for CIFAR-100.

We adopted ResNet-18 (He et al., 2016a) as the backbone network for feature extraction. We employed the SGD-M optimizer with a momentum of 0.9 to train the model for 800 epochs. We set the batch size to 256. We used an initial learning rate of 0.01 with the cosine learning rate scheduler. We ran the experiments 5 times and reported the average performance with standard deviations.

Semi-Supervised Learning. We strictly followed the evaluation protocol of existing methods (Berthelot et al., 2019b;a; Sohn et al., 2020; Li et al., 2021) for fair comparisons. We randomly

Table 1: Experimental results of learning with noisy labels methods (%) on the CIFAR-10 and CIFAR-100 datasets with symmetric and asymmetric noise.

Method	Venue	CIFAR-10					CIFAR-100			
		Symm.				Asym.	Symm.			
		20%	50%	80%	90%	40%	20%	50%	80%	90%
Standard CE	-	86.8	79.4	62.9	42.7	85.0	62.0	46.7	19.9	10.1
Bootstrap	ICLRW 15	86.8	79.8	63.3	42.9	91.2	62.1	46.6	19.9	10.2
F-correction	CVPR 17	86.8	79.8	63.3	42.9	87.2	61.5	46.6	19.9	10.2
Mixup	ICLR 18	95.6	87.1	71.6	52.2	-	67.8	57.3	31.1	15.3
Co-teaching+	ICML 19	89.5	85.7	67.4	47.9	-	65.6	51.8	27.9	13.7
P-correction	CVPR 19	92.4	89.1	77.5	58.9	88.1	69.4	57.5	31.1	15.3
Meta-Learning	CVPR 19	92.9	89.3	77.4	58.7	88.6	68.5	59.2	42.4	19.5
M-correction	ICML 19	94.0	92.0	86.8	69.1	87.4	73.9	66.1	48.2	24.3
ELR+	NeurIPS 20	95.8	94.8	93.3	78.7	93.0	77.6	73.6	60.8	33.4
Co-learning	MM 21	92.5	84.8	63.5	-	81.4	66.7	55.0	36.2	-
LongReMix	Arxiv 21	96.2	95.0	93.9	82.0	94.7	77.8	75.6	62.9	33.8
Tripartite	CVPR 22	96.3	94.9	92.6	-	-	78.7	74.7	59.8	-
DivideMix	ICLR 20	96.1	94.6	93.2	76.0	93.4	77.3	74.6	60.2	31.5
DDL (DivideMix)	-	**96.4**	**95.0**	**93.3**	**81.5**	**93.6**	**77.5**	**74.9**	60.2	**31.9**
AugDesc	CVPR 21	96.3	95.6	93.8	91.9	94.6	79.6	77.6	66.4	41.2
DDL (AugDesc)	-	**96.4**	**95.7**	**94.7**	91.9	**94.8**	**79.9**	**77.8**	66.3	**41.7**

Table 2: Experimental results (%) of partial label learning methods on the CIFAR-10 dataset with different numbers of candidate labels.

Method	Venue	1 label	3 labels	5 labels
Fully Supervised	-	94.91 ± 0.07 (0 labels)		
EXP	ICML 20	79.23 ± 0.10	75.79 ± 0.21	70.34 ± 1.32
MSE	ICML 20	79.97 ± 0.45	75.64 ± 0.28	67.09 ± 0.66
CC	NeurIPS 20	82.30 ± 0.21	79.08 ± 0.07	74.05 ± 0.35
PRODEN	ICML 20	90.24 ± 0.32	89.38 ± 0.31	87.78 ± 0.07
LWS	ICML 21	90.30 ± 0.60	88.99 ± 1.43	86.16 ± 0.85
PiCO	ICLR 22	94.39 ± 0.18	94.18 ± 0.12	93.58 ± 0.06
DDL (PiCO)	-	$\mathbf{94.66 \pm 0.16}$	$\mathbf{94.45 \pm 0.06}$	$\mathbf{93.77 \pm 0.09}$

selected $\{20, 40, 80, 250\}$ training samples per class to provide them with labels and regard other samples as unlabeled for CIFAR-10. Note that each class contains 5,000 samples for training, so we only exploited a small portion (≤ 0.05) of labeled data.

