Directional Label Diffusion Model for Learning from Noisy Labels

Senyu Hou[1], Gaoxia Jiang[1], Jia Zhang[1], Shangrong Yang[1], Husheng Guo[1],
Yaqing Guo[1], Wenjian Wang[2*]
[1]School of Computer and Information Technology, Shanxi University, China
[2]Key Laboratory of Computational Intelligence and Chinese Information Processing
Ministry of Education, Shanxi University, China
{202312407005, 202412407035}@email.sxu.edu.cn
{jianggaoxia, sryang, guohusheng, yqguo, wjwang}@sxu.edu.cn

Abstract

In image classification, the label quality of training data critically influences model generalization, especially for deep neural networks (DNNs). Traditionally, learning from noisy labels (LNL) can improve the generalization of DNNs through complex architectures or a series of robust techniques, but its performance improvement is limited by the discriminative paradigm. Unlike traditional ways, we resolve the LNL problems from the perspective of robust label generation, based on diffusion models within the generative paradigm. To expand the diffusion model into a robust classifier that explicitly accommodates more noise knowledge, we propose a Directional Label Diffusion (DLD) model. It disentangles the diffusion process into two paths, i.e., directional diffusion and random diffusion. Specifically, directional diffusion simulates the corruption of true labels into a directed noise distribution, prioritizing the removal of likely noise, whereas random diffusion introduces inherent randomness to support label recovery. This architecture enable DLD to gradually infer labels from an initial random state, interpretably diverging from the specified noise distribution. To adapt the model to diverse noisy environments, we design a low-cost label precorrection method that automatically supplies more accurate label information to the diffusion model, without requiring manual intervention or additional iterations. Our approach outperforms state-of-the-art methods on both simulated and real-world noisy datasets. Code is available at https://github.com/SenyuHou/DLD.

1. Introduction

Deep neural networks (DNNs) have achieved notable success across key areas of machine learning, driven by the growth of datasets and the increase in model parameters. These models are primarily trained using supervised learning, which relies heavily on large volumes of labeled data [5]. However, real-world data often suffer from low-quality labels due to subjective errors like indistinguishable samples or objective errors such as incorrect label recording, making noisy labels a widespread issue [13, 55]. Studies have shown that the generalization capability of deep neural networks is adversely affected when trained with noisy labels [1, 59]. Consequently, it is crucial to develop methods for learning from noisy labels (LNL) on large-scale but low-quality datasets.

Due to the varied distribution of label noise, historical research has classified it into class-conditional noise (CCN) and instance-dependent noise (IDN) [57]. Recent studies indicate that label noise in real datasets predominantly aligns with the IDN assumption, wherein noise is unevenly distributed and more likely in samples with similar features [2, 6, 10–12, 51]. To tackle these complex noise patterns, mainstream LNL methods employ strategies combining filtering and training approaches. Some studies demonstrate a phenomenon known as the memorization effect [1, 3, 29, 56], i.e., DNNs first learn from simple data before fitting to noisy samples. Based on this observation, sample selection strategies utilize confidence levels or prediction loss during training intervals to identify noisy samples [17, 24, 28, 58]. However, this strategy is inherently risky, as the timing of noise memorization varies across datasets and noise distributions, complicating the selection criteria [6, 39]. Consequently, mainline LNL methods can often be integrated with training strategies to enhance model robustness, employing techniques like label filtering [17, 36, 50, 58], label correction [47, 60, 64], robust loss [16, 48, 61], and semi-supervised learning [28]. Among these techniques, semi-supervised learning, by combining multiple approaches [34, 37, 49], often outperforms other LNL methods but is still constrained by the discriminative framework. Although some researches attempt to combine

*Corresponding author.

variational autoencoders (VAE) [25] with DivideMix [28] to address label noise [15], most methods remain focused on discriminative models.

In this paper, we shifts away from the traditional LNL perspective, proposing robust classification models using label diffusion models that focus on the generative modeling and deeply explores the generative relationship between features and labels. Under this paradigm, labels are generated by gradually refining an initial random guess. The uncertainty modeling in diffusion models inherently meets some robustness requirements, as random label generation can offset some noise interference [4, 18]. However, this type of diffusion process lacks information about noisy labels. Intrinsic uncertainty alone is insufficient to handle complex noisy environments. Inspired by the residual diffusion denoising models (RDDM) [32], we explore a novel Directional Label Diffusion (DLD) model that incorporates noisy label information into the diffusion process. In DLD, we decouple the diffusion process into two channels: directional diffusion and random diffusion (see Figure 1). The directional diffusion guides the initial label towards a more probable noise direction (e.g., a green *automobile* may be misclassified as a *frog* or other classes, as indicated by the probability distribution in the third column in Figure 1), while random diffusion retains the original diffusion mechanism. This approach injects information about which categories should not be generated into the label generation paradigm, enhancing the interpretability of the diffusion model's robust inference. Unlike conventional label diffusion, our diffusion process concludes with an endpoint that includes the potential noise distribution of the sample (i.e., $\mathbf{y}_T = \mathbf{y}_n + \epsilon$) and employs two independent coefficient schedules to control the diffusion speed of both channels.

One challenge is that only noisy labels are available in practice for training, so neither $\mathbf{y}_0$ nor $\mathbf{y}_n$ can be obtained directly. Few researchers employed neighborhood detection techniques like KNN to re-estimate $\mathbf{y}_0$ [4]. However, under the IDN assumption, neighborhood labels of noisy samples often contain similar errors, making naive label estimation methods unreliable and potentially causing $\mathbf{y}_0$ to deviate from the ground truth (see Figure 3). To address these issues, we further propose a straightforward and efficient label pre-correction method that automatically provides diverse label information for diffusion model training. This framework employs feature extractors to identify the neighborhood label distribution of samples. Note that these extractors are either trained with self-supervised methods [7, 19, 65], which require few additional training iterations, or use pre-trained models without further training, thus remaining unaffected by label noise. Additionally, this pre-correction strategy allows the seamless integration of pre-trained large-scale models such as Sim-CLR [7] and vision transformer (ViT) [40] into our framework to further boost the performance. The main contributions of this work are summarized as follows:

- We propose a novel Directional Label Diffusion (DLD) model that explicitly introduces interpretable noise information through a dual diffusion architecture, enabling robust learning of label generation patterns from noisy data.
- We design a fast label pre-correction method that automatically provides more accurate label information for the diffusion process and richer feature conditions for the generation process. By leveraging the consistency principle, this method pre-estimates the label distribution in a single iteration without incurring additional costs.
- We refine the paradigm for introducing feature conditions with rigorous theoretical deduction and provide complementary training strategies to enhance robustness further.
- We experimentally show that our model surpasses the state-of-the-art (SOTA) methods on both simulated and real-world noisy datasets, e.g., 6%∼21% accuracy improvement on the CIFAR-100 dataset with IDN.

