DECONFOUNDED NOISY LABELS LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Noisy labels are practical in real-world applications and cause severe performance degeneration. In this paper, first the validity of the small loss trick which plenty of noisy methods utilize is challenged. Then an empirical phenomenon named malignant bias is studied which results from the spurious correlation between noisy labels and background representation. To address this problem, unlike previous works based on statistical and regularization methods, we revisit the task from a causal perspective. A causal intervention model named deconfounded noisy labels learning (DeNLL) is applied to explicitly deconfound noisy label learning with causal adjustment, which eliminates the spurious correlation between labels and background representation and preserves true causal effect between labels and foreground representation. DeNLL implements the derived adjustment by a localization module (LM) and a debiased interaction module (DIM). LM adaptively discriminates foreground from background, and DIM dynamically encourages the interaction between the original representation and a debiased factor of the representation, which accords with the causal intervention. Experiments are carried out on five public noisy datasets including synthetic label noise, human label noise and real-world label noise. The proposed method achieves the state-of-the-art accuracy and exhibits clear improvements. Also, the proposed method is model-agnostic which improves the performances consistently on different backbones.

1 INTRODUCTION

Noisy labels are ubiquitous in practical datasets partly because of the large expense of clean human annotation and the popularity of crowdsourcing and online queries Frénay & Verleysen (2013); Algan & Ulusoy (2021). Noisy labels inevitably degenerate the robustness of deep neural networks and cause severe decrease in model performances Cordeiro & Carneiro (2020); Karimi et al. (2020). Thus, noisy labels learning is a vibrant and significant topic in recent years.

Plenty of noisy labels learning works Han et al. (2018); Yu et al. (2019); Chen et al. (2019); Jiang et al. (2018); Zhang et al. (2019); Li et al. (2019); Wei et al. (2020); Karim et al. (2022); Nishi et al. (2021) depend on an empirical trick named the *small loss trick*, which assumes that the samples with smaller loss have a higher possibility to be clean. Based on this trick, the previous works utilize the loss value to divide the dataset into a clean set and a noisy set, and further discard the noisy samples Han et al. (2018); Yu et al. (2019); Chen et al. (2019); Jiang et al. (2018), correct the noisy samples Zhang et al. (2019) or remove the labels and keep the images Li et al. (2019); Wei et al. (2020); Karim et al. (2022); Nishi et al. (2021), which turns out to be a semi-supervised learning problem. Small loss trick is reasonable (Fig.1(c)) since the network tends to first learn simple and clean patterns and then learn complicated and probably noisy patterns later Arpit et al. (2017); Zhang et al. (2021a).

However, small loss trick loses its validity in the following two circumstances (Fig.1(c)). 1) Datasets with high ratio of label noise. With large proportion of noisy samples, clean patterns can hardly be studied at early period of training and clean samples exhibit large loss. Simply using this trick results in the incorrect dataset division and degeneration (30% as in Tab.2) 2) Harder classification datasets with large intra-class variance and small inter-class variance. For example, in gait recognition Wang et al. (2003) or person-re-identification, samples from the same class are visually dissimilar since the different viewing angles, while samples from different classes are similar due to similar body proportions. In these cases, samples within a class distribute evenly. Thus the samples with large loss not necessarily be noisy. Further, we challenge that recent works depend highly on this empirical

effects and singularize the noisy labels learning problem into a fixed paradigm. Addressing this issue is essential for a healthy ecosystem of noisy labels learning in the long run.

In this paper, a different empirical phenomenon named malignant bias is studied. As illustrated in Fig.1(a), we compute the average correlation coefficient between the image representation and background representation on clean dataset, noisy dataset with baseline method Bai et al. (2021), noisy dataset with DeNLL. On the clean dataset, feature representation depends more on foreground. However, when on noisy datasets, it turns out that feature representation depends highly on background representation, and a spurious correlation between background information and noisy labels is established. We name it malignant bias since we would like to discriminate this bias from the benign bias in clean dataset. Bias denotes the correlation on background representation. In ideal conditions, the dataset is unbiased on background and the prediction depends only on foreground. In clean datasets, although the bias exists, it is benign since data distribution shares the same in train and test sets. Malignant bias degenerates the model and prevent the model to learn from true correlation. Similar observation (Fig.1(b)) is made in a different perspective in previous works Yi et al. (2022); Rao et al. (2021).

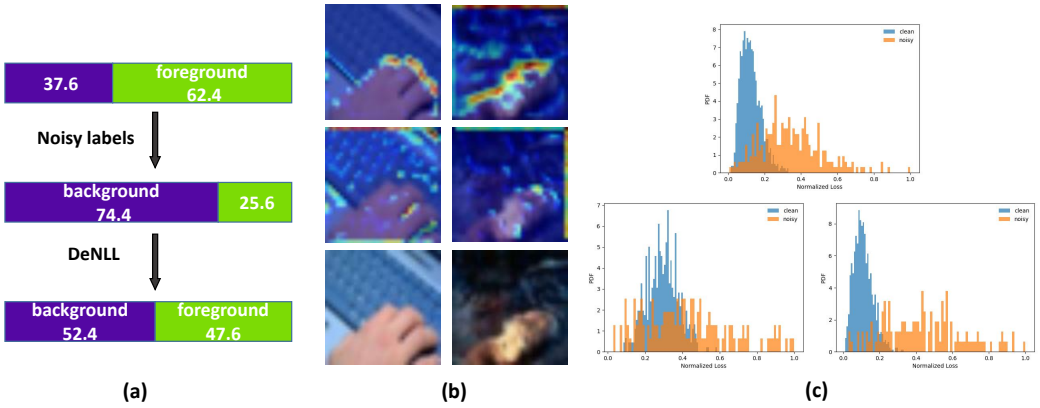

Figure 1: (a) Spurious correlation caused by noisy labels. The correlation coefficient between background representation and image representation (purple), between foreground representation and image representation. Respectively on clean CIFAR-100 (first row), synthetic label noise CIFAR-100 with baseline (second row), and synthetic label noise CIFAR-100 with DeNLL (third row). (b) Grad-Cam visualization. DeNLL (first row), baseline (second row), orginal image (third row). (c) The histogram of loss values on different datasets and different ratios. X-axis is the normalized loss values since absolute loss is not comparable among different datasets. We randomly sample 1000 clean samples and $1000 * noise\_ratio$ noisy samples for depicting the loss distribution.

