CLIPCleaner: Cleaning Noisy Labels with CLIP

Chen Feng
Queen Mary University of London
London, UK
chen.feng@qmul.ac.uk

Georgios Tzimiropoulos
Queen Mary University of London
London, UK
g.tzimiropoulos@qmul.ac.uk

Ioannis Patras
Queen Mary University of London
London, UK
i.patras@qmul.ac.uk

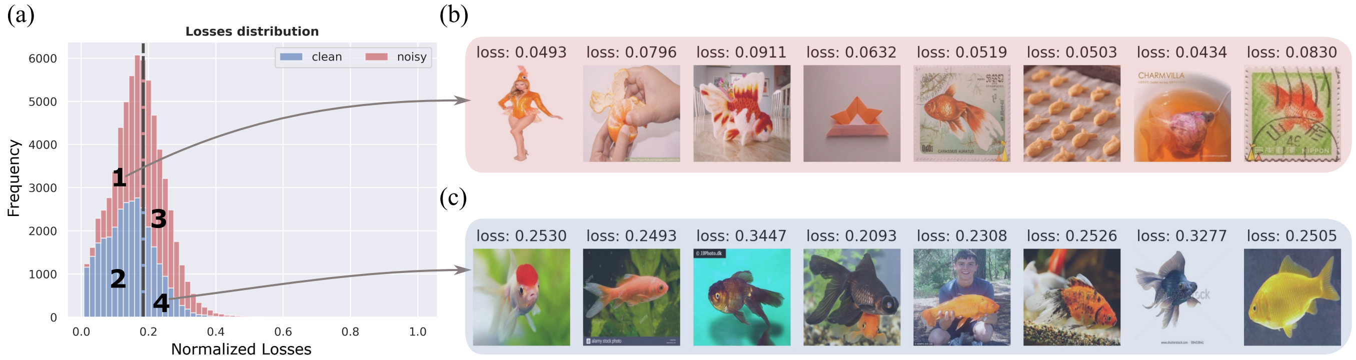

Figure 1: The normalized losses distribution of WebVision dataset after one epoch warm-up training, i.e., training with whole dataset and cross-entropy loss. In (a), 'clean'/'noisy' denotes samples been identified as clean/noisy by *CLIPCleaner* while the 'gray vertical line' denotes the sample selection boundary induced by 'small-loss' mechanism. We show some example images on *part 1* in (b) and *part 4* in (c) which represents samples identified as 'clean' by 'small-loss' while rejected by *CLIPCleaner* and vice versa. For example, in (b) we can find that many images with small losses due to its similar color or textures to 'tench' class, thus been wrongly identified as 'clean' by 'small-loss' but been correctly rejected by *CLIPCleaner*.

Abstract

Learning with Noisy labels (LNL) poses a significant challenge for the Machine Learning community. Some of the most widely used approaches that select as clean samples for which the model itself (the in-training model) has high confidence, e.g., 'small loss', can suffer from the so called 'self-confirmation' bias. This bias arises because the in-training model, is at least partially trained on the noisy labels. Furthermore, in the classification case, an additional challenge arises because some of the label noise is between classes that are visually very similar ('hard noise'). This paper addresses these challenges by proposing a method (*CLIPCleaner*) that leverages CLIP, a powerful Vision-Language (VL) model for constructing a zero-shot classifier for efficient, offline, clean sample selection. This has the advantage that the sample selection is decoupled from the in-training model and that the sample selection is aware of the semantic and visual similarities between the classes due to the way that CLIP is trained. We provide theoretical justifications and empirical evidence to demonstrate the advantages of CLIP for LNL compared to conventional pre-trained models. Compared to current methods that combine iterative sample selection with various techniques, *CLIPCleaner* offers a simple, single-step approach that achieves competitive or superior performance on benchmark datasets. To the best of our knowledge, this is the first time a VL model has been used for sample selection to address the problem of Learning with Noisy Labels (LNL), highlighting their potential in the domain.

CCS Concepts

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

Keywords

Sample selection, Noisy Labels, CLIP

ACM Reference Format:
Chen Feng, Georgios Tzimiropoulos, and Ioannis Patras. 2024. CLIPCleaner: Cleaning Noisy Labels with CLIP. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28-November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3664647.3680664

1 Introduction

Over the past two decades, deep neural networks have demonstrated exceptional success in various vision tasks, partly due to the existence of accurately labelled, large-scale datasets such as ImageNet-1K. However, collecting high-quality labels for such datasets is generally time-consuming and labour-intensive. Noisy labels, stemming from human error, ambiguity in labelling criteria,

MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3680664

or inherent noise in data collection processes, introduce a critical challenge that traditional learning algorithms must deal with.

