Learning Causal Transition Matrix for Instance-dependent Label Noise

Jiahui Li[1, 2*], Tai-Wei Chang[2*], Kun Kuang[1†], Ximing Li[2], Long Chen[3], Jun Zhou[2]
[1]Zhejiang University
[2]Ant Group
[3]The Hong Kong University of Science and Technology
jiahuil@zju.edu.cn, taiwei.twc@antgroup.com, kunkuang@zju.edu.cn
xili.lxm@antgroup.com, longchen@ust.hk, jun.zhoujun@antgroup.com

Abstract

Noisy labels are both inevitable and problematic in machine learning methods, as they negatively impact models' generalization ability by causing overfitting. In the context of learning with noise, the transition matrix plays a crucial role in the design of statistically consistent algorithms. However, the transition matrix is often considered unidentifiable. One strand of methods typically addresses this problem by assuming that the transition matrix is instance-independent; that is, the probability of mislabeling a particular instance is not influenced by its characteristics or attributes. This assumption is clearly invalid in complex real-world scenarios. To better understand the transition relationship and relax this assumption, we propose to study the data generation process of noisy labels from a causal perspective. We discover that an unobservable latent variable can affect either the instance itself, the label annotation procedure, or both, which complicates the identification of the transition matrix. To address various scenarios, we have unified these observations within a new causal graph. In this graph, the input instance is divided into a noise-resistant component and a noise-sensitive component based on whether they are affected by the latent variable. These two components contribute to identifying the "causal transition matrix", which approximates the true transition matrix with theoretical guarantee. In line with this, we have designed a novel training framework that explicitly models this causal relationship and, as a result, achieves a more accurate model for inferring the clean label.

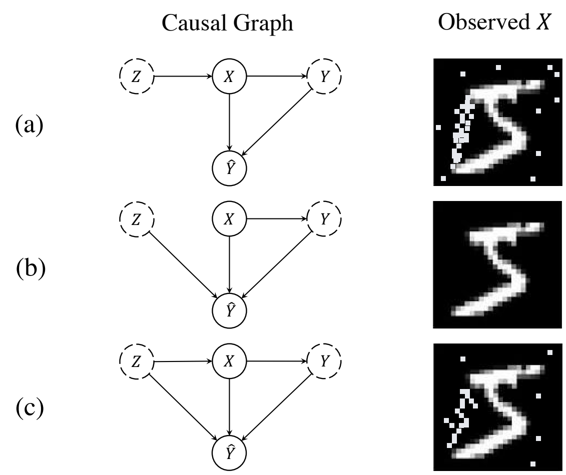


Figure 1: Examples of three causal graphs illustrating the mislabeling of "5" as "6" in MNIST, where X denotes instance(image), Y denotes the ground truth label, $\hat{Y}$ denotes the noisy label, and Z denotes the latent variable. The dashed circles represent the unobservable variable. (a) The instance is perturbed by noise, making "5" looks like "6". (b) The instance is clean, but it is mislabeled by an annotator. (c) The instance exhibits a mixture situation of (a) and (b).

Introduction

The success of deep neural networks is heavily based on large-scale annotation datasets (Daniely and Granot 2019; Yao et al. 2020a). However, data annotation inevitably introduces label noise, and the models are prone to overfitting on the mislabeled data, which hampers their performance in terms of generalization. Cleaning up corrupted labels is extremely expensive and time-consuming. Therefore, finding effective algorithms to mitigate the impact of label noise and improve model performance is of utmost importance.

Most of the main research (Patrini et al. 2017; Vahdat 2017; Xiao et al. 2015; Cheng et al. 2020; Yao et al. 2021; Bae et al. 2022; Zhang et al. 2024) in label-noise learning focuses on estimating the transition matrix $T_{ij} = P(\hat{Y} = j|Y = i, X = x)$, which captures the relationship between the true (clean) labels Y and the observed (noisy) labels $\hat{Y}$. This transition matrix provides valuable information about the noise distribution and can be used to guide the learning process. Typically, the identification of the transition matrix becomes a challenge due to the inaccessibility of the clean label Y for direct observation. In order to facilitate learning in the presence of noise, one strand of methods (Natarajan et al. 2013; Berthon et al. 2021; Xia et al. 2020b) assumes that the transition relationship between the noisy label $\hat{Y}$ and the clean label Y is **instant-independent**, wherein $P(\hat{Y}|Y, X) = P(\hat{Y}|Y)$. However, in complex real-world scenarios, label noise can be attributed to various factors and may violate the instance-independent assumption.

*These authors contributed equally.
†Corresponding Author

Therefore, it is crucial to find a more rational and effective transition relationship to handle label noise in realistic and complex settings.

To explore the factors that hinder the inference of the clean label, we approach our study from a causal perspective. Depending on the source of noise, we can get three distinct causal graphs of the data generation procedure, which are depicted in Figure 1. Firstly, in scenario (a), certain environmental factors Z, *e.g.*lighting, noise, shadow, impact the instance X. This, in turn, influences the annotators, leading to the production of a noisy label $\hat{Y}$. Secondly, in scenario (b), some factors Z do not influence X but directly cause the generation of the noisy label $\hat{Y}$, such as the annotator's negligence. Lastly, both scenarios (a) and (b) can occur simultaneously, resulting in a noisy label $\hat{Y}$. In all the cases shown in the figure, fulfilling the instance-independent assumption is challenging. It is evident that X serves as a common cause for both Y and $\hat{Y}$, introducing confounding bias when attempting to estimate the transition matrix $P(\hat{Y}|Y, X)$. Furthermore, the presence of an unobservable latent variable Z further complicates the estimation process, resulting in inaccuracies in learning the clean labels.