Based on the empirical malignant bias, we revisit the noisy labels learning from a causal perspective. Noisy labels learning is formulated as a causal model Pearl (2009) (Fig.2), and noisy labels affects the representation of background and foreground representations and serves as their parent node. To address the spurious correlation and deconfound the biased estimation of the noisy labels on representations we propose a deconfounded noisy labels learning method (DeNLL). First, we derive the traditional noisy labels learning and show why it is biased by the spurious correlation. Then, DeNLL is proposed to address this by causal intervention, and corresponding probability is estimated. DeNLL implement the deconfounded solution in deep neural networks by a localization module (LM) and a debiased interaction module (DIM). LM adaptively discriminates foreground from background, and DIM dynamically encourages the interaction between the original representation and a debiased factor of the representation. The proposed method achieves the state-of-the-art accuracy on five popular benchmarks, showing that DeNLL establishes a better debiased correlation among variables.

Our main contributions are summarized as below:
• We challenge and do not depend on a ubiquitous small loss trick. Instead, an empirical phenomenon malignant bias is studied. Addressing this issue is essential for a healthy ecosystem of noisy labels learning in the long run.
• To the best of our knowledge, for the first time, the problem of noisy labels learning is studied from a causal-effect view. A deconfounded noisy labels learning method named DeNLL is proposed,

which derives from a causal intervention between noisy labels and background representation and unbias the spurious correlation. DeNLL contains a localization module adaptively determining the foreground from the background and a debiased interaction module encouraging the mutual information exchange between the debiased factor and original representations.
• Extensive experiments are conducted on five popular benchmarks, and DeNLL achieves the state-of-the-art performances, which demonstrates its effectiveness. DeNLL is also model-agnostic and improves the performances on different backbones consistently.

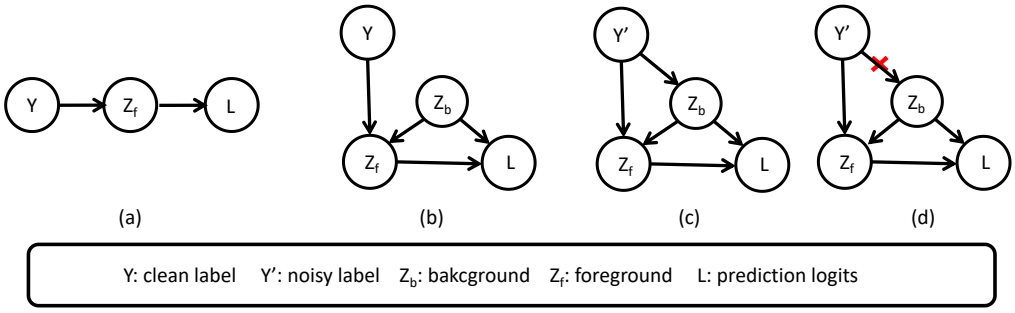

Figure 2: The causal graph of noisy image classification. (a) Ideal image classification, where the output logits depends solely on the foreground object. (b) Image classification with clean labels. Although the model depends on the background information, the test set and train set are sampled from the same distribution, and the bias on background representation is treated as the benign bias. (c) Image classification with noisy labels, where the bias on background representation is malignant bias, and the spurious relation between Z_b, Z_f and $Y^{'}$ appears as in Eq.1. (d) Deconfounded noisy labels learning, where the spurious correlation is debiased.

2 Related Works

Noisy labels learning. Due to the existence crowd-sourcing, data acquired on the web and social media, label noise is common and inescapable in real-world datasets. Noisy labels lead to model degeneration and overfitting Wang et al. (2019b); Berthelot et al. (2019); Li et al. (2019). Previous noisy labels learning methods address these from multiple aspects Song et al.; Algan & Ulusoy (2021) and can be divided into three categories. The first type of methods Jiang et al. (2018); Lee et al. (2018); Jaehwan et al. (2019) utilizes a re-annotated subset to help obtain noise-identifying ability. The second type doesn't require a clean subset, but have assumptions or prior knowledge on noise pattern, and tend to add additional noise adaption layers to implicitly or explicitly construct noise transition matrix Goldberger & Ben-Reuven; Patrini et al. (2017). The third type without the need of neither clean sets nor noise knowledge mainly focuses on regularization methods Zhang et al. (2018), robust loss design Wang et al. (2019b), meta-learning Algan & Ulusoy (2020) or ensemble methods.

Previous noisy labels learning methods greatly depend on an empirical trick named small loss trick. However, small loss trick is not always applicable and a new perspective to model noisy labels learning is necessary. Thus, we propose a different empirical phenomenon and a corresponding method to address the problem. Note that the proposed method needs neither clean sets nor noise knowledge and falls in the third type.