To learn with noisy labels (LNL), various methods have been proposed. Some methods aim to develop robust loss functions [12, 17, 35, 48, 50, 61, 67, 71] or model the labeling error patterns with a label transition matrix [18, 21, 32, 41, 55, 60]. However, these methods are often sub-optimal in dealing with high noise ratios and complicated noise patterns.

More recently, methods based on sample selection [25–27, 39, 40, 46, 49, 52, 65] that aim to identify samples with clean labels have become perhaps the dominant paradigm. Among them, the most common sample selection strategies are the 'small-loss' mechanism motivated by the fact that the model tends to fit clean samples earlier than noisy samples in the training process – this results in relatively smaller losses for the clean samples. Following this, most of methods focus primarily on further improving such sample selection mechanisms. This includes different variants of the 'small-loss' strategy [1, 29, 56], and utilizing kNN [2, 10, 38] or graph models [53, 54] based on the samples' feature space for sample selection. However, these methods are inherently affected by the label noise, as losses, or the features used for sample selections are extracted from the model that is being trained (i.e., the in-training model) – this leads to the infamous 'self-confirmation' bias. Some methods [19, 63] attempt to *alleviate* 'self-confirmation' bias through model co-training, but this approach introduces additional computational overhead. Moreover, these methods solely rely on the visual information within the images, and therefore have difficulty dealing with 'hard noise', that is labelling errors between classes with high visual similarity.

To address the aforementioned issue, we propose a novel method, namely *CLIPCleaner*, that leverages the popular visual-language model CLIP [42] for sample selection. Specifically, we propose using a CLIP-based zero-shot classifier with descriptive class prompts that are generated automatically using a Large Language Model for sample selection. Given that CLIP is trained with massive vision-language pairs, this leads to a sample selection scheme that has two advantages: 1. the sample selection is aware of visual and semantic similarities between the classes and therefore compensates for biases that may arise from relying solely on visual information for sample selection (fig. 1); 2. the sample selection is independent of the in-training model, and therefore immune to the influence of noisy labels and the 'self-confirmation' bias. **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.**

Furthermore, we introduce a very simple semi-supervised learning method tailored for noisy datasets without common advanced modules such as co-training or multi-task training, namely *MixFix*. The proposed semi-supervised method, gradually introduces more clean samples and re-labels noisy samples to expand the initial clean subset selected by *CLIPCleaner*. Let us note that in the proposed scheme, the in-training model, i.e., the final classifier, is different from the VL model that is used for sample selection. More specifically, unlike common transfer learning techniques such as model fine-tuning [13], knowledge distillation [51], and prompt-based learning [3, 69], we adhere to using CLIP solely for sample selection and refrain from training/fine-tuning it. This has the distinct advantage that the proposed scheme allows for computationally, or parameter-wise light in-training model, and allows the use as sample selectors of VL models to which one does not necessarily have full access.

We demonstrate the effectiveness and advantages of the proposed method both theoretically and empirically. 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 learning with noisy labels. Most sample selection methods usually rely on model classifiers, such as the widely-applied 'small-loss' mechanism [1, 19, 24, 29] or model predictions [36, 44, 62]. More recent works focus on further improving the sample selection quality by modelling the loss with markov process [56] or dynamically selecting samples with multiple metrics [70]. In addition, some works try to utilize the feature representations for sample selection. Wu et al. [53] and Wu et al. [54] try to build a kNN graph and identify clean samples through connected sub-graphs, while Feng et al. [10, 11], Ortego et al. [38] propose to utilize a kNN in feature space to alleviate the effect of noisy labels. Some recent methods involving contrastive learning also identify clean sample pairs based on neighbourhood relationships in the feature space [31] or fit Gaussian distributions to model the clean distribution [22]. However, these methods remain unstable and prone to 'self-confirmation' bias, especially in high-ratio noise scenarios, due to their intrinsic reliance on the in-training model based on noisy datasets.

Utilization of auxiliary model. The utilization of an auxiliary noise-free model is reasonable and straightforward for LNL. Related to us, some methods also try to use pre-trained noise-free models for learning with noisy labels. Cheng et al. [7], Zheltonozhskii et al. [68] propose to utilize self-supervised learning [5, 8, 9, 14, 15, 20, 37, 45] since it can learn good representations in the label-free case. Bahri et al. [2] utilize the pre-logit space of the pre-trained model along with the kNN classifier for sample selection. Zhu et al. [72] follow the same idea and also involve CLIP, but they only utilize its vision encoder as a common pre-trained encoder without utilizing the language encoder. In this work, we emphasize that language modality is critical as a supplementary modality and show the unique advantage of VL models for sample selection, both theoretically and empirically.