2. Related Work

2.1. Learning from Noisy Labels

Screening and training strategies are central to LNL methods, aiming to distinguish noisy from clean samples and to optimize model learning under noisy conditions. Screening strategies, like Co-teaching [17], select low-loss samples as clean, while DivideMix [28] uses a Gaussian mixture model to automatically identify clean samples during training. Confidence-based methods like CC [62] and DISC [31] classify samples with unreliable predictions as noisy, whereas neighbor consistency regularization (NCR) [22] promotes learning from samples with high neighbor consistency. Our pre-correction method is based on similarity constraints across multi-view label distributions and leverages a feature extractor to provide high-quality information to the diffusion model without requiring additional iterative training. For training, models such as SEAL [6] iteratively correct noisy labels, but this often risks worsening outcomes [23]. Robust loss functions, like mean absolute error (MAE) [16] and generalized cross entropy (GCE) [61], enhance resilience to noisy labels, while DAC [45] and DMI [54] offer advanced mutual information-based losses. Approaches like DivideMix [28] treat LNL as semi-supervised learning, supported by methods like CC [62], C2D [63], and ELR [33] to refine the process. Without complex iterations or loss design, our method utilizes a lightweight diffusion network and a simple L2 loss to quickly learn the feature-label generation relationship.

2.2. Label Denoising Diffusion Models

Diffusion models were initially designed for image generation tasks and can be understood through the lenses of score matching and dynamics. Ho et al. [21] introduced a

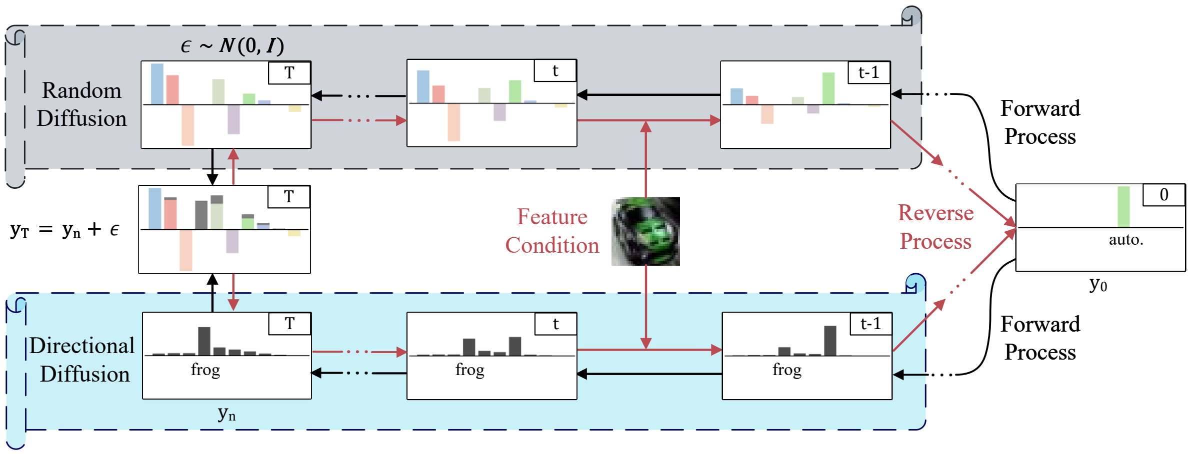


Figure 1. The proposed Directional Label Diffusion (DLD) model includes diagrams illustrating both the forward and reverse processes. In the forward process, diffusion is decoupled into two channels: random diffusion and directional diffusion. Random diffusion represents the transformation of $\mathbf{y}_0$ into random Gaussian noise ϵ, while directional diffusion describes the transformation of $\mathbf{y}_0$ into a specific sample noise label distribution $\mathbf{y}_n$. Ultimately, $\mathbf{y}_0$ diffuses to the endpoint $\mathbf{y}_T = \mathbf{y}_0 + \epsilon$. Notably, the features of the samples serve as control conditions, guiding the generation of labels only during the reverse process.

more classical interpretative framework known as denoising diffusion probabilistic models (DDPM), which have been widely applied across diverse domains. RDDM expands the denoising-based diffusion models, initially uninterpretable for image restoration, into a unified and interpretable model for both image generation and restoration by introducing residuals. Classification and regression diffusion models (CARD) [18] reconstruct the classification as a conditional generative task, generating labels from dimensionally reduced image features via a pre-trained encoder. Building on CARD, label-retrieval-augmented (LRA) diffusion model [4] re-estimates the generation targets through neighbor label retrieval, improving label accuracy. Our model disentangles the diffusion process into directional and random diffusion channels, enhancing diffusion quality through pre-correction, which leads to train a more robust label diffusion model.

3. Preliminary

In this section, we primarily introduce the label denoising diffusion models (LDDM), which is constructed based on the theoretical framework of the DDPM [21]. We start by defining the data distribution $\mathbf{y}_0 \sim q(\mathbf{y}_0)$ and a Markovian forward (diffusion) process q, which progressively adds Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ to the labels based on diffusion strength β_t, diffusing label $\mathbf{y}_0$ to $\mathbf{y}_T$:

$$q(\mathbf{y}_t \mid \mathbf{y}_{t-1}) = \mathcal{N}\left(\mathbf{y}_t; \sqrt{1-\beta_t}\mathbf{y}_{t-1}, \beta_t \mathbf{I}\right), \quad (1)$$

Under Bayes' theorem, one finds that the posterior distribution of the forward process can be represented as a Gaussian distribution with mean $\tilde{\mu}_t(\mathbf{y}_t, \mathbf{y}_0)$ and variance $\tilde{\beta}_t$:

$$q(\mathbf{y}_{t-1} \mid \mathbf{y}_t, \mathbf{y}_0) = \mathcal{N}\left(\mathbf{y}_{t-1}; \tilde{\mu}_t(\mathbf{y}_t, \mathbf{y}_0), \tilde{\beta}_t \mathbf{I}\right), \quad (2)$$

where $\tilde{\mu}_t(\mathbf{y}_t, \mathbf{y}_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{y}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{y}_t$ and $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$, In the reverse process, we naturally employ a neural network to perform Markovian inference from $q(\mathbf{y}_T)$, fitting the posterior distribution $q(\mathbf{y}_{t-1} \mid \mathbf{y}_t)$ through to $\mathbf{y}_0$. Since the variance of the posterior Gaussian distribution depends only on the diffusion strength corresponding to time t, the neural network focuses primarily on predicting and learning the mean:

$$p_\theta(\mathbf{y}_{t-1} \mid \mathbf{y}_t) := \mathcal{N}\left(\mathbf{y}_{t-1}; \mu_\theta(\mathbf{y}_t, t), \tilde{\beta}_t \mathbf{I}\right). \quad (3)$$

