

CLIPCleaner: Cleaning Noisy Labels with CLIP

Anonymous Authors

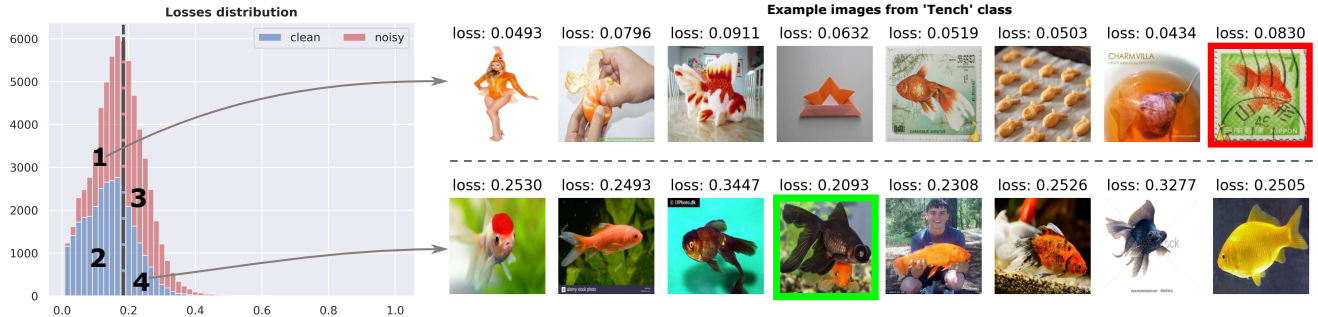


Figure 1: Loss distribution of the WebVision dataset after one epoch of warmup training with the entire dataset is shown here. Samples are categorized as ‘clean’ or ‘noisy’ based on CLIPCleaner’s identification. The vertical gray line represents the sample selection boundary imposed by the ‘small-loss’ mechanism (further details in introduction part). We visualize examples in part 1 and part 4. These examples represent samples classified as ‘clean’ by ‘small-loss’ but rejected by CLIPCleaner, and vice versa. Specifically, two samples from the ‘Tench’ class are highlighted in red and green. We can not find the specific source of red image, but highly-related images can be found with keyword: 1966 Japanese Goldfish Stamp Postage, while the green one can be originated back to: https://acn1.fandom.com/wiki/Pop-Eyed_Goldfish. The red sample is a tench fish postage stamp, semantically similar to real tench images, resulting in a lower loss. The green sample, however, is a photo of a black pop-eyed goldfish, which deviates more from the typical golden tench fish visually.

ABSTRACT

Noisy labels pose a significant challenge for machine learning models. Existing sample selection methods for Learning with Noisy Labels (LNL), often based on a strategy like selecting samples with ‘small loss’, can suffer from ‘self-confirmation bias’. This bias arises because these methods rely on the in-training model, which itself might be misled by the noisy labels. Furthermore, solely relying on visual information can introduce biases and challenges like ‘hard noise’, where noisy labels incorrectly assign samples to semantically similar categories. This paper proposes addressing these challenges by leveraging CLIP, a powerful vision-language model, for sample selection. We introduce CLIPCleaner, which utilizes CLIP’s pre-trained zero-shot classifier along with a classifier based on CLIP’s vision encoder and the noisy labels themselves. Our approach enables effective offline sample selection. We also provide theoretical justifications and empirical evidence to demonstrate the advantages of CLIP compared to conventional pre-trained models. Compared to current methods that combine iterative sample selection with various techniques, CLIPCleaner offers a streamlined approach

while achieving competitive or superior performance on benchmark datasets. Our work highlights the potential of large-scale vision-language models for tackling LNL problems.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning**: *Computer vision representations*; *Learning under covariate shift*.

KEYWORDS

Sample selection, Noisy Labels, CLIP

1 INTRODUCTION

Over the past two decades, deep neural networks have demonstrated exceptional success in various vision tasks, attributed to the existence of high-precision, large-scale datasets such as ImageNet-1K. However, collecting high-quality labels for such datasets is generally time-consuming and labor-intensive. To mitigate the cost, an alternative is automatic labeling (e.g. “webly-labeled” dataset by web-crawling the images and labels). While reducing the cost of manual labeling, it inevitably leads to low-quality noisy labels.

To address the problem of label noise, a variety of methods have been proposed. Some methods aim to develop robust loss functions [8, 11, 29, 41, 43, 54, 60, 64] or noise transition matrix [12, 15, 26, 34, 48, 53]. However, in practice, these methods are often sub-optimal dealing with high noise ratio and complicated noise.

More recently, methods based on sample selection [19–21, 32, 33, 39, 42, 45, 58] to filter out samples with noisy labels become perhaps the dominant paradigm. For example, the most common sample selection strategy is the ‘small-loss’ mechanism motivated by the

memorization effect, that is, the model tends to fit clean samples earlier than noisy samples in the training process thus resulting in relatively smaller losses for the clean ones. Following this, most of methods focus primarily on improving sample selection mechanisms, including different variants of ‘small-loss’ strategy [1, 23, 49], and utilizing kNN [2, 7, 31] or graph models [46, 47] based on samples’ feature space for sample selection. However, these methods are inherently affected by the label noise as they still rely on the current in-training model, leading to the infamous self-confirmation bias. Some methods [13, 56] attempt to alleviate self-confirmation bias through model co-training, but this approach noticeably introduces additional computational overhead. Moreover, these methods solely rely on the visual information within the images, which can readily lead to biased sample selection outcomes, as exemplified in cases of ‘hard noise’ - noisy sample exhibits a highly visual similarity with its incorrectly labeled class, as illustrated in fig. 1.

To address the aforementioned issues, this paper proposes utilizing popular vision-language model - CLIP [35], for sample selection. Like any pre-trained models, CLIP is unaffected by the label noise in the collected dataset thus avoiding ‘self-confirmation’ bias. More importantly, CLIP’s distinctive language modality and zero-shot classifier allow us to compensate for the biases that may arise from solely relying on visual information for sample selection. For instance, this allows us to identify ‘hard noise’ (fig. 1) that is difficult to distinguish using only the vision modality.

To the best of our knowledge, we are the first to employ a large-scale vision-language model, particularly leveraging its language modality, for sample selection. Specifically, we simultaneously utilize CLIP’s zero-shot classifier and an easily-inducible classifier based on noisy labels and CLIP’s vision encoder. We name this method *CLIPCleaner* and theoretically and empirically demonstrate its effectiveness and unique advantages. To ensure the efficiency of the method and facilitate seamless comparison with existing methods, unlike common transfer learning techniques such as model fine-tuning [9], knowledge distillation [44], and prompt-based learning [3, 62], ***CLIPCleaner* does not involve training/fine-tuning the CLIP model.** Furthermore, to evaluate the performance of *CLIPCleaner* on existing datasets, **we introduce a minimal semi-supervised learning method tailored for noisy datasets, namely *MixFix*.** In detail, we gradually introduce ignored clean samples and re-label noisy samples to expand the initial clean subset selected by *CLIPCleaner*.