Causal Inference in Deep Learning. Causal inference is now a critical research topic which could endow deep learning models the ability to learn the true casual effects instead of the statistical spurious correlation. Causal inference can be applied to many domains, such as computer vision (CV) Lopez-Paz et al. (2017); Niu et al. (2021); Wang et al. (2020); Tang et al. (2020), natural language processing (NLP) Wei & Zou (2019); Schuster et al. (2019); Mahabadi et al. (2019), recommendations Schnabel et al. (2016); Wang et al. (2019a); Zhang et al. (2021b) and so on.

For example, Niu et al. (2021) proposed to eliminate the language bias, and only capture the direct causal effect of questions on answers by subtracting the direct language effect from the total causal effect. Tang et al. (2020) utilized counterfactual inference to remove the bias introduced by the image content in scene graph generation task. Mahabadi et al. (2019) employed a bias-only model to

identify dataset biases. Schnabel et al. (2016) utilized the traditional Inverse Propensity Weighting (IPW) method to handle selection biases in recommendations.

3 METHODOLOGY

3.1 BENIGN BIAS AND MALIGNANT BIAS IN NOISY LABELS LEARNING

In this section, firstly the small loss experimental phenomenon that most works Han et al. (2018); Yu et al. (2019); Chen et al. (2019); Jiang et al. (2018); Zhang et al. (2019); Li et al. (2019); Wei et al. (2020); Karim et al. (2022); Nishi et al. (2021) utilize is challenged. Then an empirical phenomenon named malignant bias is studied and it results from spurious relation caused by noisy labels. After this, we formulate the noisy labels learning from the causal perspective and a causal graph is built up to explicitly analyze the causal relations in the conventional noisy labels learning task. Then we derive a solution In the Section 3.2 and propose the implementation in deep neural network in the Section 3.3.

As shown in the Fig.1(c), small loss trick works with a prerequisite, where the loss values of clean samples are clustered at low value interval while the loss of noisy samples are clustered at high value interval. However, with two counterexamples of different noisy datasets and different noise settings, small loss trick loses its effectiveness under two circumstances. First, in the large noise ratio case, the network hardly learns any meaningful patterns at all training periods, where the network easily overfits to the noise and the differences between loss of clean samples and noisy samples are narrowed down (Fig.1(c)). Almost 40% noisy data also has small loss values. And the large performance degeneration (25% on CIFAR-10) Zhang et al. (2018); Li et al. (2019) also demonstrates the failure of the effectiveness. Second, in the datasets with large intra-class variance and small inter-class variance, due to the visual dissimilarity within a class (e.g., images of the same person but taken from different view points), a proportion of the clean samples have large loss values while the noisy samples can be of small loss values. In these cases, the small loss trick does not necessarily hold and the methods based on the trick do not work well.

Instead of utilizing small loss trick, a different empirical phenomenon is observed and modeled. 1) **Unbiased dataset**. In the ideal image classification, we expect the output image class logits L to be irrelevant to the background feature representation B. Whatever the background is, the foreground information is enough for the network to determine the output class. Thus, the data and the correlation is unbiased, as in $P(L|f,b) = P(L|f)$. 2) **Benign biased dataset**. Obviously, it's impractical for the dataset to be totally unbiased on the background due to the limited number of samples, as well as the uneven distribution of backgrounds within classes. E.g. class 'dog' has a preference for grass ground and class 'cat' has a preference for indoor environments. However, the bias in background can be **benign bias** for the whole task since the data distribution shares the same when in the train set and test set. In this case, $P_{train}(L|f,b) = P_{test}(L|f,b) \neq P(L|f)$, which indeed improves the performance.

Malignant biased dataset. However, in noisy labels learning, as shown in Fig.2, the situation becomes complicated. Firstly, the bias becomes malignant since the noisy train set and the clean test set no more shares the same data distribution. Malignant bias causes overfitting and results in severe performance degeneration. Secondly, an empirical phenomenon resulting from spurious relation among noisy labels and background representation is observed. Since the foreground shares little similarity with the noisy label, the network tends to focus on background instead and extract spurious background features to represent the class of the given image. Through our experiments, training on clean dataset, the feature representation and the foreground representation are highly correlated, while in noisy label learning, the feature representation and the background representation are highly correlated. This spurious correlation caused by noisy labels confuses the learning period and causes overfitting.

Thus, we consider the following structural model to describe the spurious correlation mechanism. Uppercase character (e.g. B) denotes a random variable and lowercase character (e.g. b) denotes its specific value; calligraphic font (e.g., $\mathcal{B}$) denotes the sample space of the corresponding random variable, and $f(\cdot)$ represents probability distribution of a random variable.

As shown in Fig.2, in the directed acyclic graph $G = \{N, E\}$, there are four variables: $Y^{'}$, Z_b, Z_f, L, and the edges in the graph describe the causal relations between variables, e.g., $Y^{'} \rightarrow Z_f$ means that $Y^{'}$ has a direct causal effect Pearl (2009) on Z_f, i.e., changes on $Y^{'}$ will affect the representation of Z_f.

• Y and $Y^{'}$ is respectively the clean label and the noisy label. Y is unknown in the task of noisy labels learning. Here, we do not make any assumptions about the label noise distribution or label noise rate.

• Z_b and Z_f is the representation feature of the background and foreground of the given image.

• L is the output classification prediction logits of the given image.

• $Y^{'} \rightarrow Z_f$ and $Y^{'} \rightarrow Z_b$. The noisy labels affect the representation of the background/foreground representation since the noisy labels supervise the network to update its parameters.

• $Z_b \rightarrow Z_f$, the background representation affects the foreground representation, which accords with malignant bias. When the label is clean, the network tends to focus on the right foreground and the background features are close to a random variable. Otherwise when the label is noisy, the network tends to focus on the background and thus id-irrelevant background details are learnt.