3 Method

In section 3.1, we cast the learning with noisy labels problem in a formulation that covers mainstream sample selection methods. 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 pretrained models. In section 3.5, we provide further discussions on the topics of sample selection and the usage of the CLIP model.

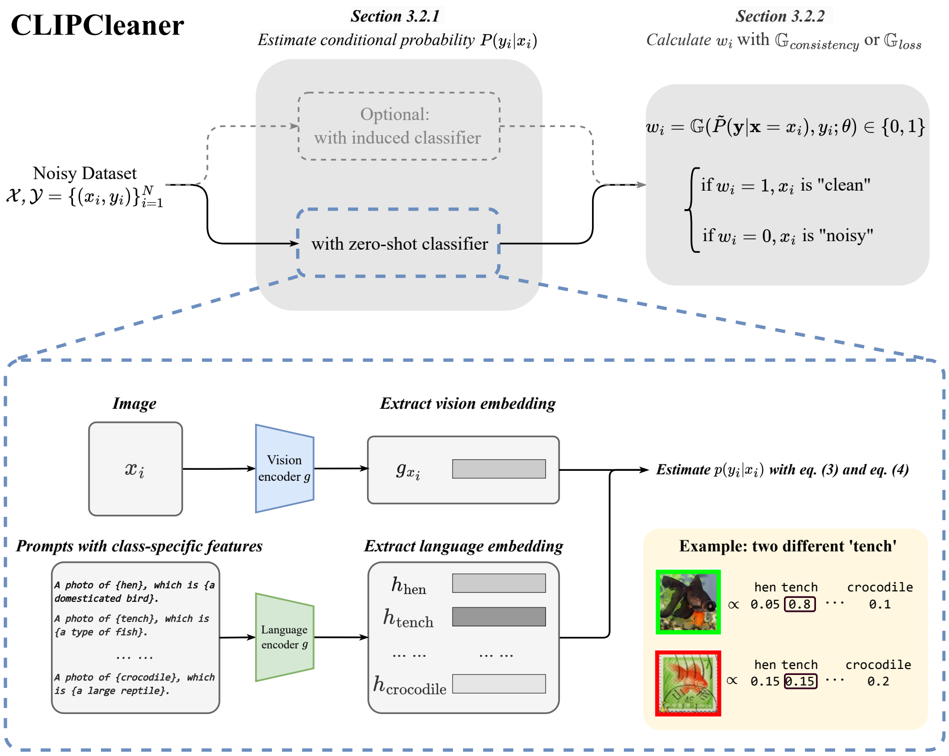
CLIPCleaner
Section 3.2.1
Estimate conditional probability $P(y_i|x_i)$
Section 3.2.2
Calculate w_i with $\mathbb{G}_{consistency}$ or $\mathbb{G}_{loss}$
Optional:
with induced classifier
Noisy Dataset
$\mathcal{X}, \mathcal{Y} = \{(x_i, y_i)\}_{i=1}^N$
with zero-shot classifier
$w_i = \mathbb{G}(\tilde{P}(\mathbf{y}|\mathbf{x}=x_i), y_i; \theta) \in \{0,1\}$
if $w_i = 1, x_i$ is "clean"
if $w_i = 0, x_i$ is "noisy"
Image
x_i
Vision encoder g
Extract vision embedding
g_{x_i}
Estimate $p(y_i|x_i)$ with eq. (3) and eq. (4)
Prompts with class-specific features
A photo of {hen}, which is {a domesticated bird}.
A photo of {tench}, which is {a type of fish}.
... ...
A photo of {crocodile}, which is {a large reptile}.
Language encoder g
Extract language embedding
h_{hen}
h_{tench}
... ...
$h_{\text{crocodile}}$
Example: two different 'tench'
hen tench crocodile
∝ 0.05 0.8 ··· 0.1
hen tench crocodile
∝ 0.15 0.15 ··· 0.2

Figure 2: Workflow of *CLIPCleaner*. We highlight the sections corresponding to the two main steps of *CLIPCleaner*, and particularly visualize the intuition of the probability estimation step based on the CLIP zero-shot classifier.