By leveraging *CLIPCleaner* and *MixFix* we establish a simple two-step framework for LNL: initiating with sample selection using *CLIPCleaner* and then perform semi-supervised learning solely using *MixFix*. Compared to existing methods involving iterations of sample selection and model training, **our approach features a simpler structure and aligns better with end-to-end training logic** when the noise information in the dataset is unknown. Moreover, ***CLIPCleaner* can serve as a plug-in module for existing methods**, which further shows great potential of CLIP in learning with noisy labels. **Despite its simplicity, our method achieves competitive and superior performance on various datasets**, including CIFAR10/CIFAR100 with synthetic noise (symmetric, asymmetric, and instance-dependent), as well as real-world noisy datasets like Red Mini-ImageNet, WebVision, Clothing1M, and ANIMAL-10N.

2 RELATED WORKS

Sample selection for noisy dataset. Most of the recent sample selection methods do so, by relying on the in-training model, for example the per-sample losses [1, 13, 18, 23] or model predictions [30, 38, 55]. A few works focus on further improving the sample selection quality by modelling the loss with markov process [49] or dynamically select samples with multiple metrics [63]. In addition to selecting samples based on the model classifier, some works also try to utilize the feature representations for sample selection. Wu et al. [46] and Wu et al. [47] try to build a kNN graph and identify clean samples through connected sub-graphs, while Feng et al. [7] and Ortego et al. [31] propose to utilize a simpler kNN in feature space to alleviate the effect of noisy labels. Some recent methods involving contrastive learning also identify clean sample pairs based on neighborhood relationships in the feature space [25] or fit Gaussian distributions to model the clean distribution [16]. However, these methods remain unstable and prone to self-confirmation bias, especially in strong noise scenarios, due to its intrinsic reliance on the in-training model based on noisy dataset.

Utilization of auxiliary model. To alleviate self-confirmation bias, the utilization of an auxiliary noise-free model is reasonable and straightforward. Related to us, some methods also try to use pre-trained noise-free models for learning with noisy labels. Cheng et al. [6], Zheltonozhskii et al. [61] propose to utilize self-supervised pre-training since it can learn good representations in the label-free case. Bahri et al. [2] utilize the pre-logit space of the pretrained model along with the kNN classifier for sample selection. Zhu et al. [65] follow the same idea and also involve CLIP, but they only utilize its vision encoder as a common pretrained encoder without utilizing the language encoder. We emphasize that language modality is critical as a supplementary modality.

3 METHOD

In section 3.1, we cast the learning with noisy labels problem in a formulation that covers mainstream sample selection methods. We also provide essential details about the CLIP model. In section 3.2, we elaborate our sample selection method, namely *CLIPCleaner*. In section 3.3, we introduce our semi-supervised learning method, namely *MixFix*. In section 3.4, we theoretically analyze the unique advantage of using CLIP for sample selection over common pre-trained models. In section 3.5, we provide further discussions on the topics of sample selection and the use of the CLIP model for LNL.

3.1 Preliminary

Sample selection with noisy labels. Given a dataset of training samples $(\mathbf{x}_i, y_i)_{i=1}^N$ i.i.d sampled from a noisy joint distribution $P(\mathbf{x}, y)$ with support as $\text{sup}(P) = \{\mathbf{x} \in \mathbb{R}^{C \times H \times W}, y \in \{1, \dots, K\}\}$ where K denotes the number of semantic classes, the goal of our method is to learn a classifier f that can accurately predict the true labels y for new, unseen examples. Let us denote the clean joint distribution as $P^{\text{true}}(\mathbf{x}, y)$. Most sample selection methods aim to approximate and optimize the unbiased empirical risk of f on the clean joint distribution $P^{\text{true}}(\mathbf{x}, y)$ with samples from noisy joint distribution $P(\mathbf{x}, y)$: $\hat{R}^{\text{true}}(f) = \frac{1}{N} \sum_{i=1}^N w_i L(\mathbf{x}_i, y_i; f)$, where w_i are the sample weights.

Particularly, with optimal weights ($w_i = P^{true}(y_i|x_i)/P(y_i|x_i)$) we can achieve risk-consistent learning¹. However, since $P^{true}(y_i|x_i)$ and $P(y_i|x_i)$ are typically both unknown for x_i , the objective of sample selection methods often revolves around estimating these two to subsequently estimate the optimal weights. In general, the noisy label y_i can serve as a confident proxy of the noisy distribution $P(y_i|x_i)$, making our focus on utilizing an additional auxiliary classifier $\tilde{P}(y_i|x_i)$ to estimate $P^{true}(y_i|x_i)$. In addition, it is commonly accepted to restrict the weights as binary since for most classification datasets, $P^{true}(y_i|x_i)$ tends to be highly centered around only one class. As a result, the optimal weight w_i usually leans towards either 0 or 1 for most samples. Here, we propose a concise form sufficient to comprehensively represent most existing sample selection methods:

$$\tilde{w}_i = \mathbb{G}(\tilde{P}(y_i|x_i), y_i) \in \{0, 1\}, \quad (1)$$

where $\mathbb{G}$ denotes a specific sample selection mechanism, such as the ‘small loss’ strategy, to further refining the estimation.

CLIP. We briefly introduce the CLIP model [35], which is currently one of the most prevalent vision-language models. CLIP aims to learn from a dataset of image-text pairs, denoted as $(x'_i, z_i)_{i=1}^M$ (we use x' here for the CLIP training images to discriminate from above in-question dataset), which is *i.i.d.* sampled from a hidden joint distribution $Q(x, z)$ with support as $\text{sup}(Q) = \{x \in \mathbb{R}^{C \times H \times W}, z \in \mathbb{R}^d\}$. We have below as CLIP training loss:

$$L(x'_i, z_i; g, h) = \frac{1}{2} \left(-\log \frac{\exp(g(x'_i)^T h(z_i))}{\sum_{j=1}^M \exp(g(x'_i)^T h(z_j))} - \log \frac{\exp(g(x'_i)^T h(z_i))}{\sum_{j=1}^M \exp(g(x'_j)^T h(z_i))} \right). \quad (2)$$

Here, g and h denote the vision and language encoder, respectively. Intuitively, the CLIP model tries to maximize the correspondence between related image-text pairs.

3.2 CLIPCleaner: sample selection with vision-language models

In this section, we propose a new sample selection method based on CLIP, namely *CLIPCleaner*. According to eq. (1), our method (actually nearly all sample selection methods) is divided into two main steps: 1. estimate $\tilde{P}(y_i|x_i)$; 2. calculate weight $\tilde{w}_i$ with specific $\mathbb{G}$. To enable the analysis between text z , image x and label y , we consistent the notations for CLIP’s training dataset and the in-question noisy dataset. Specifically, we extend the in-question noisy dataset to be *i.i.d.* sampled from $P(x, y, z)$ (actually from its marginalization), where $\text{sup}(P) = \{x \in \mathbb{R}^{C \times H \times W}, y \in [0, 1, \dots, K], z \in \mathbb{R}^d\}$; similarly, we extend the sampling distribution of CLIP’s training dataset to $Q(x, y, z)$. Here we assume $\text{sup}(P) \subset \text{sup}(Q)$.

3.2.1 Estimate $\tilde{P}(y_i|x_i)$. We consider two options for estimation: directly utilizing CLIP’s zero-shot classifier, or, ignoring CLIP’s language modality and treating its vision encoder as a regular pre-trained model and training a new classifier atop it with in-question noisy dataset.

¹Please refer to SUPPLEMENTARY E for details. We omit the variables for brevity, e.g., $P(y = y_i | x = x_i)$ as $P(y_i|x_i)$.