• $Z_b \rightarrow L$ and $Z_f \rightarrow L$, background and foreground representations both affect the output classification predictions.

According to the causal theory Pearl (2009), $Y^{'}$ is a confounder between Z_f and Y, resulting in the spurious correlation.

The malignant bias is formulated as:

$$Z_f = m_1(Y^{'}, Z_b) = m_2(Y, Z_b, \eta), \eta \perp\!\!\!\perp Z_b, Z_b \perp\!\!\!\perp Y^{'} \mid Y, \eta \sim F \tag{1}$$

where Z_f and Z_b are the representations of foreground and background, η denotes the noise, $m_1(\cdot, \cdot)$ and $m_2(\cdot, \cdot, \cdot)$ demotes the unknown structual functions. Note that every variable here are a random variable with a certain distribution. The noise η is independent of the representation. And given the clean unknown label Y, the foreground feature Z_f is is independent of noisy label $Y^{'}$. F indicates the noise distribution.

Although the function $m_2(\cdot)$ has better properties of independence, in the noisy labels setting, the clean label Y is unknown, and in our setting, the noisy distribution F is also unknown, which further make the estimation of $m(\cdot)$ intractable. To address the malignant bias caused by noisy labels, we then propose the following causal model.

3.2 Causal Intervention in Noisy labels Learning

In this section, first we estimate the conditional probability based on the conventional noisy labels learning, as follows.

$$\begin{aligned} P(L|Z_f = f, Z_b = b) &\overset{(a)}{=} \frac{\sum_{y^{'} \in \mathcal{Y}^{'}} P(y^{'}) P(b|y^{'}) P(f|y^{'}, b) P(L|f, b)}{P(b) P(f|b)} \\ &\overset{(b)}{=} \frac{\sum_{y^{'} \in \mathcal{Y}^{'}} P(y^{'}|b) P(f|y^{'}, b) P(L|f, b)}{P(f|b)} \\ &\overset{(c)}{=} \frac{\sum_{y^{'} \in \mathcal{Y}^{'}} P(y^{'}|b) P(m_1(y^{'}, Z_b) = f) P(L|m_1(y^{'}, Z_b), b)}{\sum_{y^{'} \in \mathcal{Y}^{'}} P(m_1(y^{'}, Z_b) = f) P(y^{'})} \end{aligned} \tag{2}$$

where (a) is the law of total probability, (b) follows the Bayes rule. In (c), we substitute the probability with the Eq.1.

The term $P(y^{'}|b)P(m_1(y^{'}, Z_b) = f)$ is a weighting coefficient of the probability $P(L|f, b)$, and the weighting term is corrupted by the spurious correlation $m(\cdot)$. To be specific, the noisy labels upon distribution of background information serves as an incorrect confusing term of correct probabilities. For example, given a representation of a dog and a grassland, the correct prediction $P(L|f, b)$ depends mainly on the dog foreground and predict the logits $(0.8, 0.2)$ with dog and cat class. However, given the $P(m_1(y^{'}, Z_b) = f)$, which is corrupted by noisy samples with grassland background labeled as

cat. $P(m_1(y^{'}, Z_b) = f)$ confuses the network by making incorrect weighting $(0.5, 3)$ and results in a logits $(0.4, 0.6)$. The spurious correlation is the caused by the misleading noisy labels and can be alleviated by utilizing the tool of causal intervention.

To unbias the correlation between noisy labels and the background information, we estimate the causal effect of labels and background information and adapt causal adjustment, which can be derived as:

$$\begin{aligned} P(Y|do(Z_b = b), Z_f = f) &\overset{(a)}{=} \frac{\sum_{y' \in \mathcal{Y}'} P(y^{'})P(b)P(f|y^{'}, do(b))P(L|f, do(b))}{P(f|b)P(b)} \\ &\overset{(b)}{=} \frac{\sum_{y' \in \mathcal{Y}'} P(y^{'})P(f|y^{'}, do(b))P(L|f, do(b))}{P(f|b)} \\ &\overset{(c)}{=} \frac{\sum_{y' \in \mathcal{Y}'} P(f, y^{'}|b)P(L|m_1(y^{'}, b), b)}{\sum_{y' \in \mathcal{Y}'} P(m_1(y', Z_b) = f)P(y')} \end{aligned} \tag{3}$$

where similar to the statistical equation, (a) is the law of total probability, (b) follows the Bayes rule.(c) is because the independence after the do operation: $P(f, y^{'}|b) = P(f|b, y^{'})P(y^{'}|b) = P(f|b, y^{'})P(y^{'})$.

Compare Eq.2(c) with Eq.3(c), the probability is deconfounded without a $m(\cdot)$ as a weighting term and $P(f, y^{'}|b)$ can be estimated by sampling strategies.