3.1 Revisiting sample selection for LNL

Given a dataset of training samples $\{x_i, y_i\}_{i=1}^N$ *i.i.d* sampled from a noisy joint distribution $P^n(\mathbf{x}, \mathrm{y})$, the goal is to learn a classifier $f : \mathbf{x} \to \mathrm{y}$ that can accurately predict the true labels y for new, unseen clean examples. Let us denote the clean (but unknown) joint distribution as $P(\mathbf{x}, \mathrm{y})$. Sample selection methods aim to identify those samples with (possibly) clean labels. Let us also denote the sample selection results as $\{w_i \in \{0,1\}\}_{i=1}^N$ with $w_i = 1$ or 0 representing sample $\boldsymbol{x}_i$ been selected or not. For specific sample $\boldsymbol{x}_i$, here, allowing us to propose a concise form to represent most existing sample selection methods:

$$w_i = \mathbb{G}(\tilde{P}(\mathrm{y}|\mathbf{x} = \boldsymbol{x}_i), y_i; \theta) \in \{0,1\}. \tag{1}$$

We define as $\tilde{P}(\mathrm{y}|\mathbf{x} = \boldsymbol{x}_i)$ an estimation of the clean conditional probability $P(\mathrm{y}|\mathbf{x} = \boldsymbol{x}_i)$, and abbreviate here as $\mathbb{G}$ a specific selection mechanism, with its hyperparameter as θ. Intuitively speaking, we conceptualize a sample selection method into two steps: firstly estimating clean conditional probability $\tilde{P}(\mathrm{y}|\mathbf{x} = \boldsymbol{x}_i)$ for sample $\boldsymbol{x}_i$. Then, applying $\mathbb{G}$ to compare $\tilde{P}(\mathrm{y}|\mathbf{x} = \boldsymbol{x}_i)$ with the annotated label y_i, to decide/measure if the annotated label is (likely) clean or not[1].

However, most sample selection methods inherently and inevitably lead to the 'self-confirmation' bias as they commonly (more or less) rely on the in-training model f in estimating the conditional probability: $\tilde{P}(\mathrm{y}|\mathbf{x} = \boldsymbol{x}_i) = P_f(\mathrm{y}|\mathbf{x} = \boldsymbol{x}_i)$. To fully avoid such 'self-confirmation' bias - the reliance of sample selection on in-training model f, utilizing another pre-trained classifier naturally fits. Specifically, in this work, we consider to utilize the CLIP model for sample selection.

3.2 CLIPCleaner: sample selection with vision-language models

3.2.1 Preliminary on CLIP. We first briefly introduce the CLIP model [42], which is currently one of the most prevalent vision-language models. CLIP aims to learn from a dataset of image-text pairs, denoted as $(\boldsymbol{x}'_i, z_i)_{i=1}^M$, which is *i.i.d.* sampled from a hidden joint distribution $Q(\mathbf{x}, \mathbf{z})$. Specifically, we consider $Q(\mathbf{x}, \mathbf{z})$ as the marginalization of $Q(\mathbf{x}, \mathrm{y}, \mathbf{z})$ for ease of later analysis. We denote as $\boldsymbol{x}'$ the images in CLIP training dataset to discriminate from above in-question noisy dataset, and $\boldsymbol{z}$ the corresponding text descriptions. Then, we have below as CLIP training loss:

$$\begin{aligned} L(\boldsymbol{x}'_i, z_i; g, h) = \frac{1}{2}\Big(&- \log \frac{\exp(g(\boldsymbol{x}'_i)^T h(z_i))}{\sum_{j=1}^M \exp(g(\boldsymbol{x}'_i)^T h(z_j))} \\ &- \log \frac{\exp(g(\boldsymbol{x}'_i)^T h(z_i))}{\sum_{j=1}^M \exp(g(\boldsymbol{x}'_j)^T h(z_i))}\Big). \end{aligned} \tag{2}$$

[1] Please note, there are indeed some methods such as TopoFilter [53] and FINE [28] relying on graph models or eigenvectors rather than probability estimations for sample selection. With eq. (1) we are not attempting to cover all possible sample selection mechanisms but to motivate our proposed method.

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.

3.2.2 Estimate $P(\mathrm{y}|\mathbf{x}=\boldsymbol{x}_i)$ with CLIP zero-shot classifier. Due to its multimodal nature, CLIP naturally possesses the ability for zero-shot classification. As a relatively new technology for the LNL community, here we revisit CLIP's zero-shot classification from a probabilistic perspective, which will also serve as our method for estimating true conditional probabilities with CLIP.