To address these challenges and improve the understanding of the transition relationship, we have integrated the three causal graphs presented in Figure 1 into a unified diagram, shown in Figure 1(c), and we introduce a new causal graph in Figure 2(a). In this proposed causal graph, we consider Y as the treatment and $\hat{Y}$ as the outcome and introduce a novel concept: the "causal transition matrix" $P(\hat{Y}|do(Y), X)$, which serves to approximate the true transition matrix. The proposed causal graph suggests the presence of two components within X: a **noise-resistant component** X_1 and a **noise-sensitive component** X_2. The noise-resistant component X_1 is not influenced by the latent variable Z but plays a role in determining the clean label Y. In contrast, the noise-sensitive component X_2 is subject to interaction with both the latent variable Z and the clean label Y, contributing to the generation of the noisy label $\hat{Y}$. The clear distinction of these two components enables the "causal transition matrix" $P(\hat{Y}|do(Y), X)$ to be identified with a theoretical guarantee, which in turn contributes to the inference of the clean label Y from X.

Building upon this, we have developed a novel training framework that explicitly models the variables and causal relationships within the proposed causal graph. Initially, the framework separates the noise-resistant component X_1 and the noise-sensitive component X_2 from the variable X. Given that the variable Y is unobservable, we use confidence sampling techniques to make it partially observable. In addition, we have designed a transition model to capture the relationships among various variables effectively. Our design ensures that the causal relationship is accurately established, leading to more robust and precise learning outcomes in the presence of label noise.

Our contribution can be summarized as follows:

(1) We propose a novel causal graph that integrates multiple scenarios of label-noise learning, thereby facilitating a more comprehensive understanding of the transition dynamics within this field.

(2) Based on the causal graph, we relax the instance-independent assumption and introduce the concept of the "causal transition matrix" $P(\hat{Y}|do(Y), X)$, which serves as an approximation for the transition matrix and provides a theoretical guarantee for its identification.

(3) We propose a novel framework that explicitly models the variables and their causal relationships as delineated in the proposed causal graph. This framework can be trained end-to-end, leading to a more precise model to infer clean labels.

(4) Experiments on both synthetic and real-world label-noise datasets highlight the superiority of our method.

Related Work

Due to the susceptibility of the traditional cross entropy (CE) loss to overfit noisy labels, and the high cost and time involved in cleaning up corrupted labels, there has been a notable surge in interest in learning with noisy labels.

Transition matrix in label-noise learning. The transition matrix plays a crucial role in the learning of label noise as it captures the relationships between true labels and observed noisy labels. One line of approaches (Patrini et al. 2017; Vahdat 2017; Xiao et al. 2015; Cheng et al. 2020) involves developing algorithms that initially estimate the transition matrix roughly and then correct it using posteriors. However, recent studies (Cheng et al. 2020; Xia et al. 2019; Yao et al. 2020b) have revealed that the transition matrix is unidentifiable and thus challenging to learn. To overcome this challenge, researchers (Natarajan et al. 2013; Berthon et al. 2021; Xia et al. 2020b) often assume that the noisy label is independent of the instance, denoted $P(\hat{Y}|Y, X) = P(\hat{Y}|Y)$. Additionally, some approaches (Zhang and Sabuncu 2018; Wang et al. 2019; Yao et al. 2020a; Ma et al. 2020) focus on designing robust loss functions that mitigate the label noise issue without the need for estimating the transition matrix. However, as pointed out by Zhang, Niu, and Sugiyama (2021), these methods may only perform adequately under simple conditions, such as symmetric noise, and may struggle when dealing with heavy and complex noise. Another line of research (Han et al. 2018, 2020; Yu et al. 2019; Mirzasoleiman, Cao, and Leskovec 2020; Wu et al. 2020) involves performing sample selection during training, some of which can also eliminate the instance-independent assumption. These methods utilize techniques such as learning with rejection (El-Yaniv et al. 2010; Thulasidasan et al. 2019; Mozannar and Sontag 2020; Charoenphakdee et al. 2021), meta-learning (Shu et al. 2019; Li et al. 2019), contrastive learning (Li et al. 2022), or semi-supervised learning (Nguyen et al. 2019; Li, Socher, and Hoi 2020) to select training samples with high confidence. However, as argued by Cheng et al. (2020), these methods cannot guarantee the avoidance of overfitting to label noise because they are trained based on CE loss.

Causal methods for label-noise learning. The causal perspective serves as a valuable tool to improve our understanding of the noise generation process. Yao et al. (2021) propose CausalNL which uses variational inference (Kingma and Welling 2013) to model the causal structure underlying the generation of noisy data. Likewise, Bae et al. (2022) propose a noisy prediction calibration method based on the causal structure, which also adopts the generative model to explicitly model the relationship between the output of a classifier and the true label. However, both rely on the assumption of instance independence in the posterior probability, $P(Y|\hat{Y}, X) = P(Y|\hat{Y})$, and model such a relation via the variational auto-encoder (Kingma and Welling 2013).

Compared with the other causal-based methods. To establish a more accurate and reliable transition relationship, we propose a novel causal structure and framework for learning with noise. This approach does not require the use of variational inference tools and eliminates the need for the posterior probability assumption. Within this structure, we introduce the concept of "causal transition matrix", which approximates the true transition matrix with theoretical guarantees, thus enhancing the efficacy of learning under noise conditions.

Method

Problem Setup

Considering a classification problem with k classes. Let X and Y be the random variables for the input instance and the clean label Y, respectively. The training examples (x_n, y_n) are sampled from the joint probability distribution P over the random variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, where $n \in \{1, 2, ..., N\}$ and N represent the number of training examples. The goal of the classification task is to find a classifier $f : \mathcal{X} \rightarrow \mathcal{Y}$ that accurately maps X to Y. Mainstream methods train the neural network by minimizing the empirical risk $\mathbb{E}_P[\mathcal{L}(f(x), y)]$, where $\mathcal{L}$ denotes the loss function. In real-world applications, we can only observe the sample $(x_n, \hat{y}_n)$, where $\hat{y}_n$ represents the noisy label. The noise of the label of each instance is characterized by a transition matrix $T(X)$, where each element $T_{ij}(X) = P(\hat{Y} = j|Y = i, X = x)$ denotes a probability from Y to $\hat{Y}$ when given X. When learning with noisy labels, the loss function becomes $\mathcal{L}(f(x), \hat{y})$, but our objective is still to train a clean classifier that minimizes the empirical risk on the clean label y.