Approximation of Causal Intervention. To implement the causal intervention in the noisy labels learning tasks, approximation is made, which accords with the practical situations in most cases. Suppose the number of samples of different classes is balanced, then the denominator can be simplified as $1/N * \sum_{y' \in \mathcal{Y}'} P(m_1(y^{'}, Z_b) = f)$. Further, by uniformly sampling $y^{'}$, the denominator can be approximated by the following equation $1/N(\sum P(m_1(C_0, Z_b) = f) + \sum P(m_1(C_1, Z_b) = f) + ... + \sum P(m_1(C_N, Z_b) = f)$, where C_i denotes the i^{th} class. By replacing the possibility with a designed network $h(\cdot)$ and moving the expectation into the input of the function, $\sum P(m_1(C_0, Z_b) = f) \approx \sum h(C_0, Z_b) \approx h(C_0, \sum_{b \in \hat{\mathcal{B}}} Z_b)$, where $\hat{\mathcal{B}}$ denotes the class background, which is a subset of $\mathcal{B}$. Hence, the denominator is approximated by a designed module with a averaged background representation.

$$\sum_{y' \in \mathcal{Y}'} P(m_1(y^{'}, Z_b) = f)P(y^{'}) \approx 1/N(\sum_{i=0}^{N} h(C_i, \sum_{b \in \hat{\mathcal{B}}} Z_b)) \tag{4}$$

where N denotes the class numbers. And the error of the approximation can be bounded by $\frac{|h(x)-h(\mu)|}{|x-\mu|^a+|x-\mu|^b}$, where the μ is the expectation of x, a is the big-O upper-bound of $h(x)$ approximating $h(\mu)$, b is the big-O upper-bound of $h(x)$ approximating $h(\infty)$, and the proof can be found in Appendix and similar approximations Wang et al. (2021); Gao et al. (2017).

Given the background representation Z_b, the class condition serves as a unbiased weighting parameter, and is estimated by sampling.

Advantages of Causal Intervention. In spite of making approximations and using frequency to approximate probability, the causal intervention benefits noisy labels learning from the following aspects.

• After the do operation, the bias of background representation on noisy label is alleviated, as can be shown in Eq.2(b) and Eq.3(b). The background information is disentangled from the noisy labels by an expectation of the class background in Eq.4.

• By cutting off the undesirable edge between labels and background representation and preserving the useful edge between labels and foreground representation, the model mines the true correlation and is able to alleviate malignant bias.

• The causal-effect based method does not require prior knowledge about noise distribution or noise rate, which makes it more practical and easier to utilize. The causal-effect based method is model agnostic, which can be implemented differently with flexibility. The causal-effect based method does not rely on the small loss trick.

Table 1: Experimental results on CIFAR-10 and CIFAR-100 (without semi-supervised learning). The mean and standard deviation are computed over five runs.

Dataset	Method	Symmetric		Pairflip	Instance	
		20%	50%	45%	20%	40%
CIFAR10	CE	84.00±0.66	75.51±1.24	63.34±6.03	85.10±0.68	77.00±2.17
	Co-teaching	87.16±0.11	72.80±0.45	70.11±1.16	86.54±0.11	80.98±0.39
	Forward	85.63±0.52	77.92±0.66	60.15±1.97	85.29±0.38	74.72±3.24
	Joint Optim	89.70±0.11	85.00±0.17	82.63±1.38	89.69±0.42	82.62±0.57
	T-revision	89.63±0.13	83.40±0.65	77.06±6.47	90.46±0.13	85.37±3.36
	DMI	88.18±0.36	78.28±0.48	57.60±14.56	89.14±0.36	84.78±1.97
	CDR	89.72±0.38	82.64±0.89	73.67±0.54	90.41±0.34	83.07±1.33
	PES	92.38±0.40	87.45±0.35	88.43±1.08	92.69±0.44	89.73±0.51
	Ours	**93.96±0.21**	**89.10±0.77**	**90.73±0.17**	**93.24±0.32**	**91.20±0.58**
CIFAR100	CE	51.43±0.58	37.69±3.45	34.10±2.04	52.19±1.42	42.26±1.29
	Co-teaching	59.28±0.47	41.37±0.08	33.22±0.48	57.24±0.69	45.69±0.99
	Forward	57.75±0.37	44.66±1.01	27.88±0.80	58.76±0.66	44.50±0.72
	Joint Optim	64.55±0.38	50.22±0.41	42.61±0.61	65.15±0.31	55.57±0.41
	T-revision	65.40±1.07	50.24±1.45	41.10±1.95	60.71±0.73	51.54±0.91
	DMI	58.73±0.70	44.25±1.14	26.90±0.45	58.05±0.20	47.36±0.68
	CDR	66.52±0.24	55.30±0.96	43.87±1.35	67.33±0.67	55.94±0.56
	PES	68.89±0.45	58.90±2.72	57.18±1.44	70.49±0.79	65.68±1.41
	Ours	**71.59±0.44**	**63.01±0.51**	**57.35±0.13**	**72.05±0.07**	**66.41±1.39**

3.3 Deconfounded Implementation in Deep Neural Networks

The training structure and the testing structure remains the same and trained in an end-to-end manner, which is different from some causal literature Wang et al. (2021); Zhang et al. (2021b). The accordance of training and testing though brings less space for adjusting parameters and is easier to apply and interpret. In this paper, DeNLL is implemented as follows.

As stated in the previous paragraph, the causal intervention can be intractable to directly calculated since the unavailable prior knowledge about the noise and the input distribution. In DeNLL, firstly a localization module generate input for the node f and b respectively, then the denominator and the numerator is predicted from two networks with mutual information share in replace of a direct division which accumulates the noise.

Localization Module (LM). In order to mimic the foveation of the human eye to discriminate the foreground from background, inspired by Huang et al. (2021), DeNLL contains a localization module as in Eq.5 to predict a Gaussian mask with a center (p, q) and isotropic variance σ^2.

$$\begin{aligned} M(i,j) &= exp(\frac{-(i-p_x)^2-(j-p_y)^2}{2\sigma^2}) \\ G_f[x,y] &= \text{Relu}(M(i,j) * I[i,j] - m(M(i,j) * I[i,j])) \\ G_b[x,y] &= \text{Relu}(-M(i,j) * I[i,j] + m(M(i,j) * I[i,j])) \end{aligned} \tag{5}$$

where (i, j) is the spatial index of a point in the input image. $m(\cdot)$ denotes a function to get a median value of a set.