Option 1: Estimate $\tilde{P}(y_i|x_i)$ with CLIP zero-shot classifier. Firstly, we assume the causal mechanism for P and Q as: $x \rightarrow z \rightarrow y$ where z denotes the description text and y denotes the semantic label thus we have $y \perp x | z$. Roughly speaking, we assume that the semantic label y_i can be independently generated based on a decent image description z_i alone for each image x_i . We thus have:

$$\begin{aligned} \tilde{P}_{\text{zeroshot}}(y_i|x_i) &= \int Q(y_i|z_i)Q(z_i|x_i)dz \\ &\propto \int Q(y_i|z_i)Q(z_i, x_i)dz. \end{aligned} \quad (3)$$

Thus, we can estimate $\tilde{P}_{\text{zeroshot}}(y_i|x_i)$ with above integral by sampling z_i as long as $Q(z_i, x_i)$ and $Q(y_i|z_i)$ is known. Specifically, according to eq. (2), we show that $Q(z_i, x_i)$ can be estimated by the output similarity ($\exp(g(x_i)^T h(z_i))$) of the CLIP model (see SUPPLEMENTARY E). However, $Q(y_i|z_i)$ remains unknown and cannot be learned during the CLIP training process. Most current studies customarily design a single prompt as follows: ‘A photo of class name of y_i .’, implicitly assuming that:

$$Q(y_i|z = \text{‘A photo of class name of } y_i \text{.’}) \approx 1.$$

Here, single prompt actually corresponds to sampling a single z . Obviously, it is plausible that with more high-quality samplings of z_i instead of only utilizing one single prompt the estimation would be better. In this work, we propose below template to generate multiple prompts $\{\mathcal{P}_j\}_{j=1}^J$ using class-specific features²:

$$\mathcal{P}_j = \text{‘A photo of \{class name of } y_i\}, \text{ which is/has \{class-specific feature } j \text{ of class } y_i\}.’}$$

Then we can simplify eq. (3) with above prompts as below:

$$\tilde{P}_{\text{zeroshot}}(y_i|x_i) \propto \sum_{j=1}^J \tilde{Q}(z = \mathcal{P}_j, x_i). \quad (4)$$

Option 2: Estimate $\tilde{P}(y_i|x_i)$ with CLIP vision encoder and noisy dataset. By treating the CLIP model as an ordinary large-scale pre-trained model, we can also leverage its vision encoder g solely along with the in-question noisy dataset $(x_i, y_i)_{i=1}^N$ to train a new classifier f' (to discriminate it with the original classifier f in section 3.1) for estimation. With the common cross-entropy loss, it is straightforward that the normalized prediction logits serve as an estimate of $\tilde{P}(y|x)$:

$$\tilde{P}_{\text{trained}}(y_i|x_i) = \text{softmax}(f'(g(x_i)))_{y_i}. \quad (5)$$

By default, we train a *LogisticRegression* classifier as f' with fixed extracted features and noisy dataset. Empirically, we also consider non-parametric *kNN* in ablations Section 4.2.

3.2.2 Calculate weight w_i . With $\tilde{P}(y_i|x_i)$ estimated above, we can estimate weight w_i for each sample with any applicable sample selection mechanism $\mathbb{G}$. In this work, we consider two simple and popular mechanisms, named $\mathbb{G}_{\text{loss}}$ and $\mathbb{G}_{\text{consistency}}$. For $\mathbb{G}_{\text{loss}}$, we firstly model the per-sample cross-entropy losses ($\{-\log \tilde{P}(y = y_i|x_i)\}_{i=1}^N$) with GMM and then select samples by thresholding its probability belonging to the smaller component. Due to the possible class imbalances and the various semantic diversity of different classes, slightly different than the common approach utilizing a single GMM, we model the losses of samples from each class by a

²Please refer to SUPPLEMENTARY B for more details about how to generate prompts.

separate GMM model. Please refer to SUPPLEMENTARY D for specific comparisons on separate GMM and single GMM.

$$\mathbb{G}_{loss} = \mathbb{1}(\mathbb{P}(-\log \tilde{P}(y = y_i | x_i) \in \text{GMM}_{small}) \geq \theta_{loss}).$$

For $\mathbb{G}_{consistency}$, we calculate a consistency measure (defined as the ratio of the probability of noisy label class to the highest class probability) and select samples with high consistency:

$$\mathbb{G}_{consistency} = \mathbb{1}(\tilde{P}(y = y_i | x_i) / \max_k \tilde{P}(y = k | x_i) \geq \theta_{cons}).$$

3.3 MixFix: Efficient semi-supervised training by absorbing and relabelling

To evaluate our method on widely-acknowledged benchmarks, we propose a simple semi-supervised learning method for noisy dataset — namely *MixFix*. Please note, the notations employed in this section are defined independently. Specifically, we denote the selected subset and non-selected subset as $(\mathcal{X}_c, \mathcal{Y}_c)$ and $(\mathcal{X}_n, \mathcal{Y}_n)$. Motivated by pseudo-labelling [22] and FixMatch [37], we then inspect each sample's current prediction $\mathbf{p}_i$ in non-selected subset with:

$$(w_i, y_i) = \begin{cases} (0, y_i), & \text{if } p_m < \theta_r \text{ and } p_m < \theta'_r \text{ *Drop*} \\ (1, y_i), & \text{if } p_m > \theta_r \text{ and } y_i = y_m \text{ *Absorb*} \\ (1, y_m), & \text{if } p_m > \theta'_r \text{ and } y_i \neq y_m \text{ *Relabel*} \end{cases} \quad (6)$$

Here we denote as $p_m \triangleq \max_l \mathbf{p}_i(l)$ and $y_m \triangleq \arg \max_l \mathbf{p}_i(l)$. Intuitively, we 'absorb' ignored clean samples ($y_i = y_m$) and 'relabel' noisy samples ($y_i \neq y_m$) with different thresholds in non-selected subset, and progressively append it to initial selected subset to form a dynamic larger training set. Different from existing semi-supervised learning techniques, we typically set $\theta_r \leq \theta'_r$. This helps us make full use of noisy labels to differentiate the 'absorb' and 'relabel' process. To further counter the class imbalance in this new training set, the minority class is over-sampled. Then, we apply a common cross-entropy loss for training with this expanded and class-balanced training set, along with Mixup interpolation [57]. The detailed process is presented in Algorithm 1.

Please note, with selected subset only, *CLIPCleaner* can also be utilized along with existing methods - see SUPPLEMENTARY C for more results. The rationale behind formulating the *MixFix* method, tailored explicitly for noisy datasets, stems from our belief that in scenarios where noise information remains unknown, an end-to-end learning approach is not only more efficient but also stands out as an intuitive and primary choice. This is in contrast to the common style of iterative sample selection and model training.

Algorithm 1: MixFix.