We adopted a Wide ResNet-28-2 (Zagoruyko & Komodakis, 2016) as the backbone network and used the exponential-moving-average model for evaluation following existing methods (Berthelot et al., 2019b;a; Sohn et al., 2020; Li et al., 2021). We used the SGD-M optimizer with a momentum of 0.9 and a weight decay of 0.0005. We trained the model for 512 epochs with an initial learning rate of 0.03 and the cosine scheduler. We fixed the batch size to 64. We conducted all experiments with 5 different seeds and reported both the average accuracy and standard deviations.

Deep Duplex Learning. For our DDL, we employ the serial structure with one additional fully connected layer to obtain the HR unless otherwise stated. We empirically find that using a parallel structure yields better performance, but we still adopt the serial structure due to its simplicity and efficiency. We use the cosine similarity and set $\alpha = 1$ and $\beta = 2$ for all the experiments.

4.3 Learning with Noisy Labels

For learning with noisy labels, we applied the proposed DDL method to two state-of-the-art methods: DivideMix (Li et al., 2020) and AugDesc (Nishi et al., 2021). DivideMix (Li et al., 2020) dynamically partitions the training data into a labeled set with clean labels and an unlabeled set with noisy labels and then performs semi-supervised learning on them. AugDesc (Nishi et al., 2021) further improves DivideMix by using different data augmentation strategies for loss modeling and representation learning. For deep duplex learning, we replaced the similarity computing in the classifier of the original methods with the proposed duplex similarity. We adopted the same hyperparameters with the original methods without further tuning.

Table 3: Experimental results (%) of partial label learning methods on the CIFAR-100 dataset with different numbers of candidate labels.

Method	Venue	1 label	5 labels	10 labels
Fully Supervised	-	73.56 ± 0.10 (0 labels)		
EXP	ICML 20	44.45 ± 1.50	41.05 ± 1.40	29.27 ± 2.81
MSE	ICML 20	49.17 ± 0.05	46.02 ± 1.82	43.81 ± 0.49
CC	NeurIPS 20	49.76 ± 0.45	47.62 ± 0.08	35.72 ± 0.47
PRODEN	ICML 20	62.60 ± 0.02	60.73 ± 0.03	56.80 ± 0.29
LWS	ICML 21	65.78 ± 0.02	59.56 ± 0.33	53.53 ± 0.08
PiCO	ICLR 22	73.09 ± 0.34	72.74 ± 0.30	69.91 ± 0.24
DDL (PiCO)	-	**73.21 ± 0.15**	**73.04 ± 0.11**	**70.20 ± 0.16**

Table 4: Experimental results (%) of semi-supervised learning methods on the CIFAR-10 dataset with different numbers of labeled samples per class.

Method	Venue	20 labels	40 labels	80 labels	250 labels
MixMatch	NeurIPS 19	27.84±10.63	51.90±11.76	80.79±1.28	88.97±0.85
ReMixMatch	ICLR 19	-	80.90±9.64	-	94.56±0.05
FixMatch	NeurIPS 20	82.32±9.77	86.12±3.53	92.06±0.88	94.90±0.67
FixMatch w. DA	NeurIPS 20	83.81±9.35	86.98±3.40	92.29±0.86	94.95±0.66
CCSSL (FixMatch)	CVPR 22	-	90.83±2.78	-	94.86±0.55
CoMatch	ICCV 21	87.67±8.47	93.09±1.39	93.97±0.62	95.09±0.33
DDL (CoMatch)	-	**91.01±3.22**	**93.55±1.39**	**94.28±0.69**	**95.21±0.34**

In addition to the two baseline methods, we also compared the proposed DDL with other methods including the standard cross-entropy loss, Bootstrap (Reed et al., 2015), F-correction (Patrini et al., 2017), Mixup (Zhang et al., 2018), Co-teaching (Han et al., 2018), P-correction (Yi & Wu, 2019), Meta-Learning (Li et al., 2019), M-correction (Arazo et al., 2019), JoCoR (Wei et al., 2020), DivideMix (Li et al., 2020), ELR+ (Liu et al., 2020), Co-learning (Tan et al., 2021), LongReMix (Cordeiro et al., 2021), and Tripartite (Liang et al., 2022).