Debiased Interaction Module (DIM). Direct division in Eq.3(c) causes two problems in experimental performances. One is that the denominator prediction with a near-zero value leads to unstable training. The other is the noise aggregation due to the sum of multiple terms in both denominator and numerator. Thus, instead of calculating a division here, DeNLL utilizes two networks for predicting the probability as well as enforcing two levels of feature interaction, as in Eq.6. One network learns the represention as in Eq.4 for debiased prediction.

$$\mathcal{L}(g_1(f,b), g_2(f, \sum_{C_i} b), y') = \mathcal{L}(dy_1(g_1(f,b), g_2(f, \sum_{C_i} b), dy_2(g_1(f,b), g_2(f, \sum_{C_i} b)), y') \tag{6}$$

where $g_1(\cdot)$ and $g_2(\cdot)$ denotes two networks for predicting probability given a fixed background and a deconfounded background (Eq.4). $dy_1(\cdot,\cdot)$ and $dy_2(\cdot,\cdot)$ denotes two mutual information dynamic gate for estimating the output logits.

Table 2: Experimental results on CIFAR-10 and CIFAR-100 with symmetric, instance-dependent and pairflip label noise from different levels (semi-supervised learning). Results are token from Bai et al. (2021). The mean and standard deviation are computed over three runs.

Dataset	CIFAR-10				CIFAR-100			
Methods / Noise	Sym-20%	Sym-50%	Sym-80%	Pair-45%	Sym-20%	Sym-50%	Sym-80%	Pair-45%
CE	86.5±0.6	80.6±0.2	63.7±0.8	74.9±1.7	57.9±0.4	47.3±0.2	22.3±1.2	38.5±0.6
MixUp	93.2±0.3	88.2±0.3	73.3±0.3	82.4±1.0	69.5±0.2	57.1±0.6	34.1±0.6	44.2±0.5
M-correction*	94.0	92.0	86.8	-	73.9	66.1	48.2	-
DivideMix*	95.2	94.2	93.0	-	75.2	72.8	58.3	-
DivideMix	95.6±0.1	94.6±0.1	92.9±0.3	85.6±1.7	75.3±0.1	72.7±0.6	56.4±0.3	48.2±1.0
ELR+	94.9±0.2	93.6±0.1	90.4±0.2	86.1±1.2	75.5±0.2	71.0±0.2	50.4±0.8	65.3±1.3
PES	**95.9±0.1**	95.1±0.2	93.1±0.2	94.5±0.3	77.4±0.3	74.3±0.6	61.6±0.6	73.6±1.7
Ours (Semi)	95.57±0.15	**95.59±0.05**	**94.38±0.06**	**94.9±0.10**	**78.75±0.07**	**76.58±0.28**	**64.06±0.41**	**75.83±0.85**

4 EXPERIMENTS

Datasets. We verify the effectiveness of our approach on five benchmark datasets: synthetic noise on CIFAR10 and CIFAR-100 Krizhevsky et al. (2009), human-annotated real-world noisy labels CIFAR-10N and CIFAR-100N Wei et al. (2022), and Clothing-1M Xiao et al. (2015), which are widely used for evaluation of noisy labels in previous literature Bai et al. (2021); Li et al. (2019). The detail description of the noisy datasets are described in Appendix.

Network structures and training hyperparameters are described in detail in Appendix.

Baselines. Methods based on semi-supervised learning usually outperforms methods without semi-supervised learning, but the former takes more than 3 times computational resources/time as much as the latter. Thus, following previous works Bai et al. (2021); Liu et al. (2022) we both implement our method on semi-supervised learning and without semi-supervised learning. Note that the method is model agnostic, and we implement the method on two Reset based architecture, PES Bai et al. (2021) and SOP Liu et al. (2022). The experimental parameters remain the same as in the original works.

Compare with State-Of-The-Art Methods on Synthetic CIFAR10 and CIFAR-100. For semi-supervised learning, in Table 2, we compare DeNLL with other semi-supervised methods including Mixup Zhang et al. (2018), M-correlation Arazo et al. (2019), DivideMix Li et al. (2019), Early-learning regularization (ELR+) Liu et al. (2020), PES Bai et al. (2021). The proposed DeNLL outperforms other semi-supervised methods across nearly all noise ratios, which demonstrate the effectiveness of DeNLL. In Table 2 shows experimental results on CIFAR-10 and CIFAR-100 with instance-dependent and pairflip label noise from different levels. The proposed DeNLL exhibits a clear improvement compared with other semi-supervised methods across nearly all noise ratios.

Table 1 shows the results on synthetic CIFAR-10 and CIFAR-100 without semi-supervised learning. DeNLL outperforms state-of-the-art methods (without semi-supervised learning) including Co-teaching Han et al. (2018), Joint Optim Tanaka et al. (2018), T-revision Xia et al. (2019), DMI Xu et al. (2019), CDR Xia et al. (2020), PES Bai et al. (2021) across nearly all noise ratios. The above results show that DeNLL can make improvements on both semi-supervised and without semi-supervised pipelines.

Compare with State-Of-The-Art Methods on CIFAR-10N and CIFAR-100N. Table 3 shows the results on CIFAR-10N and CIFAR-100N with different generation methods of label noise. We report the last epoch test accuracy over the 5 independent runs. DeNLL is compared to Co-teaching, JoCoR Wei et al. (2020), ELR+, DivideMix, CORES Cheng et al. (2020), SOP, PES. We can safely draw the following observations. 1) DeNLL outperforms state-of-the-art methods across all noise settings. 2) the improvement is substantial (2.5% in accuracy) for the more challenging CIFAR-100 with high noise ratios, which accords with the analysis before. With harder noisy datasets, the small loss trick does not work so well, and a causal view of mining the real correlation is needed. 3) DeNLL exhibits a stable performance and a smaller deviation in most cases, which again demonstrate the effectiveness of the proposed method.