Input: Selected subset $(\mathcal{X}_c, \mathcal{Y}_c)$, non-selected subset $(\mathcal{X}_n, \mathcal{Y}_n)$, θ_r, θ'_r , max epochs T

while $i < T$ **do**

Generate $(\mathcal{X}_r^i, \mathcal{Y}_r^i)$ with eq. (6);

Generate $(\mathcal{X}_t^i, \mathcal{Y}_t^i)$ with $(\mathcal{X}_r^i, \mathcal{Y}_r^i)$ and $(\mathcal{X}_c, \mathcal{Y}_c)$;

Minority over-sampling with $(\mathcal{X}_t^i, \mathcal{Y}_t^i)$;

Model training with $(\mathcal{X}_t^i, \mathcal{Y}_t^i)$ and MIXUP.

end

3.4 Theoretical justification of *CLIPCleaner*

Considering above proposed two options for *CLIPCleaner*, an immediate question is: how does the zero-shot classifier (eq. (4)) compare to the trained classifier (eq. (5)) in estimating $\tilde{P}(y|x)$. If the latter demonstrates comparable or even superior performance to the former, there may be little incentive to employ the CLIP model for sample selection. Rather, pursuing further enhancements to existing large-scale visual-only pre-trained models may yield greater potential. To this end, we conduct a theoretical analysis and compare the distances between the estimated $\tilde{P}(y_i|x_i)$ and true $P^{true}(y_i|x_i)$ of the two options. Specifically, following previous notations, we have below theorems:

THEOREM 3.1 (ESTIMATION WITH ZERO-SHOT CLASSIFIER). *Let $\mathcal{G}, \mathcal{H}$ be the hypothesis space of vision encoder g and language encoder h . Let us denote the rademacher complexity as $\mathfrak{R}(\mathcal{G} \circ \mathcal{H})$ of the combined CLIP model. Supposing the range of L from eq. (2) as $[0, l_\infty^{clip}]$ for all (x, z) in $\text{sup}(Q)$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ we have the following holds:*

$$d(\tilde{P}_{zeroshot}(y_i|x_i), P^{true}(y_i|x_i)) \leq \epsilon_{domain} + \Delta \left(\lambda_1 \mathfrak{R}(\mathcal{G} \circ \mathcal{H}) + \lambda_2 l_\infty^{clip} \sqrt{\frac{\log 1/\delta}{M}} + \lambda_3 \epsilon_n \right)$$

with $\lambda_1, \lambda_2, \lambda_3 > 0$. Here, ϵ_{domain} denotes the bias term induced by the domain gap between Q and P^{true} , and $\Delta \geq 1$ denotes the bias coefficient induced in designing prompts and sampling in eq. (3).

THEOREM 3.2 (ESTIMATION WITH TRAINED CLASSIFIER). *Let $\mathcal{F}$ be the hypothesis space of trained classifier f' . Let us denote the rademacher complexity as $\mathfrak{R}(\mathcal{F})$ of the trained classifier. Supposing the range of L for training f' as $[0, l_\infty^{noisy}]$ for all (x, y) in $\text{sup}(P)$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ we have the following holds:*

$$d(\tilde{P}_{trained}(y_i|x_i), P^{true}(y_i|x_i)) \leq \epsilon_{noise} + \lambda_1 \mathfrak{R}(\mathcal{F}) + \lambda_2 l_\infty^{noisy} \sqrt{\frac{\log 1/\delta}{N}}$$

with $\lambda_1, \lambda_2 > 0$. Here, ϵ_{noise} denotes the difference term induced by the distribution difference between P and P^{true} .

Please refer to SUPPLEMENTARY F for full derivation. With theorem 3.1 and theorem 3.2, ignoring the uncontrollable and common optimization bound error terms (marked in gray), we confirm that the zero-shot classifier estimation is highly related to domain gap and prompts quality while the trained classifier estimation is affected by the noise of in-question dataset, which is intuitively consistent with our expectation. We also empirically verify that the higher the noise ratio, the greater the performance advantage of zero-shot classifier over the trained classifier (section 4.2). More importantly, ϵ_{noise} is always inevitable while Δ can be easily improved with better prompt engineering and ϵ_{domain} can be also reduced by training CLIP with more abundant dataset and thus minimizing the domain gap.

3.5 Additional discussion

To be greedy or conservative? For all sample selection methods, an inevitable challenge is how to balance the precision and recall of

sample selection. In this paper, we introduce two different classifiers for estimation and two distinct sample selection strategies. Theoretical analysis and subsequent experiments indicate that these different classifiers and selection strategies exhibit their own preferences. In this study, we adopt a conservative sample selection strategy by taking the intersection of different sample selection outcomes, prioritizing the precision of sample selection. Compared to more greedy sample selection strategies, we lean towards relying on the semi-supervised learning strategy - *MixFix* to gradually introduce more samples into training. This can avoid magnifying the influence of noisy samples due to excessively greedy sample selection, but it also has obvious weaknesses, that some ‘hard’ clean samples will inevitably be missed. We leave the exploration of the optimal sample selection strategy to future work.

To fully explore CLIP? The utilization of the CLIP model for learning with noisy labels remains an area that requires further investigation. To ensure a fair comparison with existing work, we adopt standard sample selection paradigm, refraining from training or fine-tuning the CLIP model [3, 62]. In fact, the current prominent research directions related to CLIP involve fine-tuning the model, specifically through prompt-based learning. However, as expected, recent work (CoOp) has indicated that direct fine-tuning CLIP with noisy datasets can yield poorer performance compared to the initial zero-shot classifier. Therefore, in addition to sample selection, incorporating established techniques for LNL into prompt-based learning with CLIP may also offer promising directions.

4 EXPERIMENTS

4.1 Experiment details

4.1.1 Dataset details. **CIFAR10** and **CIFAR100** datasets comprise 50,000 images. Following established conventions, we assess our method’s performance with two types of artificial noise: “symmetric noise,” wherein labels are randomly flipped across all samples using a uniform distribution, and “asymmetric noise,” wherein labels of visually similar categories, such as Horse $\leftrightarrow$ Deer and Dog $\leftrightarrow$ Cat, are randomly interchanged. Moreover, we conduct experiments with various noise levels: 20%, 50%, 80% and 90% symmetric noise, as well as 40% asymmetric noise, adhering to the settings in DivideMix ([23]). For instance-dependent noise, we utilize the label noise file provided by [4].

Red Mini-ImageNet dataset [17] is a real-world dataset containing a total of 100 categories. It is an extension of the Mini-Imagenet dataset, where noise is introduced at varying ratios. Specifically, noisy images and their respective labels are obtained by crawling the internet, and these noisy images replace the original images in the Mini-ImageNet dataset, with different noise ratios. To ensure a fair comparison with previous studies [10, 52], the images are resized from their original size of 84×84 pixels to 32×32 pixels. Moreover, in accordance with the existing literature [10, 52], we utilize noise ratios of 20%, 40%, 60%, and 80%.

WebVision [27] is an extensive dataset comprising 1,000 classes of images obtained through web crawling. In line with previous studies [18, 23, 31], we evaluate our methods using the top 50 classes from the Google Subset of WebVision. The estimated noise ratio for this subset is approximately 20%.

ANIMAL-10N [38] is a recently introduced real-world noisy dataset comprises 10 classes of animals. The dataset has undergone manual labeling, with an estimated label noise ratio of around 8%. Similar to the CIFAR datasets, ANIMAL-10N consists of 50,000 training images and 10,000 test images.

Clothing1M [50] is a large-scale dataset containing 14 classes of clothing images, obtained by crawling online shopping websites. It consists of a substantial collection of 1 million noisy images. The estimated noise ratio for this dataset is approximately 38.5%.

4.1.2 Implementation details. We use CLIP model with ViT-B/32 backbone in all experiments except for specific ablations. In all experiments, our default approach is *CLIPCleaner* + *MixFix* (**Ours**). By default, we train the network with a SGD optimizer with a momentum of 0.9 in all experiments.