Causal Viewpoint for Denoising

We have summarized the causal relationship in Figure 2(a), which involves two unobservable variables: Z and Y. In recent literature (Han et al. 2018, 2020; Yu et al. 2019; Yao et al. 2021; Bae et al. 2022), $\hat{Y}$ from reliable examples can be considered the clean label, which makes Y partially observed. Consequently, the unobservable latent variable Z becomes the primary obstacle. Mainstream studies (Cheng et al. 2020; Xia et al. 2019; Yao et al. 2020b) typically deem that the transition matrix $T(X)$ is not identifiable and difficult to learn. We introduce the concept of "causal transition matrix", which is defined as $T_{cau} = P(\hat{Y}|do(Y), X)$. This represents the outcome distribution of the noisy label $\hat{Y}$ by intervening on Y conditioned on X. By intervening on Y, we essentially eliminate the incoming edge of Y, thus ensuring unbiased estimation.

In causal theory, a natural approach to estimate $T_{cau} = P(\hat{Y}|do(Y), X)$ is backdoor adjustment (Pearl 1995), since X effectively blocks all backdoor paths. However, accurately computing an unbiased relationship necessitates sampling across X, which introduces substantial computational challenges. Moreover, computing over X is more "noisy", as it involves a lot of irrelevant information. Consequently, an accurate and efficient alternative approach is necessary to investigate.

To make the causal transition matrix identifiable, we separate each instance X into two components, the noise-resistant component X_1 and the noise-sensitive component X_2. We posit that X_1 is not influenced by the latent variable Z and directly causes the clean label Y, while X_2 may be affected by Z and serves as a contributing factor to the noise found in $\hat{Y}$. Based on the two components, we have the following theorem.

Theorem 1 *The instance-dependent causal transition matrix $P(\hat{Y}|do(Y), X)$ is identifiable if we recover the noise predictive factor X_2.*

Proof: Let $G_{\overline{Y}}$ be the graph induced by removing the incoming edges of Y. Since $\hat{Y} \perp\!\!\!\perp X_1|Y, X_2$ in $G_{\overline{Y}}$, we have $P(\hat{Y}|do(Y), X) = P(\hat{Y}|do(Y), X_1, X_2) = P(\hat{Y}|do(Y), X_2)$. Let $G_{\underline{Y}}$ be the graph induced by removing the outgoing edges of Y. Since $\hat{Y} \perp\!\!\!\perp Y|X_2$ in $G_{\underline{Y}}$, we have $P(\hat{Y}|do(Y), X_2) = P(\hat{Y}|Y, X_2)$. However, note that Y is an unobservable latent variable that we are interested in modeling, and as such we need causal estimand that gives us an unbiased estimation of Y as well.

Theorem 2 *The effect of X_1 on Y can be identified if we recover X_1.*

Proof: Since we have $P(Y|do(X_1)) = \int_{X_2} P(Y|X_1, X_2)P(X_2)dX_2$ by using the backdoor criterion and that $Y \perp\!\!\!\perp X_2|X_1$ by d-separation, we have

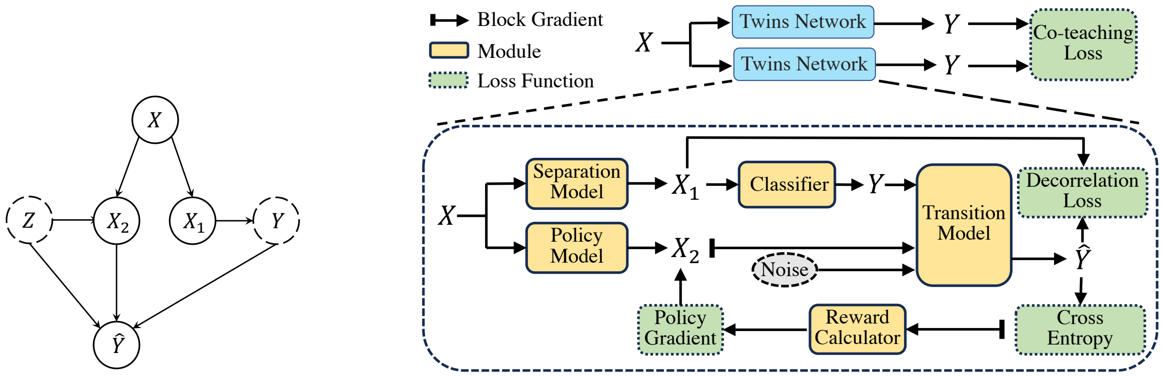
$$\int_{X_2} P(Y|X_1, X_2)P(X_2)dX_2 = \int_{X_2} P(Y|X_1)P(X_2)dX_2$$
$$= P(Y|X_1)\int_{X_2} P(X_2)dX_2 = P(Y|X_1).$$

Based on Theorem 2, it is possible to obtain an unbiased classifier based solely on X_1. Consequently, we can recover X_1 by decorrelating it from Z. According to Theorem 1, the causal transition matrix can be identified if we can identify the contributing factor $X_2 \subseteq X$ to $\hat{Y}$. Intuitively, X_1 can be omitted, since it is the parent of Y and the do operation effectively eliminates the incoming edge of Y.

Therefore, the focus of this paper is twofold: 1) **how to identify these contributing factors** and 2) **how to model causal relations between variables**.

Training Framework for Denoising

The training framework of our method is illustrated in Figure 2. In this subsection, we will provide a detailed introduction to the entire procedure.



(a) (b)

Figure 2: (a) The proposed causal graph for learning with noisy labels. (b) The training framework of our method.

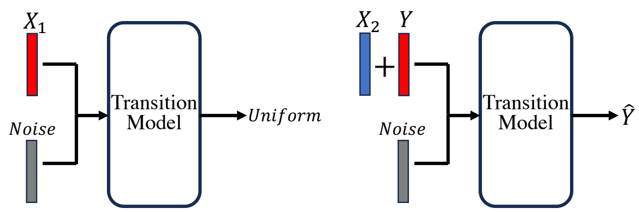


Figure 3: The design of the transition model.