To train this model such that $p_\theta(\mathbf{y}_0)$ closely approximates $q(\mathbf{y}_0)$, we can optimize the variational lower bound (VLB), however, research [7] found that directly predicting the mean was not effective, so they reparameterized the estimation target, constraining the neural network to estimate the actual noise of the diffusion process and simplified the loss function:

$$L_{\text{simple}} := E_{t\sim[1,T], \mathbf{y}_0 \sim q(\mathbf{y}_0), \epsilon \sim \mathcal{N}(0,\mathbf{I})}\left[\| \epsilon - \epsilon_\theta(\mathbf{y}_t, t) \|^2\right]. \quad (4)$$

During inference, we utilize the label noise estimated by the model, restoring the mean of the label at time $t-1$ to retrieve its distribution form, and then gradually performing random sampling until generating the label $\mathbf{y}_0$:

$$\mathbf{y}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{y}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{y}_t, t)\right) + \sqrt{\tilde{\beta}_t}\mathbf{z}, \quad (5)$$

where $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$. Unfortunately, this label generation process lacks any conditional information, resulting in random label generation for images with identical features, which fails to achieve classification. To address this, we analyze the foundational paradigm of conditional label diffusion models in Appendix A and simplify the method of feature conditioning.

4. Directional Label Diffusion Model

Based on LDDM, our goal is to develop a directional diffusion process on labels to introduce more knowledge about noise. We modify the representation of $\mathbf{y}_T = \epsilon$ in LDDM to $\mathbf{y}_T = \mathbf{y}_n + \epsilon$ in our DLD, where $\mathbf{y}_n$ is an estimation of noisy label distribution. The improved forward process is no longer a simple random noise diffusion but is disentangled into two paths of directional diffusion and random diffusion, as shown in Figure 1.

In the following subsections, we first outline the overall framework of our model in Section 4.1. Next, in Section 4.2, we present the foundational theory of the DLD diffusion process and analyze the method of conditional control. In Section 4.3, we describe the reverse process and training method, including optimizations for introducing feature conditions and an adapted accelerated inference method. Finally, in Section 4.4, we describe the rapid label pre-correction method within the DLD framework.

4.1. Model Overview

We propose the Directional Label Diffusion (DLD) framework, with the overall process illustrated in Figure 2. The framework contains two main stages: label pre-correction (Stage 1) and the directional label diffusion model (Stage 2). In stage 1, we first use two different data augmentation methods to provide two views for each sample, then use a pre-trained feature extractor f_{ext} to observe the neighborhood label distribution p_w and p_s of the same sample in two feature spaces. Based on similarity constraints and the principle of neighborhood consistency, we divide the dataset into clean, noisy, and hard sets. Then, The label pre-corrector and noise distribution estimator prepare $\mathbf{y}_0$ and $\mathbf{y}_n$ based on the sample partitioning results, using $\mathbf{y}_0$ as the initial diffusion label and $\mathbf{y}_n$ as part of the diffusion endpoint for the next stage of model training. Otherwise, we introduce both augmented image features and high-quality features extracted by f_{ext} into the training process to jointly guide the learning of the robust diffusion model.

4.2. Forward Process of DLD

To model the random diffusion of labels and the addition of directional noise information, we define the single forward (diffusion) process step in our DLD as follows:

$$\mathbf{y}_t = \mathbf{y}_{t-1} + \alpha_t \mathbf{y}_d + \beta_t \epsilon, \tag{6}$$

where $\mathbf{y}_d$ is the deviation between $\mathbf{y}_n$ and y_0 (i.e., $\mathbf{y}_d = \mathbf{y}_n - \mathbf{y}_0$), representing the directional shift contained in the diffusion from $\mathbf{y}_{t-1}$ to $\mathbf{y}_t$. Two independent coefficients,α_t and β_t control the degrees of directional and random diffusion, respectively. Based on the additivity of Gaussian noise, we derive the jump diffusion formula from $\mathbf{y}_0$ to $\mathbf{y}_t$ without requiring single-step continuous diffusion (see Appendix B for the proof):

$$\mathbf{y}_t = \mathbf{y}_0 + \bar{\alpha}_t \mathbf{y}_d + \bar{\beta}_t \epsilon, \tag{7}$$

where $\bar{\alpha}_t = \sum_{i=1}^{t} \alpha_i$ and $\bar{\beta}_t = \sqrt{\sum_{i=1}^{t} \beta_i^2}$. When $t = T$, we hope $\mathbf{y}_T = \mathbf{y}_n + \epsilon = \mathbf{y}_0 + \mathbf{y}_d + \epsilon$, so we usually set $\bar{\alpha}_T = \bar{\beta}_T = 1$. This implies that the two coefficients α_t and β_t are not arbitrary transformations; rather, they should satisfy a specific sum constraint. Then, we redefine the conditional Markov diffusion process similar to the directional label forward distribution q without condition (detailed theoretical framework in Appendix C), assuming that when conditioned on features $\mathbf{x}$, the prior distribution of $\hat{q}$ is the same as q:

$$\begin{aligned} \hat{q}\left(\mathbf{y}_t|\mathbf{y}_{t-1}, \mathbf{y}_d, \mathbf{x}\right) &:= q\left(\mathbf{y}_t|\mathbf{y}_{t-1}, \mathbf{y}_d\right) \\ &= \mathcal{N}\left(\mathbf{y}_t; \mathbf{y}_{t-1} + \alpha_t \mathbf{y}_d, \beta_t^2 \mathbf{I}\right), \end{aligned} \tag{8}$$

$$\begin{aligned} \hat{q}\left(\mathbf{y}_{1:T} \mid \mathbf{y}_0, \mathbf{y}_d, \mathbf{x}\right) &:= \prod_{t=1}^{T} \hat{q}\left(\mathbf{y}_t \mid \mathbf{y}_{t-1}, \mathbf{y}_d, \mathbf{x}\right) \\ &= \mathcal{N}\left(\mathbf{y}_t; \mathbf{y}_0 + \bar{\alpha}_t \mathbf{y}_d, \bar{\beta}_t^2 \mathbf{I}\right), \end{aligned} \tag{9}$$

which means that the diffusion process of the labels is not influenced by any image features.