Let us recall $\boldsymbol{x}, y, z$ as the image, label and text respectively. Firstly, we assume $y \perp \boldsymbol{x} \mid z$; intuitively, the semantic label y_i can be independently generated based on a decent image description z_i alone for each image $\boldsymbol{x}_i$. For zero-shot classification, we have:

$$\begin{aligned} P_{zeroshot}(\mathrm{y}=y_i|\mathbf{x}=\boldsymbol{x}_i) &= \int Q(\mathrm{y}=y_i|\mathbf{z}=z_i)Q(\mathbf{z}=z_i|\mathbf{x}=\boldsymbol{x}_i)dz \\ &\propto \int Q(\mathrm{y}=y_i|\mathbf{z}=z_i)Q(\mathbf{z}=z_i,\mathbf{x}=\boldsymbol{x}_i)dz. \end{aligned} \tag{3}$$

To calculate above integral analytically is often hard; Practically, we tend to estimate $P_{zeroshot}(\mathrm{y}=y_i|\mathbf{x}=\boldsymbol{x}_i)$ by sampling z_i, if $Q(\mathrm{y}=y_i|\mathbf{z}=z_i)$ and $Q(\mathbf{z}=z_i,\mathbf{x}=\boldsymbol{x}_i)$ is known. Firstly, according to the training loss used by CLIP, we know that[2]:

$$Q(\mathbf{z}=z_i,\mathbf{x}=\boldsymbol{x}_i) \propto \exp(g(\boldsymbol{x}_i)^T h(z_i))).$$

Still, $Q(\mathrm{y}=y_i|\mathbf{z}=z_i)$ remains unknown. Original CLIP designs a single prompt as '`A photo of class name of` y_i`.`', implicitly assuming that:

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

Then, with a single prompt we can easily sample a single z_i to estimate $P_{zeroshot}(\mathrm{y}=y_i|\mathbf{x}=\boldsymbol{x}_i)$ according to eq. (3). Moreover, 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 apply below template to generate multiple prompts $\{\mathcal{P}_j\}_{j=1}^J$ using class-specific features such as the unique color or habitat of different animal species in an animal classification task[3]:

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

Then, we can similarly estimate $P_{zeroshot}(\mathrm{y}=y_i|\mathbf{x}=\boldsymbol{x}_i)$ with above prompts as below:

$$P_{zeroshot}(\mathrm{y}=y_i|\mathbf{x}=\boldsymbol{x}_i) \underset{\sim}{\propto} \textstyle\sum_{j=1}^J \tilde{Q}(\mathbf{z}=\mathcal{P}_j,\mathbf{x}=\boldsymbol{x}_i). \tag{4}$$

3.2.3 Calculate w_i with specific $\mathbb{G}$. With the above estimated conditional probability $\tilde{P}(\mathrm{y}|\mathbf{x}=\boldsymbol{x}_i) = P_{zeroshot}(\mathrm{y}=y_i|\mathbf{x}=\boldsymbol{x}_i)$, we can apply any applicable strategy for sample selection as depicted in eq. (1). As exploration on more advanced sample selection strategy $\mathbb{G}$ is not the focus in this paper, we consider two simple sample selection strategies below.

Firstly, we consider a *consistency-based selector* - compute sample's consistency metric (defined as the ratio of the probability of noisy label class to the highest class probability) and identify samples with high consistency as clean:

$$\mathbb{G}_{consistency} = \mathbb{I}\Big(\frac{\tilde{P}(\mathrm{y}=y_i|\mathbf{x}=\boldsymbol{x}_i)}{\max_k \tilde{P}(\mathrm{y}=k|\mathbf{x}=\boldsymbol{x}_i)} \geq \theta_{consistency}\Big). \tag{5}$$

Here, $\mathbb{I}$ is the indicator function, $\theta_{consistency}$ is the manually-defined threshold, often as 1 by default.

Denoting as $\{\tilde{P}(\mathrm{y}|\mathbf{x}=\boldsymbol{x}_i)\}_{i=1}^N$ the estimated probabilities for the training dataset, we also consider the widely-applied *loss-based sample selector* - computing the sample's cross-entropy loss ($\{-\log\tilde{P}(\mathrm{y}=y_i|\mathbf{x}=\boldsymbol{x}_i)\}_{i=1}^N$) and then dividing the dataset into two parts based on a Gaussian Mixture Model (GMM), with the part having smaller losses designated as clean samples:

$$\mathbb{G}_{loss} = \mathbb{I}(\mathbb{P}(-\log\tilde{P}(\mathrm{y}=y_i|\mathbf{x}=\boldsymbol{x}_i) \in \mathtt{GMM}_{small}) \geq \theta_{loss}). \tag{6}$$

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 separate GMM model[4]. Here, θ_{loss} is also the manually-defined threshold, often as 0.5 by default.

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

With selected subset only, *CLIPCleaner* can be utilized along with any existing methods - see SUPPLEMENTARY C for results of utilizing *CLIPCleaner* with DivideMix [29]. However, the state-of-the-art methods often involve multiple modules, such as iterative sample selection and model training [10, 22, 29], model co-training [19, 63], and multi-task contrastive learning [31, 38]. While these modules can be effective, they introduce extra complexity into the learning process, requiring careful coordination and tuning. To streamline the process and enhance efficiency, we aim to avoid intricate methodologies and propose a simple semi-supervised learning method for noisy datasets — namely *MixFix*.