Table 1 shows the results on the CIFAR-10 and CIFAR-100 datasets with different noisy types and noise ratios. We use red numbers to denote the best results and bold numbers to represent improved results. We see that DivideMix (Li et al., 2020) and AugDesc (Nishi et al., 2021) demonstrates strong performance for learning with noisy labels, and the DDL further boosts the performance on nearly all the noise levels. Specifically, we observe a large performance improvement (5.5%) over DivideMix on CIFAR-10 for a large noise ratio of 90%, demonstrating the effectiveness of DDL.

4.4 Partial Label Learning

For partial label learning, we applied DDL to the best-performing method PiCO (Wang et al., 2022) to demonstrate the effectiveness of our method. PiCO (Wang et al., 2022) employs a prototype-based label disambiguation method based on a contrastive learning module to produce accurate disambiguated labels for training. For deep duplex learning with PiCO, we substitute the original representation and prototypes with the duplex ones and compute the duplex similarity for loss computation. We used the same hyperparameters with PiCO without tuning.

We also provide comparisons with other state-of-the-art methods: MSE and EXP (Feng et al., 2020a) simply adopt mean square error and exponential loss for training, respectively; CC (Feng et al., 2020b) learns a partially labeled data generation process for classifier-consistent learning; PRODEN (Lv et al., 2020) employs self-training with iterative representation training and label disambiguation; LWS (Wen et al., 2021) adopts leveraged weighted loss to consider the trade-off between losses on the candidate labels and other labels.

We present the experimental results on CIFAR-10 and CIFAR-100 with different numbers of candidate labels on Tables 2 and 3, respectively. We use red numbers to denote the best results and bold numbers to represent improved results when applying our method. We observe a consistent performance boost across all numbers of candidate labels on both datasets. Note that the standard deviations tend to decrease when equipped with the proposed DDL, which demonstrates the stability of our method. Particularly, we see that DDL (PiCO) with 3 additional candidate labels even outperforms all the other methods with only 1 additional candidate label on CIFAR-10.

Table 5: Effect of different formulations of the duplex similarity.

s^d	s^s	$s^s - 0.5s^h$	$\alpha - \frac{\alpha - s^s}{\beta - s^h}$	$\alpha - (\alpha - s^s)e^{-\frac{\beta - s^h}{2}}$	$\alpha - (\alpha - s^s)e^{-\frac{\beta - s^h}{\alpha - s^s}}$
Conditions	(1)	(1)(2)(3)	(1)(2)(4)	(1)(2)(3)(4)	(1)(2)(3)(4)(5)
Accuracy	93.8	94.2	94.2	94.5	94.7

4.5 Semi-Supervised Learning

For semi-supervised learning, we applied DDL to the state-of-the-art CoMatch (Li et al., 2021), which unifies training signals from both pseudo-labels and instance consistency for more robust training. For DDL with CoMatch, we replace the embeddings with duplex ones and compute the duplex similarities accordingly. We fix the other settings and hyperparameters the same as CoMatch.

We also compare DDL with other state-of-the-art methods. MixMatch (Berthelot et al., 2019b) produces low-entropy pseudo labels for augmented unlabeled samples and employs mixup (Zhang et al., 2018) to mix labeled and unlabeled data. ReMixMatch (Berthelot et al., 2019a) improves MixMatch with distribution alignment and augmentation anchoring. FixMatch (Sohn et al., 2020) generates pseudo labels on weakly-augmented data and employs them to perform training on strong-augmented data. CCSSL (Yang et al., 2022) further improves FixMatch by performing class-wise clustering and instance-wise contrasting on in-distribution data and out-of-distribution data.

Table 4 shows the results on CIFAR-10 with different numbers of labeled samples per class. We see that the proposed DDL uniformly improves CoMatch and achieves the best performance on all four levels of label scarcity. The performance improvement is particularly large with only 20 labeled samples (0.4%) per class, where the generated pseudo labels can be less accurate. The model benefits more from the further adaptive balance of true and false supervision signals by our DDL.

4.6 Analysis

For efficiency, we only conducted experiments on the learning with noisy labels task (80% noisy ratio) on the CIFAR-10 dataset to analyze the effect of different components of our method.