In Appendix C, using Bayes' rule and Taylor expansion, we derive the conditional posterior distribution during the forward process:

$$\hat{q}\left(\mathbf{y}_{t-1} \mid \mathbf{y}_t, \mathbf{y}_0, \mathbf{y}_d, \mathbf{x}\right) = \mathcal{N}\left(\mathbf{y}_{t-1}; \hat{\mu}_t, \hat{\sigma}_t \mathbf{I}\right), \tag{10}$$

where $\hat{\mu}_t = y_t - \alpha_t y_d - \frac{\beta_t^2}{\bar{\beta}_t}\epsilon + \hat{\sigma}_t \nabla_{y_{t-1}} \log \hat{q}\left(\mathbf{x}|\mathbf{y}_{t-1}, \mathbf{y}_d\right)$ and $\hat{\sigma}_t = \frac{\beta_t^2 \bar{\beta}_{t-1}^2}{\bar{\beta}_t^2}$. The last term in $\hat{\mu}_t$ can be considered a deviation to the unconditional distribution mean, guided by the first derivative of $\hat{q}\left(\mathbf{x}|\mathbf{y}_{t-1}, \mathbf{y}_d\right)$, indicating that the decoding information from $\mathbf{y}_{t-1}$ to $\mathbf{x}$ should be introduced in training.

4.3. Reverse and Training Process of DLD

Reverse Process and Training. In the reverse (generation) process, as indicated by Eq. 6, we need to simultaneously predict $\mathbf{y}_d$ and ϵ to progressively recover $\mathbf{y}_T$ to $\mathbf{y}_0$. Therefore, we can train two independent networks to estimate $\mathbf{y}_d^{\theta_1}$ and ϵ_{θ_2}, respectively, to compute the reverse posterior distribution. Furthermore, we directly use feature information as input for two networks:

$$p_\theta\left(\mathbf{y}_{t-1} \mid \mathbf{y}_t, \mathbf{x}\right) := \mathcal{N}\left(\mathbf{y}_{t-1}; \mu_\theta, \hat{\sigma}_t^2 \mathbf{I}\right). \tag{11}$$

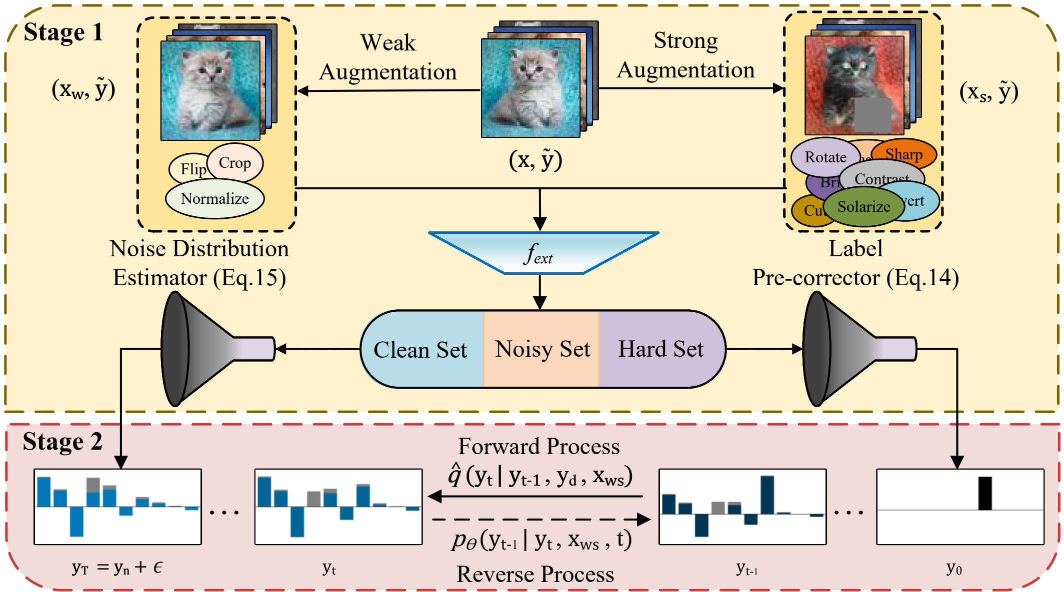


Figure 2. Overview of the DLD framework. During Stage 1, the noisy dataset is quickly divided into three subsets. Then, we estimate $\mathbf{y}_0$ and $\mathbf{y}_n$ for each samples based on Eq. 14 and Eq. 15. In Stage 2, the label information and augmented features from Stage 1 are jointly used to train a directional diffusion model (blue bars represent random diffusion and gray bars represent directional diffusion).

Based on the analysis of the forward process, we adopt a straightforward approach which naturally integrates the decoding process into the diffusion model architecture by a fixed decoding layer L_1 to map $\mathbf{y}$ into the feature space of $\mathbf{x}$. Detailed information about the specific diffusion model network structure is shown in Appendix Figure C.1.

By optimizing the VLB and reparameterization [21], we derive the following simplified loss function for training:

$$\mathcal{L}_d = ||\ \mathbf{y}_d - \mathbf{y}_d^{\theta_1}\left(\mathbf{y}_t, \mathbf{y}_n, \mathbf{x}, t\right)\ ||^2, \tag{12}$$

$$\mathcal{L}_\epsilon = ||\ \epsilon - \epsilon_{\theta_2}\left(\mathbf{y}_t, \mathbf{y}_n, \mathbf{x}, t\right)\ ||^2, \tag{13}$$

where $\epsilon \sim \mathcal{N}\left(0, \mathbf{I}\right)$, and the training input $\mathbf{y}_t$ is synthesized using Eq. 7. Algorithm 1 describes the training procedure.

Inference. Since the classification diffusion model is deterministic in inference and the dimensionality of the label vector is much lower than that of images, we mimic the inference process of DDIM [44] to generate labels in as few steps as possible without affecting the generated categories. We can gradually sample labels step-by-step following $\mathbf{y}_{t-1} = \mathbf{y}_t - (\bar{\alpha}_t - \bar{\alpha}_{t-1})\mathbf{y}_d^{\theta_1} - (\bar{\beta}_t - \bar{\beta}_{t-1})\epsilon_{\theta_2}$, where the transition from $t-1$ to t is discrete rather than continuous in time. Appendix D provides a detailed explanation of the de-Markovization process, the accelerated sampling algorithm and the visual demonstration of the sampling process. In experiments, we use 5 steps for DLD's inference, which is on the same order of magnitude as the speed of traditional classification models.

4.4. Label Pre-correction Method

Although diffusion models are powerful, under the guidance of incorrect $\mathbf{y}_0$ and $\mathbf{y}_n$, they can easily fit the wrong conditional distributions. Therefore, we propose a one-step label pre-correction method that automatically provides accurate label information and feature conditions for the DLD model, requiring no additional parameter training, making it simple, straightforward, and efficient.