For **CIFAR10** and **CIFAR100**, we use a PresActResNet-18 [14] as the backbone in all experiments following previous works. For CIFAR10, we set $\theta_{loss} = 0.5$, $\theta_{cons} = 0.8$ for *CLIPCleaner* and $\theta_r = 0.8$, $\theta'_r = 0.9$ for *MixFix*; For CIFAR100, we set $\theta_{loss} = 0.5$, $\theta_{cons} = 0.8$ for *CLIPCleaner* and $\theta_r = 0.7$, $\theta'_r = 0.8$ for *MixFix*. We train both networks with for 300 epochs with a weight decay of 5e-4. The initial learning rate is 0.02 and is controlled by a cosine annealing scheduler. The batchsize is fixed as 128.

For **Red Mini-ImageNet**, we also use a PresActResNet-18 [14] as the backbone following previous works [10, 52]. For *CLIPCleaner*, we set $\theta_{loss} = 0.5$, $\theta_{cons} = 0.8$. For *MixFix*, we set $\theta_r = 0.8$, $\theta'_r = 0.95$. We train the network for 300 epochs with a weight decay of 5e-4. The initial learning rate is 0.02 and reduced by a factor of 10 after 200 and 250 epochs. The batchsize is fixed as 64.

For **WebVision**, we use a InceptionResNetv2 as the backbone following [23]. For *CLIPCleaner*, we set $\theta_{loss} = 0.5$, $\theta_{cons} = 1$. For *MixFix*, we set $\theta_r = 0.7$, $\theta'_r = 1.0$. We train the network for 150 epochs with a weight decay of 1e-4. The initial learning rate is 0.01 and reduced by a factor of 10 after 80 and 120 epochs. The batchsize is fixed as 32.

For **Clothing1M**, we use a ResNet50 as the backbone following [23] with ImageNet pretrained weights. For *CLIPCleaner*, we set $\theta_{loss} = 0$, $\theta_{cons} = 0.5$. For *MixFix*, we set $\theta_r = 0.7$, $\theta'_r = 1.0$. We train the network for 150 epochs with a weight decay of 1e-3. The initial learning rate is 0.002 and reduced by a factor of 10 after 50 and 100 epochs. The batchsize is fixed as 32.

For **ANIMAL-10N**, we use a VGG-19 [36] as the backbone with batch-normalization following [38]. For *CLIPCleaner*, we set $\theta_{loss} = 0.5$, $\theta_{cons} = 0.8$. For *MixFix*, we set $\theta_r = 0.7$, $\theta'_r = 0.99$. We train the network with SGD optimizer for 300 epochs with a momentum of 0.9 and weight decay of 5e-4. The initial learning rate is 0.02 and reduced by a factor of 10 after 150 and 250 epochs. The batchsize is fixed as 128.

4.2 Ablations study

Hyper-parameters w.r.t MixFix. In this section, we ablate on the only two hyperparameters of our semi-supervised training strategy *MixFix*: the ‘absorb’ threshold θ_r and the ‘relabel’ threshold θ'_r . Owing to the precision-recall dilemma when doing sample selection, here we also need to weigh the precision and recall when introducing additional training samples. In table 1 we demonstrate that under different noise ratios, a too high or too low threshold

Table 1: Ablations on MixFix with synthetic CIFAR100 noisy dataset. The *top-3* results are **bolded.**

θ_r	θ'_r	Noise ratio			
		20%	50%	80%	90%
0.7	0.7	76.46	74.69	69.50	62.91
	0.8	76.63	75.23	69.72	63.11
	0.9	77.06	75.17	67.76	59.17
0.8	0.7	75.49	74.30	67.95	63.29
	0.8	76.36	74.90	68.86	63.42
	0.9	76.66	74.50	67.37	58.09
0.9	0.7	74.53	73.49	68.74	62.22
	0.8	75.98	74.25	68.94	62.81
	0.9	75.78	74.23	67.17	59.38

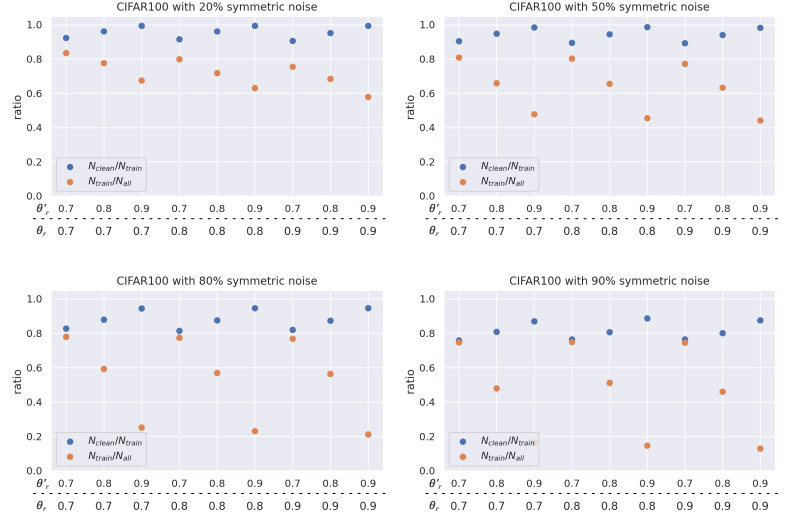


Figure 2: N_{train} denotes number of training samples, N_{clean} denotes number of clean training samples and N_{all} denotes number of clean training samples.

Table 2: Testing accuracy (%) with CLIP zero-shot classifier

Model	CIFAR10	CIFAR100	Red Mini-ImageNet	WebVision	Clothing1M	ANIMAL-10N
CLIP zero-shot	89.97	63.72	78.12	73.36	39.73	76.12
SOTA	92.68	67.7	49.55	80.9	74.84	84.6
Ours	95.15	71.17	54.21	81.56	74.87	88.14

leads to performance degradation, and $\theta_r < \theta'_r$ leads to better performance than setting same value for both thresholds. In fig. 2, we further reveal the inherent mechanism. Especially, after reducing the ‘absorb’ threshold θ'_r , the proportion of training samples increases and the accuracy of training samples decreases.

Analyzing CLIP Zero-shot classification as a baseline. In this section, we consider utilizing CLIP’s zero-shot classifier directly on the clean test set, following a procedure that we describe in Section 3.2. In table 2, we present the zero-shot classification results on six common benchmarks and compare them with current SOTA results as well as our own method. It’s worth noting that CLIP is utilized with the ViT-B/32 architecture here, while our method and the SOTA methods adopt simpler structures, such as PreResNet-18 for the CIFAR dataset. Therefore, this comparison is indeed ‘over stringent’. Even though, we observe that, when compared to directly utilizing CLIP’s zero-shot classifier, our method delivers significantly improvements on most datasets and outperforms the SOTA LNL methods on all datasets. We also consider other vision-language models other than CLIP in SUPPLEMENTARY A.

Analyzing sample selection w.r.t different classifiers and different mechanisms. In section 3.4, we theoretically conclude that the performance of the zero-shot classifier is influenced by the quality of utilized prompts and the domain gap between CLIP training dataset

and the in-question noisy dataset, while the performance of the easily-inducible classifier trained based on CLIP’s vision encoder and the in-question noisy dataset is influenced by the noise of the in-question dataset. To validate this, we empirically test with two datasets with controllable noise ratios, that is, the CIFAR10/100 dataset with synthetic noise and the Red Mini-ImageNet dataset with real-world noise.