Let us denote the selected subset and non-selected subset as $\mathcal{X}_c, \mathcal{Y}_c$ and $\mathcal{X}_n, \mathcal{Y}_n$. Motivated by FixMatch [43], we also inspect in unlabeled subset $(\mathcal{X}_n, \mathcal{Y}_n)$ each sample's current prediction $\boldsymbol{p}_i$ based on the in-training model f:

$$(w_i, y_i) = \begin{cases} (0, y_i), \text{ if } p_m < \theta_r \text{ and } p_m < \theta_r' & \texttt{*Drop*} \\ (1, y_i), \text{ if } p_m > \theta_r \text{ and } y_i = y_m & \texttt{*Absorb*} \\ (1, y_m), \text{ if } p_m > \theta_r' \text{ and } y_i \neq y_m & \texttt{*Relabel*} \end{cases} \tag{7}$$

Here we denote as $p_m \triangleq \max_l \boldsymbol{p}_i(l)$ and $y_m \triangleq \arg\max_l \boldsymbol{p}_i(l)$. Please note the difference of $\boldsymbol{p}_i$ here with our previous estimated probabilities for sample selection. Intuitively, we '`absorb`' more clean samples $(y_i = y_m)$ (not been selected in the sample selection step) and '`relabel`' noisy samples $(y_i \neq y_m)$ with different thresholds (θ_r and θ_r') in non-selected subset, and progressively append it to initial selected subset to form a dynamic larger training set $\mathcal{X}_t, \mathcal{Y}_t$. Different from FixMatch [43] using one threshold for all samples, we typically set $\theta_r \leq \theta_r'$. This allows us to fully leverage noisy labels to distinguish between the '`absorb`' and '`relabel`' processes. Then, we apply a common cross-entropy loss for training with this expanded training set $\mathcal{X}_t, \mathcal{Y}_t$.

[2] Please refer to SUPPLEMENTARY E for full derivation.

[3] Please refer to SUPPLEMENTARY B for more details about prompts generation.

[4] Please refer to SUPPLEMENTARY D for specific ablations.

3.4 Theoretical justification of *CLIPCleaner*

Ignoring the language modality and treating the CLIP model as an ordinary pre-trained model, we can also leverage its vision encoder g solely along with the interested noisy dataset $(\boldsymbol{x}_i, y_i)_{i=1}^{N}$ to induce a new classifier f' (to discriminate it with the model f in section 3.1) for estimating the clean conditional probability in eq. (1). For example, we can simply freeze the weights of g, use it as a fixed feature encoder, and train a linear classifier f' upon it based on the interested noisy dataset. Then the predicted logits after softmax normalization can be used as an estimate of $P(\mathrm{y}|\mathbf{x} = \boldsymbol{x}_i)$:

$$P_{induced}(\mathrm{y}|\mathbf{x} = \boldsymbol{x}_i) = \mathrm{softmax}(f'(g(\boldsymbol{x}_i))). \tag{8}$$

An immediate question is: how does the zero-shot classifier (eq. (4)*) compare to the induced classifier here (eq.* (8)*) in estimating the clean conditional probability?* In fact, the induced classifier can be based on any visual pre-trained model. If such easily-induced classifier demonstrates performance comparable to or even better than the zero-shot classifier, then we have no motivation to specifically adopt the CLIP model for sample selection. To this end, we conduct a theoretical analysis and compare the estimated $\tilde{P}(\mathrm{y}|\mathbf{x} = \boldsymbol{x}_i)$ in both options with the true/unknown $P(\mathrm{y}|\mathbf{x} = \boldsymbol{x}_i)$. 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 $(\boldsymbol{x}, \boldsymbol{z})$ in $\sup(Q)$ with $g, h \in \mathcal{G}, \mathcal{H}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ we have the following hold:*

$$d(P_{zeroshot}, P) \leq \varepsilon_{domain} + \Delta\left(\lambda_0\varepsilon_{clip} + \lambda_1\mathfrak{R}(\mathcal{G} \circ \mathcal{H}) + \lambda_2 l_\infty^{clip}\sqrt{\frac{\log 1/\delta}{M}} + \lambda_3\varepsilon_n\right)$$

with $\lambda_0, \lambda_1, \lambda_2, \lambda_3 > 0$. Here, ε_{domain} denotes the bias term induced by the domain gap between Q and P^{true}, ε_{clip} denotes the expected risk of the Bayes optimal CLIP model, and $\Delta \geq 1$ denotes the bias coefficient induced by designing prompts and sampling in eq. (3).