Label Pre-correction for $\mathbf{y}_0$. Inspired by the concepts of multi-view[42, 53] and neighbor consistency regularization [22], our primary assumption is that the same sample should have similar neighborhood label distributions in different feature spaces. We define the original noisy dataset as $\mathcal{D} = \left\{(x_i, \tilde{y}_i) \mid x_i \in \mathbb{R}^d, \tilde{y}_i \in \mathbb{R}^m, i = 1, \ldots, n\right\}$. To obtain different feature spaces by a shared feature extractor f_{ext}, we first apply two different data augmentation methods and map them into $\mathbb{R}^w$ and $\mathbb{R}^s$, with the augmented datasets respectively denoted as $\mathcal{D}^w$ and $\mathcal{D}^s$. Then, we use Euclidean or Cosine similarity to obtain the nearest K neighborhood labels $y_{k=1}^K$ and feature space distances $v_{k=1}^K$, then we can get two weighted label distributions p_w and p_s by $p(y_i) = \left(\sum_{k=1}^K w_k \cdot \mathbf{1}(y_k = y_i)\right) \cdot \left(\sum_{k=1}^K w_k\right)^{-1}$, where $w_k = \frac{1}{v_k + \delta}$, where δ is a small constant to prevent division by zero. Next, we calculate the divergence between the two distributions using the Kullback-Leibler divergence and a via Gaussian mixture model (GMM) [38] to re-partitioning three types of samples: Samples with high means are reassigned to the hard set $\mathcal{D}_H$, which represents samples whose correctness is hard to determine and typically lie at the boundaries of classes. Samples with low means indicate that similar neighborhood distributions under both views. Using the principle of neighborhood consistency, we compare $argmax\,(p_{ws})$ with the original label $\tilde{y}$, where $p_{ws} = \frac{p_w + p_s}{2}$. If the class with the highest frequency in the average neighborhood distribution matches the original label, then the sample is assigned to the clean set $\mathcal{D}_C$; otherwise, it is assigned to the noisy set $\mathcal{D}_N$.

Based on the repartitioned sample subsets and the original noisy labels $\tilde{y}$, we define a label pre-corrector, which functions similarly to a transition matrix. It can automatically perform label flipping or refinement:

$$P\left(y_0 \mid \tilde{y}, x\right) = \begin{cases} \tilde{y} & \text{if } x \in \mathcal{D}_C, \\ \operatorname{argmax}\left(p_{ws}\right) & \text{if } x \in \mathcal{D}_N, \\ p_{ws} & \text{if } x \in \mathcal{D}_H. \end{cases} \tag{14}$$

According to Eq. 14, for samples with deterministic neighborhood label distributions, we use hard labels to minimize the interference of redundant information on y_0. For samples with more contentious neighborhood label distributions, we adopt a soft strategy for label refinement. Instead of point estimation, we estimate through a label vector, retaining most of the neighborhood information. This strategy enhances the accuracy of the y_0 while preserving as much robust information as possible.

Estimate Noise distribution for $\mathbf{y}_n$. Similar to obtaining $\mathbf{y}_0$, we use the subset of divided samples to estimate the noise distribution of the samples. According to Eq. 15, when a sample is identified as clean, the directed diffusion term is used to gradually fade out $\mathbf{y}_0$ ($\mathbf{y}_d = -\mathbf{y}_0$), allowing the model to learn to recover a label closer to the ground truth from an uninformed ϵ. When the sample contains noise, the endpoint of diffusion should include a clear representation of noise, enabling the model to strictly avoid recovering such noisy vector values during reverse generation. For uncertain samples, we simulate potential noise distributions by retaining the difference in label distributions between two views. The greater the difference in the distribution values for a particular class in the two views, the more likely the sample is incorrectly labeled as that class.

$$P\left(y_n \mid \tilde{y}, x\right) = \begin{cases} 0 & \text{if } x \in \mathcal{D}_C, \\ \tilde{y} & \text{if } x \in \mathcal{D}_N, \\ \frac{\left|p_w - p_s\right|}{\sum \left|p_w - p_s\right|} & \text{if } x \in \mathcal{D}_H. \end{cases} \tag{15}$$

The complete training process is outlined in Algorithm 1, consisting of the following phases: Steps $1 \sim 3$ for sample re-partitioning, Steps $4 \sim 9$ for diffusion model training, with Step 6 dedicated to $\mathbf{y}_0$ pre-correction and $\mathbf{y}_n$ estimation. Notably, in Step 5, instead of sampling features from the original noisy set $\mathcal{D}$, we use a mixture of sample features from $\mathcal{D}^w$ and $\mathcal{D}^s$ as the training input for the diffusion

Algorithm 1 DLD Training

Input: Noisy training set $\mathcal{D} = \{\mathbf{X}, \tilde{\mathbf{Y}}\}$, feature extractor f_{ext}, two network parameters θ_1 and θ_2
Output: θ_1 and θ_2

1: Obtain two augmented features $\mathcal{D}^w$ and $\mathcal{D}^s$ using the feature extractor f_{ext}.
2: Calculate K-nearest label distributions p_w and p_s.
3: Re-partitioning samples into $\mathcal{D}_C$, $\mathcal{D}_N$ and $\mathcal{D}_H$.
4: **while** not converged **do**
5: Sample data $(\mathbf{x}_{ws}, \tilde{\mathbf{y}}) \sim \{\mathcal{D}^w, \mathcal{D}^s\}$; time slice $t \sim \{1, \ldots, T\}$; and noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$.
6: Pre-estimate $\mathbf{y}_0$ and $\mathbf{y}_n$ based on Eq. 14 and Eq. 15, respectively.
7: Accumulate losses through Eq. 12 and Eq. 13.
8: Take gradient descent step on the losses to update network parameters.
9: **end while**

model, where $\mathbf{x}_{ws} = (1 - \gamma)\,\mathbf{x}_w + \gamma \mathbf{x}_s$, with $\gamma = \frac{|\mathcal{D}_H|}{|\mathcal{D}|}$ representing the proportion of uncertain samples, which means that the more difficult the sample classification, the richer the feature conditions the model automatically introduces. This reflects an implicit regularization method that mitigates the issue of the diffusion model being overly confident when still learning incorrect information.

5. Experimental Results and Analysis

In this section, we first evaluate the DLD's effectiveness on CIFAR datasets with varying levels of IDN in Section 5.1. Then, We conduct ablation studies in Section 5.2 to analyze the contributions of the pre-trained feature extractor, the label pre-correction technology, and other components. In Section 5.3, we compare the DLD to the SOTA methods across four real-world noisy datasets. Appendix E provides more details about our experiments.