Effect of Different Formulations of the Duplex Similarity. We adopted different formulations to compute the duplex similarity and present the results in Table 5. We also show the satisfied conditions in Eq. (7) for different formulations. We see that the performance improves when satisfying more conditions and the proposed formulation attains the best result.

Effect of Hyperparameters. We fixed $\alpha = 1$ and evaluate the effect of different values of βs, as shown in Figure 4. We see that the proposed method is not sensitive to the choice of β. We observe a similar phenomenon on different αs.

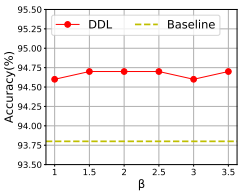

Figure 4: Effect of different values of βs.

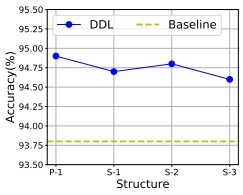

Figure 5: Effect of different structures.

Effect of Different Structures. Figure 5 shows the effect of different structures of the proposed DDL. P-1 denotes the parallel structure with a parallel convolution layer to obtain each representation. S-n denotes the serial structure with n fully connected layers following the superficial representation to obtain the hypocritical representation. We observe that the parallel structure achieves the best results. Still, due to its simplicity, we adopted S-1 as default for the main experiments.

5 Conclusion

In this paper, we have presented a deep duplex learning method for weak supervision. Unlike existing weakly-supervised learning methods, we exploit both supervision and counter-supervision signals and rely on the network itself to adaptively balance the two training signals. We have applied our DDL method to the best-performing methods of learning with noisy labels, partial label learning, and semi-supervised learning and set new state-of-the-arts on the respective tasks. Still, we only conducted experiments with synthetic weak supervision. The generalization of our method to real-world weak supervision remains unknown and is an interesting future work.

REFERENCES

Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *ICML*, pp. 312–321, 2019.

Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *ICJNN*, pp. 1–8, 2020.

David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *ICLR*, 2019a.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *NeurIPS*, 32, 2019b.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pp. 213–229, 2020.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020.

Yutian Chen, Matthew W Hoffman, Sergio Gómez Colmenarejo, Misha Denil, Timothy P Lillicrap, Matt Botvinick, and Nando Freitas. Learning to learn without gradient descent by gradient descent. In *ICML*, pp. 748–756, 2017.

Filipe R Cordeiro, Ragav Sachdeva, Vasileios Belagiannis, Ian Reid, and Gustavo Carneiro. Longremix: Robust learning with high confidence samples in a noisy label environment. *arXiv preprint arXiv:2103.04173*, 2021.

Timothee Cour, Ben Sapp, and Ben Taskar. Learning from partial labels. *JMLR*, 12:1501–1536, 2011.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020.

Lei Feng, Takuo Kaneko, Bo Han, Gang Niu, Bo An, and Masashi Sugiyama. Learning with multiple complementary labels. In *ICML*, pp. 3072–3081, 2020a.

Lei Feng, Jiaqi Lv, Bo Han, Miao Xu, Gang Niu, Xin Geng, Bo An, and Masashi Sugiyama. Provably consistent partial-label learning. *NeurIPS*, 33:10948–10960, 2020b.

Aritra Ghosh, Himanshu Kumar, and PS Sastry. Robust loss functions under label noise for deep neural networks. In *AAAI*, volume 31, pp. 1919–1925, 2017.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, pp. 630–645, 2016b.

Eyke Hüllermeier and Jürgen Beringer. Learning from ambiguously labeled examples. *Intelligent Data Analysis*, 10(5):419–439, 2006.

Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *CVPR*, pp. 5070–5079, 2019.

Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. *NeurIPS*, 32:10759–10768, 2019.

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*, pp. 2304–2313, 2018.

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICMLW*, volume 3, pp. 2, 2013.

Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Learning to learn from noisy labeled data. In *CVPR*, pp. 5051–5059, 2019.

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

Junnan Li, Caiming Xiong, and Steven CH Hoi. Comatch: Semi-supervised learning with contrastive graph regularization. In *ICCV*, pp. 9475–9484, 2021.

Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017a.