Table 3: Experimental results on human label noise CIFAR-10N and CIFAR-100N. Mean and standard deviation over 5 independent runs are reported. The results of the baseline methods are taken from Wei et al. (2022) which all use ResNet34 as the architecture.

Methods	CIFAR-10N					CIFAR-100N
	Random 1	Random 2	Random 3	Aggregate	Worst	Noisy Fine
CE	$85.02_{\pm 0.65}$	$86.46_{\pm 1.79}$	$85.16_{\pm 0.61}$	$87.77_{\pm 0.38}$	$77.69_{\pm 1.55}$	$55.50_{\pm 0.66}$
Forward	$86.88_{\pm 0.50}$	$86.14_{\pm 0.24}$	$87.04_{\pm 0.35}$	$88.24_{\pm 0.22}$	$79.79_{\pm 0.46}$	$57.01_{\pm 1.03}$
Co-teaching	$90.33_{\pm 0.13}$	$90.30_{\pm 0.17}$	$90.15_{\pm 0.18}$	$91.20_{\pm 0.13}$	$83.83_{\pm 0.13}$	$60.37_{\pm 0.27}$
JoCoR	$90.30_{\pm 0.20}$	$90.21_{\pm 0.19}$	$90.11_{\pm 0.21}$	$91.44_{\pm 0.05}$	$83.37_{\pm 0.30}$	$59.95_{\pm 0.24}$
ELR+	$94.43_{\pm 0.41}$	$94.20_{\pm 0.24}$	$94.34_{\pm 0.22}$	$94.83_{\pm 0.10}$	$91.09_{\pm 1.60}$	$66.72_{\pm 0.07}$
DivideMix	$95.16_{\pm 0.19}$	$95.23_{\pm 0.07}$	$95.21_{\pm 0.14}$	$95.01_{\pm 0.71}$	$92.56_{\pm 0.42}$	$71.13_{\pm 0.48}$
CORES*	$94.45_{\pm 0.14}$	$94.88_{\pm 0.31}$	$94.74_{\pm 0.03}$	$95.25_{\pm 0.09}$	$91.66_{\pm 0.09}$	$55.72_{\pm 0.42}$
SOP	$95.28_{\pm 0.13}$	$95.31_{\pm 0.10}$	$95.39_{\pm 0.11}$	$95.61_{\pm 0.13}$	$93.24_{\pm 0.21}$	$67.81_{\pm 0.23}$
PES(semi)	$95.06_{\pm 0.15}$	$95.19_{\pm 0.23}$	$95.22_{\pm 0.13}$	$94.66_{\pm 0.18}$	$92.68_{\pm 0.22}$	$70.36_{\pm 0.33}$
Ours	$\mathbf{96.01}_{\pm 0.09}$	$\mathbf{96.11}_{\pm 0.09}$	$\mathbf{96.16}_{\pm 0.14}$	$\mathbf{95.81}_{\pm 0.11}$	$\mathbf{94.52}_{\pm 0.14}$	$\mathbf{72.96}_{\pm 0.34}$

Table 4: Experimental results of abaltion study on synthetic noise CIFAR-100 (without semi-supervised learning).

		Ours	w/o LM	w/o DIM
Symm	20%	$71.59_{\pm 0.44}$	$68.47_{\pm 1.61}$	$70.95_{\pm 0.16}$
	50%	$63.01_{\pm 0.51}$	$57.34_{\pm 1.41}$	$55.53_{\pm 2.60}$
Pair	45%	$57.35_{\pm 0.13}$	$48.16_{\pm 0.23}$	$53.04_{\pm 1.40}$
Inst	20%	$72.05_{\pm 0.07}$	$67.46_{\pm 1.37}$	$68.43_{\pm 0.32}$
	40%	$66.41_{\pm 1.39}$	$62.61_{\pm 1.04}$	$64.26_{\pm 0.39}$

Table 5: Experimental results on Clothing-1M.

Method	Accuracy
CE	69.21
Forward	69.84
ELR+	**74.81**
SOP	73.55
Baseline	73.68
Ours	74.45

Compare with State-Of-The-Art Methods on Clothing1M. Table 5 shows the results on a real-world noisy labels dataset Clothing1M. DeNLL compares with other recent methods including CE, Forward, Joint-Optim, DMI, and T-revision, DivideMix, ELR+ and PES. The baseline results are our reproduction of PES. DeNLL improves the accuracy about 0.8%, showing that it works well with real-world noise problem.

Model-agnostic results. Since DeNLL does not have constraints on the baseline architecture, we conduct experiments on different baselines methods. Experimental results based on PES without semi-supervised (Table.2) and based on DivideMix with semi-supervised (Table.1) shows the method improve the results regardless of the backbones. In SOP implementation, DeNLL achieves an accuracy of 80.7% with asymmetric (with 40%) label noise on CIFAR-100, with a significant improvement of 2.7%, which again demonstrates the effectiveness of causal intervention.

Ablation Study As shown in Table. 4, the following conclusions are made. 1) In the experiment without a localization module, the foreground and background are generated with a fixed mask $24 * 24$ in the center. Compare the DeNLL with DeNLL without LM, we can see the localization module improves the performance, since it is automatic and can adaptively adjust to the input. 2) Experiment without DIM lacks one factor in the causal intervention, which shows a decrease compared to the DeNLL. This again demonstrates causal inference is beneficial in noisy labels learning.