In fig. 3, we show the sample selection performance and find that: i) As the noise ratio increases, regardless of the dataset, noise types, CLIP backbones or empirical variants of the trained classifier in option 2 (*LogisticRegression* VS *kNN*), the zero-shot classifier (option 1) gradually outperforms the trained classifier. This further validates our theoretical findings in section 3.4; ii) Additionally, we notice that when comparing two different modes for obtaining the training classifier, the *LogisticRegression* classifier empirically exhibits superior performance to the *kNN* classifier. Therefore, we choose the *LogisticRegression* classifier as our default choice for trained classifier; iii) Furthermore, we find that different sample selection mechanisms ($\mathbb{G}_{consistency}$ VS $\mathbb{G}_{loss}$) show distinct advantages and disadvantages on different datasets. Given that noise information is typically unknown in real-world scenarios, as analyzed in section 3.5, we default to a conservative sample selection strategy, which involves utilizing both sample selection strategies and choosing their intersection as final selected subset.

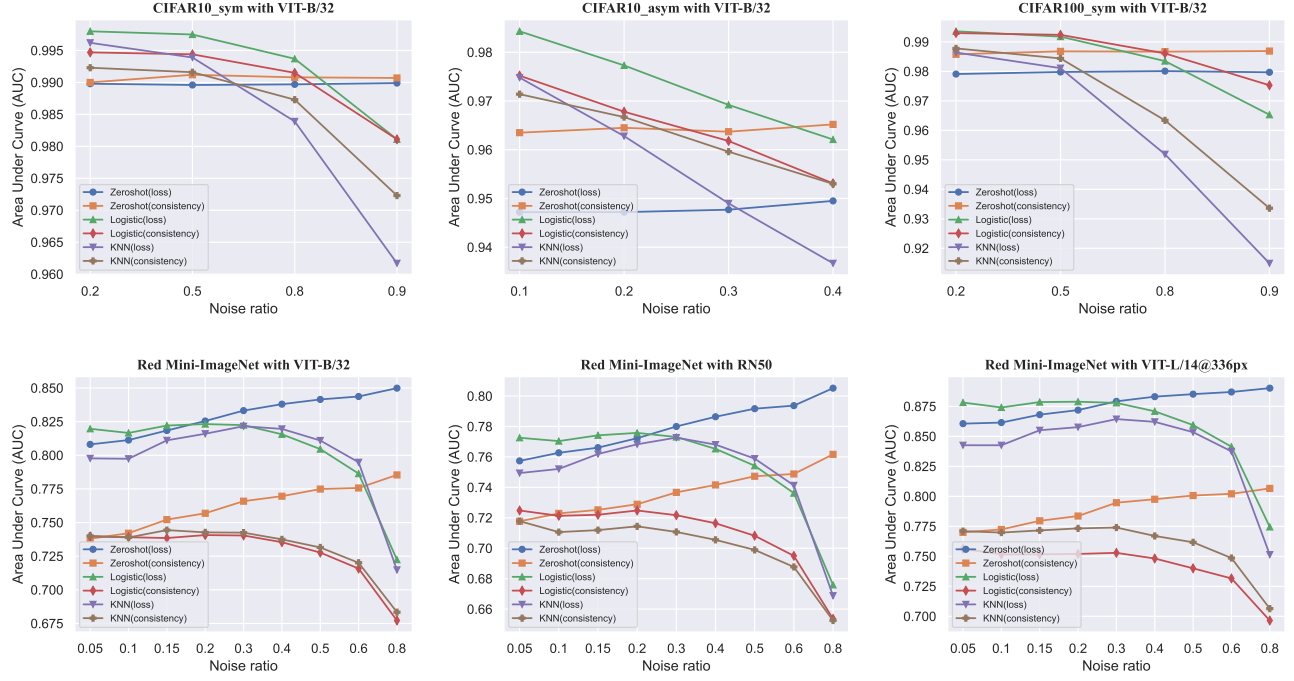


Figure 3: Comparisons of various sample selection methods w.r.t different dataset/noise type/noise ratio. Here, we show the ROC AUC score of binary identification of clean samples.

Table 3: Testing accuracy (%) on on CIFAR-10 and CIFAR-100 with synthetic noise.

Dataset	CIFAR10					CIFAR100			
	Symmetric		Assymmetric			Symmetric			
Noise type									
Noise ratio	20%	50%	80%	90%	40%	20%	50%	80%	90%
CE	86.8	79.4	62.9	42.7	85.0	62.0	46.7	19.9	10.1
Co-teaching+ [56]	89.5	85.7	67.4	47.9	-	65.6	51.8	27.9	13.7
F-correction [34]	86.8	79.8	63.3	42.9	87.2	61.5	46.6	19.9	10.2
PENCIL [55]	92.4	89.1	77.5	58.9	88.5	69.4	57.5	31.1	15.3
LossModelling [1]	94.0	92.0	86.8	69.1	87.4	73.9	66.1	48.2	24.3
DivideMix [23]	96.1	94.6	93.2	76.0	93.4	77.3	74.6	60.2	31.5
ELR+ [28]	95.8	94.8	93.3	78.7	93.0	77.6	73.6	60.8	33.4
MOIT [31]	93.1	90.0	79.0	69.6	92.0	73.0	64.6	46.5	36.0
SeLCL+ [25]	95.5	93.9	89.2	81.9	93.4	76.5	72.4	59.6	48.8
TCL [16]	95.0	93.9	92.5	89.4	92.6	78.0	73.3	65.0	54.5
Ours	95.92±0.15	95.67±0.28	95.04±0.37	94.23±0.54	94.89±0.16	78.20±0.45	75.23±0.29	69.72±0.61	63.11±0.89

4.3 Results on synthetic noisy dataset

In this section, we firstly evaluate our method on the CIFAR datasets with synthetic symmetric/asymmetric noise. In table 3, We can see that our method gets competitive and better performance in all experiment settings, especially when the noise ratio is high (63.11% testing accuracy with 90% symmetric noise on CIFAR100 dataset). Also, we would like to emphasize that we keep hyper-parameters fixed for all experiments here as we believe the method robustness in a noise agnostic scenario is critical.

To further validate the performance of our method in handling the ‘hard noise’, we also conduct experiments on instance-dependent noise in table 5. Different from symmetric or asymmetric noise, instance-dependent noise assumes that semantic-similar samples are more prone to get mislabelled, aligning better with our earlier definition of ‘hard noise’. Besides, here we here exclude *MixFix* and employ the selected samples for training with cross-entropy loss solely. This exclusion serves to provide an additional proof of the superior sample selection performance of *CLIPCleaner*.

Table 4: Testing accuracy (%) on Clothing1M.

CE	F-correction [34]	RRL [24]	C2D [61]	DivideMix [23]	ELR+ [28]	SSR+ [7]	TCL [16]	Ours	Ours (Co-training)	CLIPCleaner + DivideMix
69.21	69.84	74.30	74.84	74.76	74.81	74.83	74.80	73.41±0.65	74.01±0.47	74.87±0.44

Table 5: Testing accuracy (%) on CIFAR10 with instance-dependent noise.

Method	Noise ratio			
	10%	20%	30%	40%
CE	91.25	86.34	80.87	75.68
F-correction [34]	91.06	86.35	78.87	71.12
Co-teaching [13]	91.22	87.28	84.33	78.72
GCE [60]	90.97	86.44	81.54	76.71
DAC [40]	90.94	86.16	80.88	74.80
DMI [51]	91.26	86.57	81.98	77.81
SEAL [4]	91.32	87.79	85.30	82.98
CE*	90.76	86.08	80.64	75.27
CLIPCleaner + CE	92.33±0.37	91.06±0.37	89.71±0.37	88.26±0.37

Table 7: Testing accuracy (%) on on Red Mini-ImageNet.