5.1. Results on Simulated Noisy Datasets

We conduct experiments on the CIFAR-10 and CIFAR-100 datasets [27] with simulated IDN [6]. Details of the IDN generation algorithm and examples are provided in Appendix E.1. In our experiments, we generate noise levels of 20%, 40%, and 60%. To obtain the feature space for label pre-correction, we provide multiple configurations of feature extractors, including DLD+ResNet, DLD+SimCLR and DLD+ViT (see Appendix F.2 for detailed settings).

As shown in Table 1, the DLD+ResNet configuration outperforms the baseline across most noise levels on CIFAR, especially in high-noise environments. We analyze that the main factor affecting its performance in low noise environment is the introduction of too much robust information, which leads to the interference of the model. When using the DLD+ViT configuration, its performance surpasses

Table 1. Classification accuracies (%) of CIFAR datasets under different levels of IDN.

Method	CIFAR-10			CIFAR-100		
	20%	40%	60%	20%	40%	60%
Co-teaching [17]	87.28±0.20	78.72±0.47	62.51±1.98	69.26±0.39	59.53±0.57	46.80±0.51
GCE [61]	86.44±0.23	76.71±0.39	60.76±3.08	68.24±0.27	55.76±1.18	46.75±2.46
DAC [45]	86.16±0.13	74.80±0.32	59.97±0.90	68.70±0.78	55.71±1.01	46.29±3.35
DMI [54]	86.57±0.16	77.81±0.85	69.94±1.31	68.71±1.21	57.99±0.92	50.96±2.08
SEAL [6]	87.79±0.09	82.98±0.05	70.07±1.54	70.61±0.19	61.18±0.35	51.86±1.32
NCR [22]	87.74±0.15	83.03±0.21	69.44±1.01	70.68±0.37	61.93±0.54	49.13±2.77
CC+DivideMix [62]	88.36±0.11	84.18±0.40	71.52±1.73	70.09±0.19	64.35±0.35	53.20±0.66
DISC [31]	90.68±0.10	85.61±0.32	73.59±1.13	71.49±0.13	64.46±0.29	53.62±0.51
LRA-diffusion+ResNet [4]	89.94±0.15	86.73±0.37	74.66±1.02	72.35±0.14	65.55±0.30	54.43±0.53
DLD+ResNet (Ours)	90.43±0.12	87.17±0.29	76.07±1.10	72.60±0.15	66.59±0.31	55.72±0.24
EPL+ViT [26]	95.95±0.17	93.01±0.33	76.16±1.18	72.24±0.12	67.20±0.20	55.59±0.13
LRA-diffusion+ViT [4]	96.65±0.18	93.68±0.34	78.12±1.17	73.59±0.13	67.67±0.21	57.21±0.12
DLD+ViT (Ours)	**96.98±0.12**	**96.49±0.07**	**83.14±0.81**	**78.62±0.11**	**76.03±0.14**	**69.57±0.17**

EPL+ViT and LRA-diffusion+ViT. This implies that DLD more effectively utilizes the provided prior knowledge without requiring additional training conditions. Across various datasets, DLD's superiority is more pronounced on CIFAR-100, indicating that the more complex IDN on CIFAR-100 severely impacts the performance of other methods, whereas our model maintains robust learning.

Additionally, we conducted experiments on the simulated CCN dataset. The results (see Appendix E.2) show that CCN is easier to handle than IDN, which aligns with our expectations. Compared to other methods, DLD still achieves the highest classification accuracy. Due to the lightweight architecture of the diffusion model and the single-step pre-correction process, DLD outperforms traditional LNL methods without incurring significant computational overhead. A detailed analysis of training time overhead and model parameters is provided in Appendix E.3.

Table 2. Classification accuracies (%) on Animal-10N.

Method	**Accuracy**
CE [35]	80.6
GCE [61]	81.5
NCT [8]	84.1
DISC [31]	87.3
SSR [14]	88.5
LRA-diffusion [4]	88.6
J-ViT [9]	89.0
DLD (Ours)	**89.4**

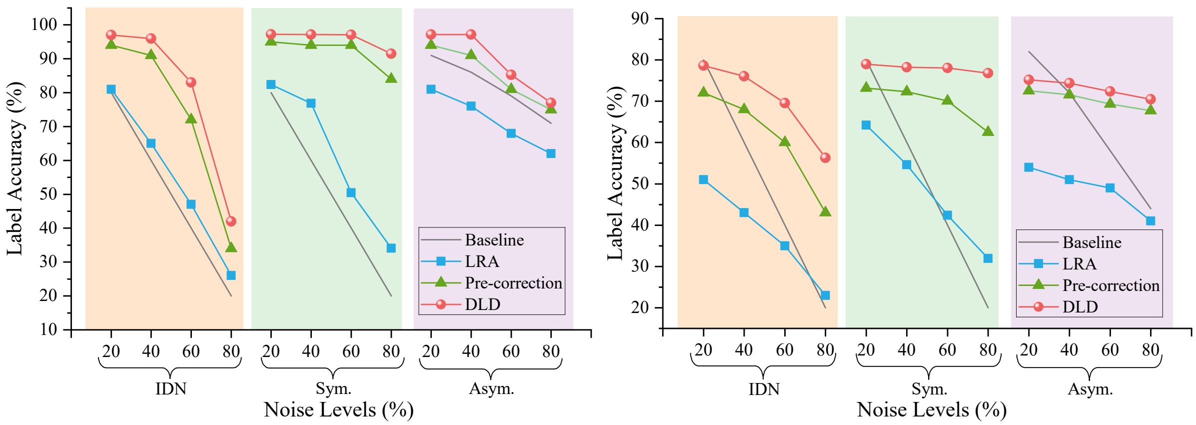
5.2. Ablation Studies

Power of label pre-correction. We validate the effectiveness of label pre-correction by freezing the diffusion model. As shown in Figure 3, our pre-correction method consistently outperforms in various noisy environments, while Figure 3b demonstrates that the LRA estimation method disrupts originally correct labels. Then, we unfroze the diffusion model to verify that DLD does not fully depend on the pre-correction method. Results show that DLD further enhances generalization performance based on pre-correction, indicating that the two methods complement each other effectively. Additionally, we design more comprehensive cross-ablation experiments, e.g., replacing the LRA technique in LRA-diffusion with the pre-correction module to demonstrate the necessity of directional diffusion. See Appendix F.1 for more details.