Modeling the observable contributing factors. Initially, we utilize a separation model to separate the noise-resistant component X_1 from X, which can be represented as $X_1 = g_1(X)$, where $g_1(\cdot)$ denotes any neural network. Subsequently, we infer the clean label Y using a classifier $Y = f(X_1)$. As mentioned above, Y is unobservable in the causal graph, so it is necessary to sample confidently Y to make it partially observable. To accomplish this, we employ co-teaching (Han et al. 2018), as it is a commonly used and straightforward technique. Two twin models are trained simultaneously and updated via:

$$\mathcal{L}_{\text{co-teaching}} = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \min(\ell_{1,i}, \ell_{2,i}), \tag{1}$$

where $\mathcal{B}$ denotes the confident examples in a batch with number $|\mathcal{B}|$, and $\ell_{j,i}$ denotes the loss of i-th example in j-th model.

Since we do not know how the instance X is affected by the latent variable Z, we directly model the noise-sensitive component X_2 as a representation through $X_2 = g_2(X)$, where $g_2(\cdot)$ denotes any type of neural network.

Modeling the causal transition relations. To ensure that the variables obey the proposed causal graph, there are two things that must be done.

(i) Decorrelating X_1 from $\hat{Y}$.

(ii) Modeling the relationship from X_2, Y, Z to $\hat{Y}$.

We propose a transition model to achieve both goals simultaneously, as depicted in Figure 3, in which the blue block indicates that the gradients do not need to be computed.

Firstly, for decorrelating X_1 and $\hat{Y}$, the transition model f_{tran} take X_1 and a Gaussian noise as inputs and output a tensor with dimension of k, which can be represented as:

$$\hat{Y}_{X_1} = f_{tran}(X_1, Z_1),\ Z_1 \sim \mathcal{N}(0,1),\ \hat{Y}_{X_1} \in \mathbb{R}^k. \tag{2}$$

We aim to ensure that the probability of $\hat{Y}$ given X_1 is uniform for each class candidate. To achieve this, we constrain $\hat{Y}_{X_1}$ so that X_1 of each instance predicts an all-one vector. This can be represented as the decorrelation loss:

$$\mathbf{Reg}(X_1) = -\sum_{i=1}^{N} \tilde{y}_{x_1} \text{log\_softmax}(\mathbf{1}_k), \tag{3}$$

where N denotes the number of the examples in a batch, $\tilde{y}_{x_1}$ denotes the output of the transition model when take x_1 as input, and $\mathbf{1}_k$ denotes all-one vector with k dimension.

Secondly, for modeling the causal transition matrix, we take Y and another noise as input, affiliated with the noise-sensitive component X_2, which is represented as:

$$\hat{Y} = f_{tran}(m(Y, X_2), Z_2),\ Z_2 \sim \mathcal{N}(0,1),\ \hat{Y} \in \mathbb{R}^k, \tag{4}$$

where $m(\cdot,\cdot)$ denotes the merging function of two tensors. In this paper, we adopt the merging function:

$$m(Y, X_2) = gs(Y + \beta * X_2), \tag{5}$$

where β denotes the scaling factor, and $gs(\cdot)$ denotes the Gumbel-Softmax function.

Since our main focus is on the causal transition matrix $P(\hat{Y}|do(Y), X)$, we block the gradient of X_2 to make the relationship from Y to $\hat{Y}$ more precise. Therefore, X_2 can be considered as the compensation tensor that bridges the gap between $P(\hat{Y}|do(Y), X)$ and $P(\hat{Y}|do(Y))$. The transition model can be optimized using the following equation:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N} \tilde{y}_i \log(\hat{y}_i), \tag{6}$$

where $\tilde{y}$ denotes the predicted noisy label, and $\hat{y}$ denotes the noisy label.

Policy model for getting the noise-sensitive component. We still need to optimize g_2 to minimize the cross-entropy between the model prediction $\tilde{Y}$ and the noisy label $\hat{Y}$. However, since we have blocked the gradient of X_2 in the transition model, we need to find an alternative approach. Drawing inspiration from reinforcement learning, we use the policy gradient (Sutton and Barto 2018) to obtain the noise-sensitive component X_2. Specifically, we model g_2 using a policy π that takes the instance X as input state and outputs a soft action X_2. After the soft action is performed in the transition model, a reward R is obtained, representing the outcome of the soft action. We design a reward calculator, which computes the reward for each instance x_i as $R = \frac{1}{1-\tilde{y}_i \log(\hat{y}_i)}$. This reward function ensures that the reward is always greater than zero and increases as the cross-entropy decreases. This naturally promotes the goal of X_2 to be achieved. The policy model is trained using the following loss function:

$$\mathcal{L}_{pg} = -\sum_{i=1}^{N} R \log \pi(x_i). \tag{7}$$

Recalling the Gumbel-Softmax function in equation. (5), it serves two main functions: (1) It acts as an exploration strategy for π and prevents the policy from settling into sub-optimal results. (2) It introduces more noise into the transition matrix to mimic the effect of Z.

Training objectives. Our framework can be trained end-to-end, and the overall loss function $\mathcal{L}$ is computed as:

$$\mathcal{L} = \mathcal{L}_{\text{co-teaching}} + \alpha_1 * \mathcal{L}_{CE} + \alpha_2 * \mathcal{L}_{PG} + \alpha_3 * \mathbf{Reg}(X_1), \tag{8}$$

where α_1, α_2 and α_3 denotes the scale factors.

Analysis of the Framework

What our framework does? Theories 1 and 2 assert that under the proposed causal graph, the causal transition matrix is identifiable and enables us to infer the clean label Y during the inference process. Our framework establishes the causal relationship through Equation (8), which comprises four terms of loss functions. The co-teaching loss $\mathcal{L}_{\text{co-teaching}}$ treats the confident noisy label $\hat{Y}$ as the true label Y and guides the training of the classifiers within the twin networks. Concurrently, the policy gradient loss $\mathcal{L}_{PG}$ is instrumental in separating the noise-sensitive component X_2 from the input instance X. Additionally, the regularization loss $\mathbf{Reg}(X_1)$ acts as a penalty to reduce the correlation between X_1 and $\hat{Y}$, ensuring that the noise-resistant component X_1 remains insensitive to noise. Our proposed causal graph accommodates the linkage between the two variables X_1 and X_2, obviating the need for additional computations to disentangle them. Once the variables are distinguished, the causal transition relationship is explicitly established through the transition model, which is refined by optimizing the cross-entropy loss $\mathcal{L}_{CE}$.