Method	Noise ratio			
	20%	40%	60%	80%
CE	47.36	42.70	37.30	29.76
Mixup [57]	49.10	46.40	40.58	33.58
DivideMix [23]	50.96	46.72	43.14	34.50
MentorMix [17]	51.02	47.14	43.80	33.46
FaMUS [52]	51.42	48.06	45.10	35.50
InstanceGM [10]	58.38	52.24	47.96	39.62
Ours	61.44±0.45	58.42±0.66	53.18±0.47	43.82±0.87

Table 8: Testing accuracy (%) on ANIMAL-10N.

Method	Accuracy
CE	79.4
SELFIE [38]	81.8
PLC [59]	83.4
NCT [5]	84.1
InstanceGM [10]	84.6
SSR+ [7]	88.5
Ours	88.85±0.61

4.4 Results on real-world noisy datasets

Finally, in table 6, table 7, and table 8 we show results on the ANIMAL-10N, Red Mini-ImageNet and WebVision datasets, respectively. In summary, our proposed method demonstrates substantial improvements compared to the current state-of-the-art approaches on both large-scale web-crawled datasets and small-scale human-annotated noisy datasets.

Table 6: Testing accuracy (%) on on WebVision.

Methods	WebVision		ILSVRC2012	
	Top1	Top5	Top1	Top5
Co-teaching [13]	63.5	85.20	61.48	84.70
DivideMix [23]	77.32	91.64	75.20	90.84
ELR+ [28]	77.78	91.68	70.29	89.76
NGC [47]	79.16	91.84	74.44	91.04
FaMUS [52]	79.4	92.8	77.0	92.8
RRL [24]	76.3	91.5	73.3	91.2
SelCL+ [25]	79.9	92.6	76.8	93.0
SSR+ [7]	80.9	92.8	75.8	91.8
TCL [16]	79.1	92.3	75.4	92.4
Ours	81.56±0.29	93.26±0.65	77.80±0.25	92.08±0.44

We note, that the proposed *CLIPCleaner* can also be used in combination with other schemes. In table 4 we show results on the Clothing1M dataset both with our default setting (*CLIPCleaner* + *MixFix*) and with it incorporated to two additional schemes: first incorporating our method with co-training, and second replacing *MixFix* with DivideMix [23]. We observe that we obtain results that are superior to the current state-of-the-art. Meanwhile, we would like to note that the majority of existing methods have small differences on the Clothing1M dataset despite the fact that they have large performance differences on other datasets. This suggests that additional training techniques may have a greater impact than

sample selection methods on this specific dataset, possibly due to the fact that the Clothing1M dataset is more fine-grained than other datasets. For such fine-grained noisy datasets, sample selection may not be the optimal strategy, as suggested in Section 3.1, where the basis of sample selection methods relies on highly concentrated conditional probabilities for the samples (eq. (1)).

5 CONCLUSION

To mitigate the issues of ‘self-confirmation bias’ and compensate for visual-only modality in current mainstream sample selection methods, in this paper we propose a method utilizing the large-scale vision-language model CLIP for sample selection, called *CLIPCleaner*. We substantiate its effectiveness through both theoretically and empirically. Furthermore, we introduce a straightforward semi-supervised learning method tailored for noisy datasets, called *MixFix*, without the need for intricate off-the-shelf techniques. We emphasize that the exploration of utilizing vision-language models for noisy datasets, such as the potential of existing prompt learning techniques, remains an open direction. Additionally, the possibility of a large domain gap between the CLIP model and the target dataset can influence results, indicating a need for more refined vision-language models. Lastly, our experiments suggest that sample selection methods may not be optimal for fine-grained noisy datasets, which presents itself also as one of our future research directions.