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

$$d(P_{induced}, P) \leq \varepsilon_{noise} + \lambda_0\varepsilon_{induced} + \lambda_1\mathfrak{R}(\mathcal{F}) + \lambda_2 l_\infty^{noisy}\sqrt{\frac{\log 1/\delta}{N}}$$

with $\lambda_0, \lambda_1, \lambda_2 > 0$. Here, ε_{noise} denotes the difference term induced by the label noise in the training dataset, and $\varepsilon_{induced}$ denotes the expected risk of the Bayes optimal induced classifier.

Please refer to Supplementary F for full derivation. With theorem 3.1 and theorem 3.2, ignoring the uncontrollable and common bound error terms (marked in gray), we find that *the zero-shot classifier is affected by domain gap and prompts quality while the induced classifier is affected by the label noise of the noisy dataset*, which is intuitively consistent with our expectation[5]. Put simply, ε_{noise} is always unavoidable in theorem 3.2, even with a perfectly-learned feature encoder; By contrast, in theorem 3.1, Δ can be reduced through better prompt engineering, and ε_{domain} can be minimized by training CLIP with a more diverse dataset, thus reducing the domain gap. *We emphasize that this is the unique advantage of CLIP for sample selection as a vision language model.*

3.5 Additional discussion

To be greedy or conservative? So far, we have mentioned two different conditional probability estimation options (eq. (4) and eq. (8)) and two different sample selection strategies (eq. (5) and eq. (6)), resulting in a total of four different combinations for a possible overall method. The theoretical analysis above and the subsequent empirical ablations show that these different combinations exhibit their preferences in different scenarios. In this work, we adopt a conservative strategy by taking the intersection of different sample selection results, prioritizing the precision of sample selection. Compared to more greedy sample selection strategies, we tend to rely on the introduced semi-supervised learning strategy - *MixFix* - to gradually incorporate more samples into training. This can avoid amplifying the impact of noisy samples due to overly greedy sample selection, but it also has the obvious weakness that it will inevitably miss some 'hard' clean samples. We leave further exploration 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, 69]. 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

In this section, we conduct extensive experiments on two standard benchmarks with synthetic label noise, CIFAR10 and CIFAR100, and four real-world noisy datasets, Red Mini-ImageNet [23], Clothing1M [57], WebVision [33], and ANIMAL-10N [44]. We mainly follow previous works [10, 16, 29] for model and training configurations, please refer to Supplementary G for full details. For comparison to other works, we report the results from most advanced SOTA methods - normally including techniques like co-training, contrastive learning, etc.

4.1 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 leads to performance degradation, and $\theta_r < \theta_r'$ leads to better performance

[5] Please note we are not aiming for strict/tight bounds but to validate the intuition: zero-shot classifier is noise-free while the induced classifier is noise-affected.

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

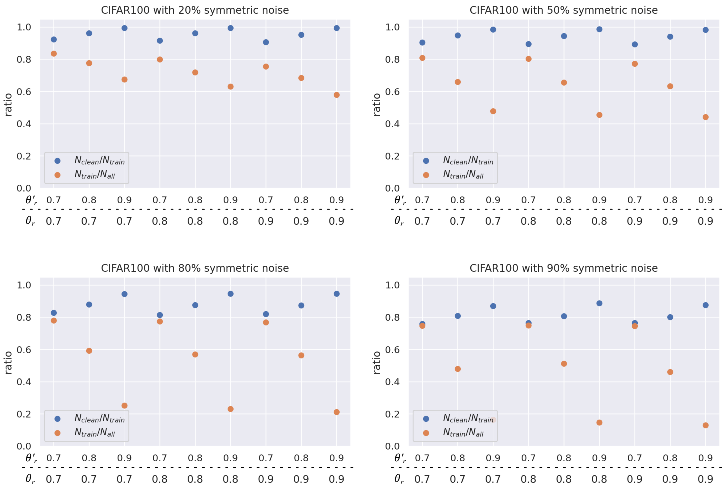

Figure 3: N_{train} denotes number of training samples, N_{clean} denotes number of clean training samples and N_{all} denotes number of clean training samples.