Additional computational costs. During the training process, we construct a twin network architecture, which requires twice the computational resources compared to the original classifier. However, for inference, only the separation model and the classifier are employed. Since the separation model can be implemented as a relatively simple layer, the increase in computational costs remains minimal. Furthermore, our method does not adopt generative models, unlike previous work (Yao et al. 2021; Bae et al. 2022; Cheng et al. 2020), which not only saves computational costs but also eliminates potential errors.

Relation to mainstream assumptions. Our method is adept at addressing both instance-independent and instance-dependent label noise. By focusing solely on the right part of the causal graph and disregarding X_2, our approach aligns with the **instance-independent assumption** that $P(\hat{Y}|Y,X) = P(\hat{Y}|Y)$. This equivalence arises due to the decorrelation between X_1 and $\hat{Y}$. Furthermore, our method does not rely on the **assumption of independence in the posterior probability** $P(Y|\hat{Y},X) = P(Y|\hat{Y})$, as seen in related work (Yao et al. 2021; Bae et al. 2022). Although most related work depends on specific assumptions, our approach is based on two relatively mild and broadly applicable assumptions: the separability of X and the partial observability of Y. These assumptions are generally reasonable and pragmatic for many real-world applications.

Flexibility of the causal graph. The proposed causal graph can also be adapted for semi-supervised methods (Wang et al. 2022; Xiao et al. 2023; Li, Socher, and Hoi 2020), which initially select clean samples and then use the remaining noisy samples to enhance performance. For further details on the implementation and experimental results, please refer to Li et al. (2024).

Experiment

We carry out experiments on both **synthetic** and **real-world datasets**, encompassing various types of label noise. We place particular emphasis on instance-dependent noise, as it represents the most challenging and significant aspect of our research.

Experiment Setup

Datasets. (i) We first perform experiments on the manually corrupted version of four **synthetic datasets**, *i.e.*, FashionMNIST (Xiao, Rasul, and Vollgraf 2017), SVHN (Yuval 2011), CIFAR10, CIFAR100 (Krizhevsky, Hinton et al. 2009). The experiments are conducted with three different types of artificial label noise, with a focus on instance-dependent noise in this paper. (a) **Symmetric Noise (SYM)**. The label of each instance is uniformly flipped to one of the other classes (Patrini et al. 2017; Han et al. 2018; Xia et al. 2020a). (b) **Asymmetric Noise (AYM)**. The label of each instance is flipped to a set of semantically similar classes (Tanaka et al. 2018; Han et al. 2018; Xia et al. 2020a). (c) **Instance Dependent Noise (IDN)**. The label of each instance is determined by the probability that it is mislabeled, and this probability is calculated based on the corresponding feature of the data instance (Berthon et al. 2021; Yao et al. 2021; Bae et al. 2022).

(ii) Meanwhile, we perform experiments on two **real-world** noisy datasets: Food101 (Bossard, Guillaumin, and Van Gool 2014) and Clothing1M (Xiao et al. 2015).

Evaluation metric. The test sets for the various datasets remain clean, ensuring that the **test accuracy** can effec-

Method	SYM		ASYM		IDN	
	20%	80%	20%	40%	20%	40%
CE	74.0	27.0	81.0	77.3	68.4	52.1
Early Stop	83.6	49.5	84.1	76.6	79.5	55.4
Co-teaching	82.5	64.2	88.2	73.6	81.8	75.4
Joint	82.0	6.0	82.1	82.3	82.7	82.4
JoCoR	86.0	27.6	88.9	79.4	86.3	83.2
CORES2	74.6	8.9	77.6	74.3	80.0	58.1
SCE	74.0	27.0	82.0	77.4	68.3	52.0
LS	73.9	27.8	81.5	77.0	69.0	52.5
REL	84.6	70.1	82.8	76.2	84.6	75.5
Forward	77.4	24.3	88.3	79.2	75.2	56.9
DualT	84.5	10.0	86.9	83.1	85.1	68.5
TVR	72.6	24.9	80.6	76.4	66.3	51.7
CausalNL	84.0	51.5	88.8	87.4	90.8	90.0
Ours w.o/ pg	**92.1**	**75.8**	89.6	82.4	**91.3**	90.3
Ours	**92.1**	71.5	**91.4**	**88.7**	91.0	**90.4**

Table 1: Results on FashionMNIST with symmetric, asymmetric, and instance-dependent label noise.

tively reflect the superior performance of the denoising methods we evaluate. For the synthetic datasets, we report the mean performance across 5 random seeds. For the real-world dataset, we report the best results for the last 10 epochs.

Baselines. We choose popular denoising methods as baselines: Co-teaching (Han et al. 2018), Joint (Tanaka et al. 2018), JoCoR (Wei et al. 2020), CORES2 (Cheng et al. 2020), SCE (Wang et al. 2019), LS (Lukasik et al. 2020), REL (Xia et al. 2020a), Forward (Patrini et al. 2017), DualT (Yao et al. 2020b), TVR (Zhang, Niu, and Sugiyama 2021), MentorNet (Jiang et al. 2018), Mixup (Zhang et al. 2018), Reweight (Liu and Tao 2015), T-Revision (Xia et al. 2019), BLTM-V (Yang et al. 2022), CausalNL (Yao et al. 2021). Among them, CausalNL is closest to our work, since it is also a causality-based method.

Ablation study. We also performed ablations on all scenarios by removing the policy model X_2 from the framework. This indicates that our model is trained with a transition relationship that follows the instance-independent assumption $P(\hat{Y}|Y,X) = P(\hat{Y}|Y)$. We represent this as “w.o/pg” in this section.

Implementation details. We use ResNet models (He et al. 2016) as backbone classifiers, selecting network depth based on the difficulty of the dataset. For the FashionMNIST dataset, we utilize ResNet18. ResNet34 is employed for both SVHN and CIFAR10 datasets. ResNet50 is applied to CIFAR100 without using pretrained checkpoints, while ResNet50 with pretrained checkpoints is used for Food101 and Clothing1M datasets. In every scenario, we assign $\alpha_1 = \alpha_3 = 0.1$. The value for α_2 varies according to scenario complexity. For datasets lacking extra perturbations on the instances, α_2 is set to 0.1. Conversely, when datasets contain perturbations, α_2 is adjusted to 0.01. Moreover, β is consistently set to 0.2 across all experiments.