Contribution of different pre-trained features and other training components. We explore the impact of various feature extractors, including ResNet34 [20], SimCLR [7] and ViT-L/14 [40], and conduct ablation studies comparing our method with CARD and LRA-diffusion. Here, DISC [31] served as a baseline that does not require any pre-trained models. Detailed results and analyses are provided in Appendix F.2. Our findings reveal that DLD outperforms other diffusion models when combined with different pre-trained models, achieving optimal performance when integrated with ViT-L/14. This underscores ViT-L/14's effectiveness in enhancing label pre-correction and facilitating diffusion model learning. Furthermore, Appendix F.3 details additional ablation studies on various hyperparameter settings. We validate the optimal training strategy for the DLD method and recommend this combination.

5.3. Results on Real-world Noisy Datasets

We select four real-world noisy datasets for our experiments to demonstrate the superior performance of DLD: Animal10N [43], Clothing1M [52], WebVision [30], and ImageNet ILSVRC2012 [41]. Comprehensive descriptions of these datasets and detailed experimental setups can be



(a) Experimental results on the CIFAR-10 dataset

(b) Experimental results on the CIFAR-100 dataset

Figure 3. Comparison of different label correction strategies' accuracies, where Baseline represents the proportion of clean labels.

Table 3. Classification accuracies (%) on Clothing1M.

Method	Accuracy
Co-teaching [17]	69.21
DMI [54]	72.46
DivideMix [28]	74.76
CC [62]	75.40
DISC [31]	74.79
EPL [26]	75.21
LRA-diffusion [4]	74.46
SANM [46]	75.63
DLD (Ours)	**75.69**

Table 4. Classification accuracies (%) on WebVision and ILSVRC2012 datasets.

Method	WebVision	ILSVRC12
Co-teaching [17]	63.58	61.48
DivideMix [28]	77.32	75.20
ELR [33]	77.78	70.29
EPL [26]	78.77	76.51
C2D [63]	79.42	78.57
CC [62]	79.36	76.08
DISC [31]	80.28	77.44
LRA-diffusion [4]	84.16	82.56
DLD (Ours)	**84.51**	**83.74**

found in Appendix E.4 and Appendix E.5, respectively. As demonstrated in Table 2, Table 3, and Table 4, DLD outperforms all SOTA methods across these datasets. Note that DLD's relative classification performance improved by 7.26% and 0.44% over J-ViT [9] and EPL [26], respectively, both employing the ViT architecture. Compared to LRA-diffusion, which also employs ViT and label diffusion models, DLD achieves varying degrees of performance improvement, demonstrating a more effective and reasonable use of prior knowledge and diffusion models. Besides, while ViT-L/14 improved performance on Clothing1M, Appendix G shows that other extractors like CC can be even more effective, highlighting our method's versatility with different LNL methods. Besides, DLD's performance advantage partly stems from the pre-correction method. In future work, we aim to integrate discriminative and generative paradigms, using diffusion models to refine pre-correction, generate label information, and enhance representational power and multimodal generalization. Appendix H discusses additional limitations and future directions.

6. Conclusion

This paper proposes DLD, a diffusion model specifically designed to learn effectively from noisy labels for robust classification. Our main idea is to decouple the label diffusion process into two channels: directional diffusion and random diffusion. This approach explicitly introduces noisy label information into the diffusion model for the first time, enabling the model to interpretably mitigate the generation of incorrect labels, thereby enhancing its ability to generalize from noisy to clean data. By incorporating our low-cost label pre-correction method, DLD can automatically select label diffusion information for each sample in diverse noisy environments, without requiring additional adjustments or iterations. Additionally, our model synergizes with various pre-trained large-scale models, like ViT, leveraging their prior knowledge more effectively than other methods. Experiments on simulated and real-world noisy datasets show that DLD surpasses conventional LNL methods in both accuracy and efficiency. By framing LNL within a generative model, our method bridges LNL with other learning frameworks. We believe this work will reduce the workload for practitioners handling large-scale, low-quality labeled data, benefiting the development of the image recognition field.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant U21A20513, Grant 62476157, Grant 62276161, Grant 62276157 and Grant 62406179, and in part by the Key Research and Development Program of Shanxi Province under Grant 202202020101003 and the Fundamental Research Program of Shanxi Province under Grant 202403021222026.

References

[1] Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017. 1

[2] Antonin Berthon, Bo Han, Gang Niu, Tongliang Liu, and Masashi Sugiyama. Confidence scores make instance-dependent label-noise learning possible. In *International conference on machine learning*, pages 825–836. PMLR, 2021. 1

[3] Stella Biderman, Usvsn Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. Emergent and predictable memorization in large language models. *Advances in Neural Information Processing Systems*, 36:28072–28090, 2023. 1

[4] Jian Chen, Ruiyi Zhang, Tong Yu, Rohan Sharma, Zhiqiang Xu, Tong Sun, and Changyou Chen. Label-retrieval-augmented diffusion models for learning from noisy labels. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 7, 8

[5] 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*, pages 1062–1070. PMLR, 2019. 1

[6] Pengfei Chen, Junjie Ye, Guangyong Chen, Jingwei Zhao, and Pheng-Ann Heng. Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11442–11450, 2021. 1, 2, 6, 7

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 3, 7

[8] Yingyi Chen, Shell Xu Hu, Xi Shen, Chunrong Ai, and Johan AK Suykens. Compressing features for learning with noisy labels. *IEEE Transactions on Neural Networks and Learning Systems*, 35(2):2124–2138, 2022. 7

[9] Yingyi Chen, Xi Shen, Yahui Liu, Qinghua Tao, and Johan AK Suykens. Jigsaw-vit: Learning jigsaw puzzles in vision transformer. *Pattern Recognition Letters*, 166:53–60, 2023. 7, 8

[10] De Cheng, Tongliang Liu, Yixiong Ning, Nannan Wang, Bo Han, Gang Niu, Xinbo Gao, and Masashi Sugiyama. Instance-dependent label-noise learning with manifold-regularized transition matrix estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16630–16639, 2022. 1

[11] Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. *arXiv preprint arXiv:2010.02347*, 2020.