5 Conclusion

In this paper, firstly a phenomenon named malignant bias causef by noisy labels is proposed, where the background representation and the noisy labels exhibit a spurious correlation. To address the spurious correlation as well as the biased dataset, noisy labels learning is revisited from a causal-effect view and the deconfounded equations are derived. Based on the direction of the causal equation, the method DeNLL is proposed, which contains a location module to discriminate foreground and background and a dynamic mutual information exchange between two networks for estimating the final probability. By using the causal deconfounded method, spurious correlation is alleviated while the real correlation is preserved. DeNLL is evaluated on five different benchmarks and achieves the state-of-the-art performances. Moreover, DeNLL is model-agnostic and improves performances regardless of the baseline models.

REFERENCES

Görkem Algan and Ilkay Ulusoy. Meta soft label generation for noisy labels. In *International Conferance on Pattern Recognition, ICPR*, 2020.

Görkem Algan and Ilkay Ulusoy. Image classification with deep learning in the presence of noisy labels: A survey. *Knowledge-Based Systems*, 215:106771, 2021.

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

Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 233–242. PMLR, 06–11 Aug 2017.

Yingbin Bai, Erkun Yang, Bo Han, Yanhua Yang, Jiatong Li, Yinian Mao, Gang Niu, and Tongliang Liu. Understanding and improving early stopping for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34:24392–24403, 2021.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Pengfei Chen, Ben Ben Liao, Guangyong Chen, and Shengyu Zhang. Understanding and utilizing deep neural networks trained with noisy labels. In *International Conference on Machine Learning*, pp. 1062–1070. PMLR, 2019.

Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. In *International Conference on Learning Representations*, 2020.

Filipe R Cordeiro and Gustavo Carneiro. A survey on deep learning with noisy labels: How to train your model when you cannot trust on the annotations? In *2020 33rd SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*, pp. 9–16. IEEE, 2020.

Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2013.

Xiang Gao, Meera Sitharam, and Adrian E Roitberg. Bounds on the jensen gap, and implications for mean-concentrated distributions. *arXiv preprint arXiv:1712.05267*, 2017.

Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer.

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. *Advances in neural information processing systems*, 31, 2018.

Zhen Huang, Dixiu Xue, Xu Shen, Xinmei Tian, Houqiang Li, Jianqiang Huang, and Xian-Sheng Hua. 3d local convolutional neural networks for gait recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14920–14929, 2021.

Lee Jaehwan, Yoo Donggeun, and Kim Hyo-Eun. Photometric transformer networks and label adjustment for breast density prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 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 *International Conference on Machine Learning*, pp. 2304–2313. PMLR, 2018.

Nazmul Karim, Mamshad Nayeem Rizve, Nazanin Rahnavard, Ajmal Mian, and Mubarak Shah. Unicon: Combating label noise through uniform selection and contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9676–9686, 2022.

Davood Karimi, Haoran Dou, Simon K Warfield, and Ali Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 65:101759, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5447–5456, 2018.

Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2019.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33:20331–20342, 2020.

Sheng Liu, Zhihui Zhu, Qing Qu, and Chong You. Robust training under label noise by over-parameterization. *arXiv preprint arXiv:2202.14026*, 2022.

David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, and Léon Bottou. Discovering causal signals in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6979–6987, 2017.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. End-to-end bias mitigation by modelling biases in corpora. *arXiv preprint arXiv:1909.06321*, 2019.

Kento Nishi, Yi Ding, Alex Rich, and Tobias Hollerer. Augmentation strategies for learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8022–8031, 2021.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12700–12710, 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1944–1952, 2017.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1025–1034, 2021.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *international conference on machine learning*, pp. 1670–1679. PMLR, 2016.

Tal Schuster, Darsh J Shah, Yun Jie Serene Yeo, Daniel Filizzola, Enrico Santus, and Regina Barzilay. Towards debiasing fact verification models. *arXiv preprint arXiv:1908.05267*, 2019.

Hwanjun Song, Minseok Kim, Dongmin Park, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey.

Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5552–5560, 2018.

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3716–3725, 2020.

Liang Wang, Tieniu Tan, Huazhong Ning, and Weiming Hu. Silhouette analysis-based gait recognition for human identification. *IEEE transactions on pattern analysis and machine intelligence*, 25(12): 1505–1518, 2003.

Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10760–10770, 2020.

Wenjie Wang, Fuli Feng, Xiangnan He, Xiang Wang, and Tat-Seng Chua. Deconfounded recommendation for alleviating bias amplification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 1717–1725, 2021.

Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. Doubly robust joint learning for recommendation on data missing not at random. In *International Conference on Machine Learning*, pp. 6638–6647. PMLR, 2019a.

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13726–13735, 2020.

Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TBWA6PLJZQm.

Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? *Advances in Neural Information Processing Systems*, 32, 2019.

Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *International conference on learning representations*, 2020.

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2691–2699, 2015.

Yilun Xu, Peng Cao, Yuqing Kong, and Yizhou Wang. L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise. *Advances in neural information processing systems*, 32, 2019.

Li Yi, Sheng Liu, Qi She, A Ian McLeod, and Boyu Wang. On learning contrastive representations for learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16682–16691, 2022.

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

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021a.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Weihe Zhang, Yali Wang, and Yu Qiao. Metacleaner: Learning to hallucinate clean representations for noisy-labeled visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7373–7382, 2019.

Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. Causal intervention for leveraging popularity bias in recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 11–20, 2021b.