Map	IDN				
	20%	30%	40%	45%	50%
CE	91.51±0.45	91.21±0.43	87.87±1.12	67.15±1.65	51.01±3.62
Co-teaching	93.93±0.31	92.06±0.31	91.93±0.81	89.33±0.71	67.62±1.99
Decoupling	90.02±0.25	91.59±0.25	88.27±0.42	84.57±0.89	65.14±2.79
MentorNet	94.08±0.12	92.73±0.37	90.41±0.49	87.45±0.75	61.23±2.82
Mixup	89.73±0.37	90.02±0.35	85.47±0.63	82.41±0.62	68.95±2.58
Forward	91.89±0.31	91.59±0.23	89.33±0.53	80.15±1.91	62.53±3.35
Reweight	92.44±0.34	92.32±0.51	91.31±0.67	85.93±0.84	64.13±3.75
T-Revision	93.14±0.53	93.51±0.74	92.65±0.76	88.54±1.58	64.51±3.42
BLTM-V	**95.12±0.40**	**94.69±0.24**	88.13±3.23	80.43±4.12	78.71±4.37
CausalNL	94.06±0.23	93.86±0.65	93.82±0.64	93.19±0.93	85.41±2.95
Ours w.o/ pg	93.86±0.17	93.82±0.18	93.50±0.25	93.19±1.25	**92.91±1.54**
Ours	94.13±0.08	93.97±0.11	**93.94±0.16**	**93.33±1.12**	92.57±1.56

Table 2: Results on SVNH with instance-dependent noise.

Results on Symmetric and Asymmetric Noise

We begin our experiments with the relatively simple FashionMNIST dataset, considering both symmetric and asymmetric noise. The results for different noise rates are shown in Table 1. For symmetric noise, we performed experiments with noise rates of 20% and 80%. For asymmetric noise, we conducted experiments with noise rates of 20% and 40%. In both scenarios, our ablation study achieves the best performance. The reason behind this is that the noisy labels are randomly assigned without considering the instance, making them almost instance-independent, where $P(\hat{Y}|Y,X) = P(\hat{Y}|Y)$. The policy model can be considered as $P(\hat{Y}|Y+\mathcal{R})$, where a real number $\mathcal{R}$ is added to Y, thus affecting the model's performance.

Results on Instance-dependent Noise

Comparison with baselines. Instance-dependent noise is crucial for evaluating the accuracy of the estimated transition matrix, as it is challenging to satisfy the instance-independent assumption. We conducted our study on four datasets: FashionMNIST (Table 1), SVHN (Table 2), CIFAR10 and CIFAR100 (Table 3), with increasing levels of difficulty. As the results demonstrate, our method achieves state-of-the-art performance in almost all scenarios, particularly when faced with high noise rates. Even the ablation study of our method yields satisfactory results. This can be attributed to the separation of the X_1 component from X, which allows explicit decorrelation between X_1 and the noisy label $\hat{Y}$, resulting in a more reliable classifier.

How the latent variable Z influences performance. In real world scenarios, the latent variable Z influences both the instance X and the noisy label $\hat{Y}$. In this subsection, we introduce additional noise to perturb the instance X, reflecting its influence along the path $Z \rightarrow X$ in the causal graph. Meanwhile, the noise rate represents the influence of Z along the path $Z \rightarrow \hat{Y}$. To investigate how the latent variable Z influences performance, we control the influence of Z on one path while adjusting the other.

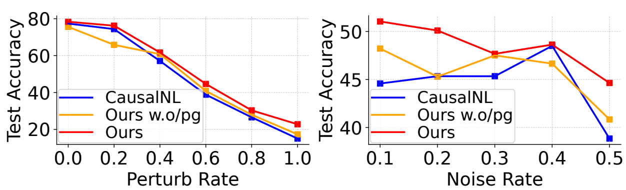
Construction of the perturbed instance. We introduce noise to each instance X in the training data by generating

Map	CIFAR10-IDN					CIFAR100-IDN				
	20%	30%	40%	45%	50%	20%	30%	40%	45%	50%
CE	75.81±0.26	69.15±0.65	62.45±0.86	51.72±1.34	39.42±2.52	30.42±0.44	24.15±0.78	21.45±0.70	15.23±1.32	14.42±2.21
Co-teaching	80.96±0.31	78.56±0.61	73.41±0.78	71.60±0.79	45.92±2.21	37.96±0.53	33.43±0.74	28.04±1.43	25.60±0.93	23.97±1.91
Decoupling	78.71±0.15	75.17±0.58	61.73±0.34	58.61±1.73	50.43±2.19	36.53±0.49	30.93±0.88	27.85±0.91	23.81±1.31	19.59±2.12
MentorNet	81.03±0.24	77.22±0.47	71.83±0.49	66.18±0.64	47.89±2.03	38.91±0.54	34.23±0.73	31.89±1.19	27.53±1.23	24.15±2.31
Mixup	73.17±0.34	70.02±0.31	61.56±0.71	56.45±0.67	48.95±2.58	32.92±0.76	29.76±0.87	25.92±1.26	23.13±2.15	21.31±1.32
Forward	76.64±0.26	69.75±0.56	60.21±0.75	48.81±2.59	46.27±1.30	36.38±0.92	33.17±0.73	26.75±0.93	21.93±1.29	19.27±2.11
Reweight	76.23±0.25	70.12±0.72	62.58±0.46	51.54±0.92	45.46±2.56	36.73±0.72	31.91±0.91	28.39±1.46	24.12±1.41	20.23±1.23
T-Revision	76.15±0.37	70.36 ±0.54	64.09±0.37	54.42±1.01	49.02±2.13	37.24±0.85	36.54±0.79	27.23±1.13	25.53±1.94	22.54±1.95
BLTM-V[1]	80.37±1.98	78.82±1.07	72.93±4.00	64.83±4.65	60.33±5.29	-	-	-	-	-
CausalNL	81.47±0.32	80.38±0.44	77.53±0.45	78.60±1.06	77.39±1.24	41.47±0.43	40.98±0.62	34.02±0.95	33.34±1.13	32.13±2.23
Ours w.o/ pg	82.57±0.33	81.24±0.36	79.36±0.81	78.43±0.51	75.59±2.07	45.50±0.99	**44.67±0.60**	38.44±1.40	34.88±2.53	33.05±1.47
Ours	**82.94±0.29**	**82.15±0.25**	**81.04±0.23**	**80.24±0.39**	**78.37±0.93**	**46.37±0.46**	43.34±0.39	**39.61±1.04**	**37.04±1.83**	**34.44±1.86**

Table 3: Results on CIFAR dataset.