REFERENCES

- [1] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. 2019. Unsupervised label noise modeling and loss correction. In *International Conference on Machine Learning*. PMLR, 312–321.
- [2] Dara Bahri, Heinrich Jiang, and Maya Gupta. 2020. Deep k-nn for noisy labels. In *International Conference on Machine Learning*. PMLR, 540–550.
- [3] Guangyi Chen, Weiran Xiao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. 2022. PLOT: Prompt Learning with Optimal Transport for Vision-Language Models. In *The Eleventh International Conference on Learning Representations*.
- [4] Pengfei Chen, Junjie Ye, Guangyong Chen, Jingwei Zhao, and Pheng-Ann Heng. 2021. Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 11442–11450.
- [5] Yingyi Chen, Xi Shen, Shell Xu Hu, and Johan AK Suykens. 2021. Boosting Co-teaching with Compression Regularization for Label Noise. *arXiv preprint arXiv:2104.13766* (2021).
- [6] Hao Cheng, Zhaowei Zhu, Xing Sun, and Yang Liu. 2021. Demystifying how self-supervised features improve training from noisy labels. *arXiv preprint arXiv:2110.09022* (2021).
- [7] Chen Feng, Georgios Tzimiropoulos, and Ioannis Patras. 2022. SSR: An Efficient and Robust Framework for Learning with Unknown Label Noise. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press.
- [8] Lei Feng, Senlin Shu, Zhuoyi Lin, Fengmao Lv, Li Li, and Bo An. 2021. Can cross entropy loss be robust to label noise?. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*. 2206–2212.
- [9] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2021. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544* (2021).
- [10] Arpit Garg, Cuong Nguyen, Rafael Felix, Thanh-Toan Do, and Gustavo Carneiro. 2023. Instance-Dependent Noisy Label Learning via Graphical Modelling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2288–2298.
- [11] Aritra Ghosh, Himanshu Kumar, and PS Sastry. 2017. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.
- [12] Jacob Goldberger and Ehud Ben-Reuven. 2016. Training deep neural-networks using a noise adaptation layer. (2016).
- [13] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *arXiv preprint arXiv:1804.06872* (2018).
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity mappings in deep residual networks. In *European conference on computer vision*. Springer, 630–645.
- [15] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. 2018. Using trusted data to train deep networks on labels corrupted by severe noise. *arXiv preprint arXiv:1802.05300* (2018).
- [16] Zhizhong Huang, Junping Zhang, and Hongming Shan. 2023. Twin Contrastive Learning with Noisy Labels. In *CVPR*.
- [17] Lu Jiang, Di Huang, Mason Liu, and Weilong Yang. 2020. Beyond synthetic noise: Deep learning on controlled noisy labels. In *International conference on machine learning*. PMLR, 4804–4815.
- [18] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. 2018. Mentor-net: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*. PMLR, 2304–2313.
- [19] Nazmul Karim, Mamshad Nayeem Rizve, Nazanin Rahnavard, Ajmal Mian, and Mubarak Shah. 2022. Unicon: Combating label noise through uniform selection and contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9676–9686.
- [20] Jihye Kim, Aristide Baratin, Yan Zhang, and Simon Lacoste-Julien. 2023. CrossSplit: mitigating label noise memorization through data splitting. In *International Conference on Machine Learning*. PMLR, 16377–16392.
- [21] Jang-Hyun Kim, Sangdoo Yun, and Hyun Oh Song. 2024. Neural Relation Graph: A Unified Framework for Identifying Label Noise and Outlier Data. *Advances in Neural Information Processing Systems* 36 (2024).
- [22] Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, Vol. 3. 896.
- [23] Junnan Li, Richard Socher, and Steven CH Hoi. 2020. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394* (2020).
- [24] Junnan Li, Caiming Xiong, and Steven Hoi. 2020. Learning from Noisy Data with Robust Representation Learning. (2020).
- [25] Shikun Li, Xiaobo Xia, Shiming Ge, and Tongliang Liu. 2022. Selective-supervised contrastive learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 316–325.
- [26] Shikun Li, Xiaobo Xia, Hansong Zhang, Yibing Zhan, Shiming Ge, and Tongliang Liu. 2022. Estimating noise transition matrix with label correlations for noisy multi-label learning. *Advances in Neural Information Processing Systems* 35 (2022), 24184–24198.
- [27] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. 2017. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862* (2017).
- [28] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. 2020. Early-learning regularization prevents memorization of noisy labels. *arXiv preprint arXiv:2007.00151* (2020).
- [29] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. 2020. Normalized loss functions for deep learning with noisy labels. In *International conference on machine learning*. PMLR, 6543–6553.
- [30] Eran Malach and Shai Shalev-Shwartz. 2017. Decoupling" when to update" from" how to update". *arXiv preprint arXiv:1706.02613* (2017).
- [31] Diego Ortego, Eric Arazo, Paul Albert, Noel O'Connor, and Kevin McGuinness. 2021. Multi-Objective Interpolation Training for Robustness to Label Noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6606–6615.
- [32] Dongmin Park, Seola Choi, Doyoung Kim, Hwanjun Song, and Jae-Gil Lee. 2024. Robust data pruning under label noise via maximizing re-labeling accuracy. *Advances in Neural Information Processing Systems* 36 (2024).
- [33] Deep Patel and PS Sastry. 2023. Adaptive sample selection for robust learning under label noise. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3932–3942.
- [34] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. 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.
- [35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [36] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [37] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685* (2020).
- [38] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. 2019. Selfie: Refurbishing unclean samples for robust deep learning. In *International Conference on Machine Learning*. PMLR, 5907–5915.
- [39] Zeren Sun, Fumin Shen, Dan Huang, Qiong Wang, Xiangbo Shu, Yazhou Yao, and Jinhui Tang. 2022. Pnp: Robust learning from noisy labels by probabilistic noise prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5311–5320.
- [40] Sunil Thulasidasan, Tanmoy Bhattacharya, Jeff Bilmes, Gopinath Chennupati, and Jamal Mohd-Yusof. 2019. Combating label noise in deep learning using abstention. *arXiv preprint arXiv:1905.10964* (2019).
- [41] Deng-Bao Wang, Yong Wen, Lujia Pan, and Min-Ling Zhang. 2021. Learning from noisy labels with complementary loss functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 10111–10119.
- [42] Haobo Wang, Ruixuan Xiao, Yiwen Dong, Lei Feng, and Junbo Zhao. 2022. ProMix: combating label noise via maximizing clean sample utility. *arXiv preprint arXiv:2207.10276* (2022).
- [43] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. 2019. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 322–330.
- [44] Zhecan Wang, Noel Codella, Yen-Chun Chen, Luowei Zhou, Jianwei Yang, Xiyang Dai, Bin Xiao, Haoxuan You, Shih-Fu Chang, and Lu Yuan. 2022. Clip-td: Clip targeted distillation for vision-language tasks. *arXiv preprint arXiv:2201.05729* (2022).
- [45] Qi Wei, Haoliang Sun, Xiankai Lu, and Yilong Yin. 2022. Self-filtering: A noise-aware sample selection for label noise with confidence penalization. In *European Conference on Computer Vision*. Springer, 516–532.
- [46] Pengxiang Wu, Songzhu Zheng, Mayank Goswami, Dimitris Metaxas, and Chao Chen. 2020. A topological filter for learning with label noise. *Advances in neural information processing systems* 33 (2020), 21382–21393.
- [47] Zhi-Fan Wu, Tong Wei, Jianwen Jiang, Chaojie Mao, Mingqian Tang, and Yu-Feng Li. 2021. NGC: A Unified Framework for Learning with Open-World Noisy Data. *arXiv preprint arXiv:2108.11035* (2021).
- [48] Xiaobo Xia, Bo Han, Nannan Wang, Jiansong Deng, Jiatong Li, Yinian Mao, and Tongliang Liu. 2022. Extended T: Learning With Mixed Closed-Set and Open-Set Noisy Labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 3 (2022), 3047–3058.
- [49] Xiaobo Xia, Tongliang Liu, Bo Han, Mingming Gong, Jun Yu, Gang Niu, and Masashi Sugiyama. 2021. Sample selection with uncertainty of losses for learning with noisy labels. *arXiv preprint arXiv:2106.00445* (2021).

- [50] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. 2015. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2691–2699.
- [51] Yilun Xu, Peng Cao, Yuqing Kong, and Yizhou Wang. 2019. L_{DMI}: A novel information-theoretic loss function for training deep nets robust to label noise. *Advances in neural information processing systems* 32 (2019).
- [52] Youjiang Xu, Linchao Zhu, Lu Jiang, and Yi Yang. 2021. Faster meta update strategy for noise-robust deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 144–153.
- [53] Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. 2020. Dual t: Reducing estimation error for transition matrix in label-noise learning. *Advances in neural information processing systems* 33 (2020), 7260–7271.
- [54] Xichen Ye, Xiaoqiang Li, Tong Liu, Yan Sun, Weiqin Tong, et al. 2024. Active Negative Loss Functions for Learning with Noisy Labels. *Advances in Neural Information Processing Systems* 36 (2024).
- [55] Kun Yi and Jianxin Wu. 2019. Probabilistic end-to-end noise correction for learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7017–7025.
- [56] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. 2019. How does disagreement help generalization against label corruption?. In *International Conference on Machine Learning*. PMLR, 7164–7173.
- [57] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412* (2017).
- [58] HaiYang Zhang, XiMing Xing, and Liang Liu. 2021. Dualgraph: A graph-based method for reasoning about label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9654–9663.
- [59] Yikai Zhang, Songzhu Zheng, Pengxiang Wu, Mayank Goswami, and Chao Chen. 2021. Learning with Feature-Dependent Label Noise: A Progressive Approach. *arXiv preprint arXiv:2103.07756* (2021).
- [60] Zhilu Zhang and Mert R Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *arXiv preprint arXiv:1805.07836* (2018).
- [61] Evgenii Zheltonozhskii, Chaim Baskin, Avi Mendelson, Alex M Bronstein, and Or Litany. 2021. Contrast to Divide: Self-Supervised Pre-Training for Learning with Noisy Labels. *arXiv preprint arXiv:2103.13646* (2021).
- [62] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision* 130, 9 (2022), 2337–2348.
- [63] Tianyi Zhou, Shengjie Wang, and Jeff Bilmes. 2020. Robust curriculum learning: from clean label detection to noisy label self-correction. In *International Conference on Learning Representations*.
- [64] Xiong Zhou, Xianming Liu, Junjun Jiang, Xin Gao, and Xiangyang Ji. 2021. Asymmetric loss functions for learning with noisy labels. In *International conference on machine learning*. PMLR, 12846–12856.
- [65] Zhaowei Zhu, Zihao Dong, and Yang Liu. 2022. Detecting corrupted labels without training a model to predict. In *International Conference on Machine Learning*. PMLR, 27412–27427.