Xiaoxiao Li, Ziwei Liu, Ping Luo, Chen Change Loy, and Xiaoou Tang. Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade. In *CVPR*, pp. 3193–3202, 2017b.

Xuefeng Liang, Longshan Yao, Xingyu Liu, and Ying Zhou. Tripartite: Tackle noisy labels by a more precise partition. In *CVPR*, 2022.

Liping Liu and Thomas Dietterich. A conditional multinomial mixture model for superset label learning. *NeurIPS*, 25, 2012.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. In *NeurIPS*, 2020.

Jiaqi Lv, Miao Xu, Lei Feng, Gang Niu, Xin Geng, and Masashi Sugiyama. Progressive identification of true labels for partial-label learning. In *ICML*, pp. 6500–6510, 2020.

Aditya Menon, Brendan Van Rooyen, Cheng Soon Ong, and Bob Williamson. Learning from corrupted binary labels via class-probability estimation. In *ICML*, pp. 125–134, 2015.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *NeurIPS*, 2013.

Nam Nguyen and Rich Caruana. Classification with partial labels. In *KDD*, pp. 551–559, 2008.

Kento Nishi, Yi Ding, Alex Rich, and Tobias Hollerer. Augmentation strategies for learning with noisy labels. In *CVPR*, pp. 8022–8031, 2021.

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*, pp. 1944–1952, 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763, 2021.

Scott E Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. In *ICLRW*, 2015.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *arXiv*, abs/2106.05974, 2021.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *NeurIPS*, 33:596–608, 2020.

Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *ICCV*, 2021.

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

Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *CVPR*, pp. 5552–5560, 2018.

Yihe Tang, Weifeng Chen, Yijun Luo, and Yuting Zhang. Humble teachers teach better students for semi-supervised object detection. In *CVPR*, pp. 3132–3141, 2021.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*, 2017.

Deng-Bao Wang, Min-Ling Zhang, and Li Li. Adaptive graph guided disambiguation for partial label learning. *TPAMI*, 2021.

Haobo Wang, Yuzhou Qiang, Chen Chen, Weiwei Liu, Tianlei Hu, Zhao Li, and Gang Chen. Online partial label learning. In *ECML PKDD*, pp. 455–470, 2020.

Haobo Wang, Ruixuan Xiao, Yixuan Li, Lei Feng, Gang Niu, Gang Chen, and Junbo Zhao. Pico: Contrastive label disambiguation for partial label learning. In *ICLR*, 2022.

Xinshao Wang, Yang Hua, Elyor Kodirov, and Neil M Robertson. Imae for noise-robust learning: Mean absolute error does not treat examples equally and gradient magnitude's variance matters. *arXiv preprint arXiv:1903.12141*, 2019a.

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*, pp. 322–330, 2019b.

Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *CVPR*, pp. 13726–13735, 2020.

Hongwei Wen, Jingyi Cui, Hanyuan Hang, Jiabin Liu, Yisen Wang, and Zhouchen Lin. Leveraged weighted loss for partial label learning. In *ICML*, pp. 11091–11100, 2021.

Xiaobo Xia, Tongliang Liu, N. Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? In *NeurIPS*, 2019.

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *CVPR*, pp. 2691–2699, 2015.

Ning Xu, Jiaqi Lv, and Xin Geng. Partial label learning via label enhancement. In *AAAI*, volume 33, pp. 5557–5564, 2019.

Fan Yang, Kai Wu, Shuyi Zhang, Guannan Jiang, Yong Liu, Feng Zheng, Wei Zhang, Chengjie Wang, and Long Zeng. Class-aware contrastive semi-supervised learning. In *CVPR*, 2022.

Kun Yi and Jianxin Wu. Probabilistic end-to-end noise correction for learning with noisy labels. In *CVPR*, pp. 7017–7025, 2019.

Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *ICML*, pp. 7164–7173, 2019.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *arXiv*, abs/2106.04560, 2021.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

Min-Ling Zhang and Fei Yu. Solving the partial label learning problem: An instance-based approach. In *IJCAI*, 2015.

Min-Ling Zhang, Bin-Bin Zhou, and Xu-Ying Liu. Partial label learning via feature-aware disambiguation. In *KDD*, pp. 1335–1344, 2016.

Zhilu Zhang and Mert R Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*, 2018.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2020.