Method	Dataset	
	Food101	Clothing1M
CE	78.37	68.88
Early Stop	73.22	67.07
Co-teaching	78.35	60.15
SCE	75.23	67.77
REL	78.96	62.53
Forward	83.76	69.91
DualT	57.46	70.18
TVR	77.37	69.44
CausalNL	85.64	68.90
Ours w.o/ pg	85.52	70.48
Ours	**85.86**	**72.25**

Table 4: Results on real-world dataset.



(a) Perturbing the training instances considering different intensities with a 50% instance-dependent noise.

(b) Considering different levels of instance-dependent noise when perturbing the instances with an intensity of 0.6.

Figure 4: Model performance on the CIFAR10 Dataset.

a noise vector Z_{noise} from a uniform distribution, $Z_{\text{noise}} \sim U([0,1])$, where $Z_{\text{noise}} \in \mathbb{R}^D$, and D denotes the dimensionality of instance X. The perturbed instance X_{per} is generated by: $X_{\text{per}} = X + \gamma \cdot Z_{\text{noise}}$, where γ is the scale factor that determines the intensity of the perturbation.

How the performance decreases. Figure 4 (a) shows the performance curve as we gradually increase the intensity of the perturbation γ from 0 to 1 on the CIFAR10 dataset with a 50% instance-dependent noisy label rate. As the perturbation increases, the performance of all methods decreases, but our method consistently exhibits the best performance. Figure 4 (b) illustrates the performance curve as we increase the noisy label rate from 0.1 to 0.5 while keeping γ fixed at 0.6. Compared to CausalNL and the ablation study, our method demonstrates not only higher effectiveness but also greater stability. In particular, the performance of both baselines experiences a significant drop when the label noise rate increases from 0.4 to 0.5, while our method is relatively unaffected. This can be attributed to our modeling of the noise-sensitive component X_2, which helps bridge the gap between $P(\hat{Y}|Y)$ and $P(\hat{Y}|Y,X)$ in certain contexts.

Results on Real-world Noisy Dataset

We also performed experiments on two popular noisy real-world label datasets, Food101 and Clothing1M, to evaluate the superiority of our method. Table 4 presents the test accuracy on the clean test set. Our method outperforms the baselines, demonstrating its effectiveness in real-world settings. Even in the ablation study, our method without the policy model still achieves competitive performance compared to the baseline. This is mainly due to our explicit separation of the component X_1 from the original instance X, which is not influenced by the latent variable X and directly contributes to the learning of the classifier. The ablation results also indicate that the label noise in Clothing1M is more likely to be instance-dependent compared to Food101.

Conclusion

In this paper, our objective is to learn an accurate transition relationship in label-noise learning to obtain a better classifier. To gain a deeper understanding of the label noise generation procedure, we approach the problem from a causal viewpoint and propose a novel causal structure for learning with label noise. Within this causal structure, we introduce the concept of a "causal transition matrix" $P(\hat{Y}|do(Y),X)$, which can approximate the original transition matrix $P(\hat{Y}|Y,X)$ without relying on the instance-independent assumption. To accomplish this, we developed a framework that enables us to estimate the causal transition matrix with a theoretical guarantee of identifiability. Experimental results on both synthetic and real-world datasets validated the superior performance of our method.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62376243, 62441605, 62037001), the Starry Night Science Fund at Shanghai Institute for Advanced Study (Zhejiang University), and Ant Group Postdoctoral Programme. Long Chen was supported by RGC Early Career Scheme (26208924), National Natural Science Foundation of China Young Scientist Fund (NSFC24EG42), and HKUST Sports Science and Technology Research Grant (SSTRG24EG04).

References

Bae, H.; Shin, S.; Na, B.; Jang, J.; Song, K.; and Moon, I.-C. 2022. From noisy prediction to true label: Noisy prediction calibration via generative model. In *International Conference on Machine Learning*, 1277–1297. PMLR.

Berthon, A.; Han, B.; Niu, G.; Liu, T.; and Sugiyama, M. 2021. Confidence scores make instance-dependent label-noise learning possible. In *International conference on machine learning*, 825–836. PMLR.

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, 446–461. Springer.

Charoenphakdee, N.; Cui, Z.; Zhang, Y.; and Sugiyama, M. 2021. Classification with rejection based on cost-sensitive classification. In *International Conference on Machine Learning*, 1507–1517. PMLR.

Cheng, J.; Liu, T.; Ramamohanarao, K.; and Tao, D. 2020. Learning with bounded instance and label-dependent label noise. In *International conference on machine learning*, 1789–1799. PMLR.

Daniely, A.; and Granot, E. 2019. Generalization bounds for neural networks via approximate description length. *Advances in Neural Information Processing Systems*, 32.

El-Yaniv, R.; et al. 2010. On the Foundations of Noise-free Selective Classification. *Journal of Machine Learning Research*, 11(5).

Han, B.; Niu, G.; Yu, X.; Yao, Q.; Xu, M.; Tsang, I.; and Sugiyama, M. 2020. Sigua: Forgetting may make learning with noisy labels more robust. In *International Conference on Machine Learning*, 4006–4016. PMLR.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Jiang, L.; Zhou, Z.; Leung, T.; Li, L.-J.; and Fei-Fei, L. 2018. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International conference on machine learning*, 2304–2313. PMLR.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

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

Li, J.; Chang, T.-W.; Kuang, K.; Li, X.; Chen, L.; and Zhou, J. 2024. Learning Causal Transition Matrix for Instance-dependent Label Noise. *arXiv preprint arxiv :2412.13516*.