[12] Jiacheng Cheng, Tongliang Liu, Kotagiri Ramamohanarao, and Dacheng Tao. Learning with bounded instance and label-dependent label noise. In *International conference on machine learning*, pages 1789–1799. PMLR, 2020. 1

[13] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5562–5570, 2016. 1

[14] Chen Feng, Georgios Tzimiropoulos, and Ioannis Patras. Ssr: An efficient and robust framework for learning with unknown label noise. *arXiv preprint arXiv:2111.11288*, 2021. 7

[15] Arpit Garg, Cuong Nguyen, Rafael Felix, Thanh-Toan Do, and Gustavo Carneiro. Instance-dependent noisy label learning via graphical modelling. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2288–2298, 2023. 2

[16] Aritra Ghosh, Himanshu Kumar, and P Shanti Sastry. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, 2017. 1, 2

[17] 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. 1, 2, 7, 8

[18] Xizewen Han, Huangjie Zheng, and Mingyuan Zhou. Card: Classification and regression diffusion models. *Advances in Neural Information Processing Systems*, 35:18100–18115, 2022. 2, 3

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 2

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3, 5

[22] Ahmet Iscen, Jack Valmadre, Anurag Arnab, and Cordelia Schmid. Learning with neighbor consistency for noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4672–4681, 2022. 2, 5, 7

[23] Gaoxia Jiang, Jia Zhang, Xuefei Bai, Wenjian Wang, and Deyu Meng. Which is more effective in label noise cleaning, correction or filtering? In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12866–12873, 2024. 2

[24] 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*, pages 2304–2313. PMLR, 2018. 1

[25] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[26] Jongwoo Ko, Sumyeong Ahn, and Se-Young Yun. Efficient utilization of pre-trained model for learning with noisy labels. In *ICLR 2023 Workshop on Pitfalls of limited data and computation for Trustworthy ML*, 2023. 7, 8

[27] Alex Krizhevsky. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, 2009. 6

[28] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020. 1, 2, 8

[29] Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *International conference on artificial intelligence and statistics*, pages 4313–4324. PMLR, 2020. 1

[30] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017. 7

[31] Yifan Li, Hu Han, Shiguang Shan, and Xilin Chen. Disc: Learning from noisy labels via dynamic instance-specific selection and correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24070–24079, 2023. 2, 7, 8

[32] Jiawei Liu, Qiang Wang, Huijie Fan, Yinong Wang, Yandong Tang, and Liangqiong Qu. Residual denoising diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2773–2783, 2024. 2

[33] 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. 2, 8

[34] Yangdi Lu, Yang Bo, and Wenbo He. Co-matching: Combating noisy labels by augmentation anchoring. *arXiv preprint arXiv:2103.12814*, 2021. 1

[35] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications. In *International conference on Machine learning*, pages 23803–23828. PMLR, 2023. 7

[36] Duc Tam Nguyen, Chaithanya Kumar Mummadi, Thi Phuong Nhung Ngo, Thi Hoai Phuong Nguyen, Laura Beggel, and Thomas Brox. Self: Learning to filter noisy labels with self-ensembling. *arXiv preprint arXiv:1910.01842*, 2019. 1

[37] 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*, pages 8022–8031, 2021. 1

[38] Haim Permuter, Joseph Francos, and Ian Jermyn. A study of gaussian mixture models of color and texture features for image classification and segmentation. *Pattern recognition*, 39(4):695–706, 2006. 5

[39] Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 2002. 1

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 7

[41] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 7

[42] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 5

[43] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. Selfie: Refurbishing unclean samples for robust deep learning. In *International conference on machine learning*, pages 5907–5915. PMLR, 2019. 7

[44] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 5

[45] Sunil Thulasidasan, Tanmoy Bhattacharya, Jeff Bilmes, Gopinath Chennupati, and Jamal Mohd-Yusof. Combating label noise in deep learning using abstention. *arXiv preprint arXiv:1905.10964*, 2019. 2, 7

[46] Yuanpeng Tu, Boshen Zhang, Yuxi Li, Liang Liu, Jian Li, Jiangning Zhang, Yabiao Wang, Chengjie Wang, and Cai Rong Zhao. Learning with noisy labels via self-supervised adversarial noisy masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16186–16195, 2023. 8

[47] Yisen Wang, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song, and Shu-Tao Xia. Iterative learning with open-set noisy labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8688–8696, 2018. 1

[48] 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 computer vision and pattern recognition*, pages 322–330, 2019. 1

[49] 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 con-*

ference on computer vision and pattern recognition, pages 13726–13735, 2020. 1

[50] Pengxiang Wu, Songzhu Zheng, Mayank Goswami, Dimitris Metaxas, and Chao Chen. A topological filter for learning with label noise. *Advances in neural information processing systems*, 33:21382–21393, 2020. 1

[51] Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. *Advances in Neural Information Processing Systems*, 33:7597–7610, 2020. 1

[52] 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*, pages 2691–2699, 2015. 7

[53] Cai Xu, Yilin Zhang, Ziyu Guan, and Wei Zhao. Trusted multi-view learning with label noise. *arXiv preprint arXiv:2404.11944*, 2024. 5

[54] 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. 2, 7, 8

[55] Yan Yan, Rómer Rosales, Glenn Fung, Ramanathan Subramanian, and Jennifer Dy. Learning from multiple annotators with varying expertise. *Machine learning*, 95:291–327, 2014. 1

[56] Quanming Yao, Hansi Yang, Bo Han, Gang Niu, and James Tin-Yau Kwok. Searching to exploit memorization effect in learning with noisy labels. In *International Conference on Machine Learning*, pages 10789–10798. PMLR, 2020. 1

[57] Yu Yao, Tongliang Liu, Mingming Gong, Bo Han, Gang Niu, and Kun Zhang. Instance-dependent label-noise learning under a structural causal model. *Advances in Neural Information Processing Systems*, 34:4409–4420, 2021. 1

[58] 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*, pages 7164–7173. PMLR, 2019. 1

[59] 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, 2021. 1

[60] Yikai Zhang, Songzhu Zheng, Pengxiang Wu, Mayank Goswami, and Chao Chen. Learning with feature-dependent label noise: A progressive approach. *arXiv preprint arXiv:2103.07756*, 2021. 1

[61] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018. 1, 2, 7

[62] Ganlong Zhao, Guanbin Li, Yipeng Qin, Feng Liu, and Yizhou Yu. Centrality and consistency: two-stage clean samples identification for learning with instance-dependent noisy labels. In *European Conference on Computer Vision*, pages 21–37. Springer, 2022. 2, 7, 8

[63] Evgenii Zheltonozhskii, Chaim Baskin, Avi Mendelson, Alex M Bronstein, and Or Litany. Contrast to divide: Self-supervised pre-training for learning with noisy labels. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1657–1667, 2022. 2, 8

[64] Songzhu Zheng, Pengxiang Wu, Aman Goswami, Mayank Goswami, Dimitris Metaxas, and Chao Chen. Error-bounded correction of noisy labels. In *International Conference on Machine Learning*, pages 11447–11457. PMLR, 2020. 1

[65] Yuanyi Zhong, Haoran Tang, Junkun Chen, Jian Peng, and Yu-Xiong Wang. Is self-supervised learning more robust than supervised learning? *arXiv preprint arXiv:2206.05259*, 2022. 2