Li, J.; Socher, R.; and Hoi, S. C. 2020. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*.

Li, J.; Wong, Y.; Zhao, Q.; and Kankanhalli, M. S. 2019. Learning to learn from noisy labeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5051–5059.

Li, S.; Xia, X.; Ge, S.; and Liu, T. 2022. Selective-supervised contrastive learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 316–325.

Liu, T.; and Tao, D. 2015. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3): 447–461.

Lukasik, M.; Bhojanapalli, S.; Menon, A.; and Kumar, S. 2020. Does label smoothing mitigate label noise? In *International Conference on Machine Learning*, 6448–6458. PMLR.

Ma, X.; Huang, H.; Wang, Y.; Romano, S.; Erfani, S.; and Bailey, J. 2020. Normalized loss functions for deep learning with noisy labels. In *International conference on machine learning*, 6543–6553. PMLR.

Mirzasoleiman, B.; Cao, K.; and Leskovec, J. 2020. Coresets for robust training of deep neural networks against noisy labels. *Advances in Neural Information Processing Systems*, 33: 11465–11477.

Mozannar, H.; and Sontag, D. 2020. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, 7076–7087. PMLR.

Natarajan, N.; Dhillon, I. S.; Ravikumar, P. K.; and Tewari, A. 2013. Learning with noisy labels. *Advances in neural information processing systems*, 26.

Nguyen, D. T.; Mummadi, C. K.; Ngo, T. P. N.; Nguyen, T. H. P.; Beggel, L.; and Brox, T. 2019. Self: Learning to filter noisy labels with self-ensembling. *arXiv preprint arXiv:1910.01842*.

Patrini, G.; Rozza, A.; Krishna Menon, A.; Nock, R.; and Qu, L. 2017. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1944–1952.

Pearl, J. 1995. Causal diagrams for empirical research. *Biometrika*, 82(4): 669–688.

Shu, J.; Xie, Q.; Yi, L.; Zhao, Q.; Zhou, S.; Xu, Z.; and Meng, D. 2019. Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in neural information processing systems*, 32.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Tanaka, D.; Ikami, D.; Yamasaki, T.; and Aizawa, K. 2018. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5552–5560.

Thulasidasan, S.; Bhattacharya, T.; Bilmes, J.; Chennupati, G.; and Mohd-Yusof, J. 2019. Combating label noise in deep learning using abstention. *arXiv preprint arXiv:1905.10964*.

Vahdat, A. 2017. Toward robustness against label noise in training deep discriminative neural networks. *Advances in neural information processing systems*, 30.

Wang, X.; Wu, Z.; Lian, L.; and Yu, S. X. 2022. Debiased learning from naturally imbalanced pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14647–14657.

Wang, Y.; Ma, X.; Chen, Z.; Luo, Y.; Yi, J.; and Bailey, J. 2019. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, 322–330.

Wei, H.; Feng, L.; Chen, X.; and An, B. 2020. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13726–13735.

Wu, P.; Zheng, S.; Goswami, M.; Metaxas, D.; and Chen, C. 2020. A topological filter for learning with label noise. *Advances in neural information processing systems*, 33: 21382–21393.

Xia, X.; Liu, T.; Han, B.; Gong, C.; Wang, N.; Ge, Z.; and Chang, Y. 2020a. Robust early-learning: Hindering the memorization of noisy labels. In *International conference on learning representations*.

Xia, X.; Liu, T.; Han, B.; Wang, N.; Gong, M.; Liu, H.; Niu, G.; Tao, D.; and Sugiyama, M. 2020b. Part-dependent label noise: Towards instance-dependent label noise. *Advances in Neural Information Processing Systems*, 33: 7597–7610.

Xia, X.; Liu, T.; Wang, N.; Han, B.; Gong, C.; Niu, G.; and Sugiyama, M. 2019. Are anchor points really indispensable in label-noise learning? *Advances in neural information processing systems*, 32.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Xiao, R.; Dong, Y.; Wang, H.; Feng, L.; Wu, R.; Chen, G.; and Zhao, J. 2023. ProMix: combating label noise via maximizing clean sample utility. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 4442–4450.

Xiao, T.; Xia, T.; Yang, Y.; Huang, C.; and Wang, X. 2015. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2691–2699.

Yang, S.; Yang, E.; Han, B.; Liu, Y.; Xu, M.; Niu, G.; and Liu, T. 2022. Estimating instance-dependent bayes-label transition matrix using a deep neural network. In *International Conference on Machine Learning*, 25302–25312. PMLR.

Yao, Q.; Yang, H.; Han, B.; Niu, G.; and Kwok, J. T.-Y. 2020a. Searching to exploit memorization effect in learning with noisy labels. In *International Conference on Machine Learning*, 10789–10798. PMLR.

Yao, Y.; Liu, T.; Gong, M.; Han, B.; Niu, G.; and Zhang, K. 2021. Instance-dependent label-noise learning under a structural causal model. *Advances in Neural Information Processing Systems*, 34: 4409–4420.

Yao, Y.; Liu, T.; Han, B.; Gong, M.; Deng, J.; Niu, G.; and Sugiyama, M. 2020b. Dual t: Reducing estimation error for transition matrix in label-noise learning. *Advances in neural information processing systems*, 33: 7260–7271.

Yu, X.; Han, B.; Yao, J.; Niu, G.; Tsang, I.; and Sugiyama, M. 2019. How does disagreement help generalization against label corruption? In *International Conference on Machine Learning*, 7164–7173. PMLR.

Yuval, N. 2011. Reading digits in natural images with unsupervised feature learning. In *Proceedings of the NIPS Workshop on Deep Learning and Unsupervised Feature Learning*.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*.

Zhang, R.; Cao, Z.; Yang, S.; Si, L.; Sun, H.; Xu, L.; and Sun, F. 2024. Cognition-Driven Structural Prior for Instance-Dependent Label Transition Matrix Estimation. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhang, Y.; Niu, G.; and Sugiyama, M. 2021. Learning noise transition matrix from only noisy labels via total variation regularization. In *International Conference on Machine Learning*, 12501–12512. PMLR.

Zhang, Z.; and Sabuncu, M. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31.