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

Model	CIFAR10	CIFAR100	Red Mini-ImageNet	WebVision	Clothing1M	ANIMAL-10N
CLIP	89.97	63.72	78.12	73.36	39.73	76.12
SOTA	92.68 [22]	67.7 [22]	49.55 [16]	80.9 SSR+ [10]	74.84 C2D [68]	88.5 SSR+ [10]
Ours	**95.15**	**71.17**	**54.21**	**81.56**	**74.87**	**88.85**

than setting the same value for both thresholds. In fig. 3, 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 involved 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 significant 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 sample selection 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-induced 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. 4, we show the sample selection result and find that:

- Firstly, as the noise ratio increases, regardless of the dataset (CIFAR10 *vs.* CIFAR100 *vs.* Red Mini-ImageNet), noise modes (symmetric *vs.* asymmetric *vs.* real-world) or CLIP backbones (VIT-B/32 *vs.* VIT-L/14@336px *vs.* RN50), the zero-shot classifier gradually outperforms the induced classifier. This further validates the unique advantage of CLIP and our theoretical findings in section 3.4 - that the latter is affected by label noise while the former is not;
- Additionally, 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;
- Furthermore, we notice that when comparing two different choices for obtaining the induced classifier, the *LogisticRegression* classifier empirically exhibits superior performance to the *kNN* classifier. Therefore, we choose the *LogisticRegression* classifier as our default choice for the induced classifier.

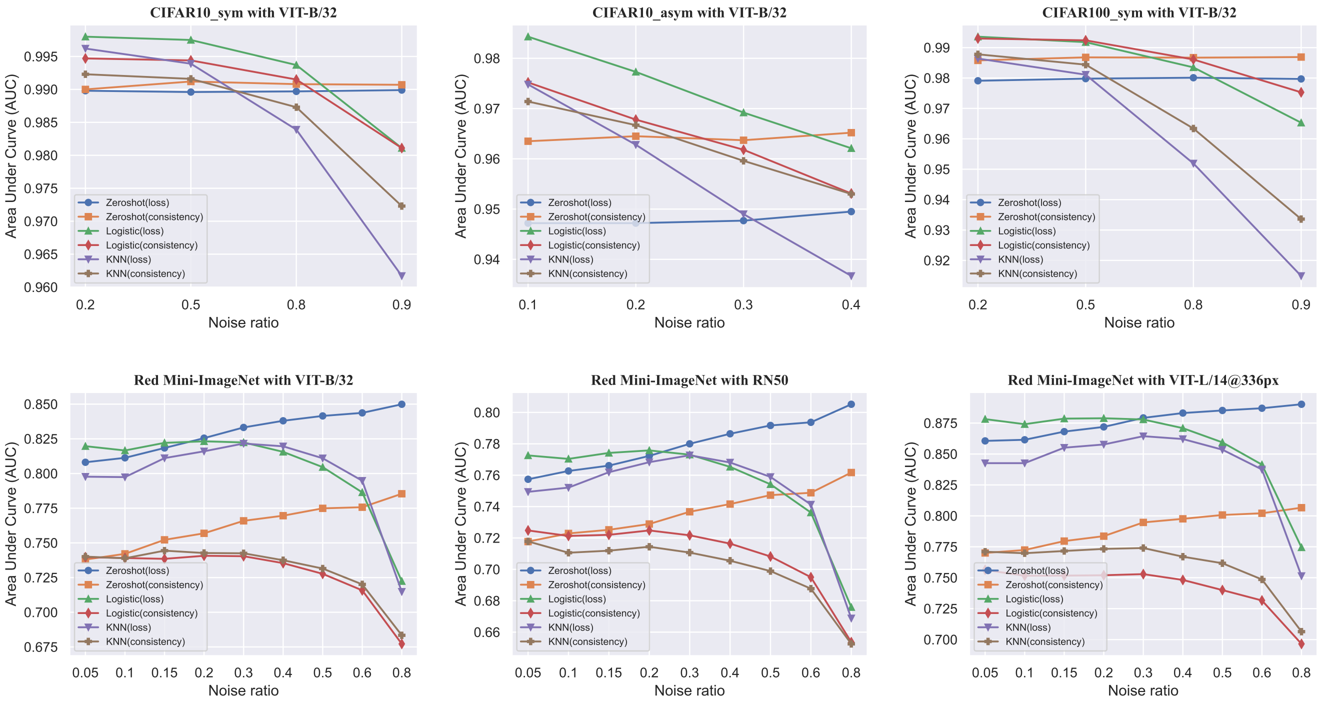

Figure 4: Comparisons of various sample selection methods *w.r.t* different dataset/noise type/noise ratio. Here, we show the ROC AUC score of binary identification of clean samples.

4.2 Results on synthetic noisy dataset

In this section, we first evaluate our method on the CIFAR datasets with synthetic symmetric/asymmetric noise. In table 4, 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 3. 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 additional proof of the superior sample selection performance of *CLIPCleaner*.

4.3 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 3: 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 [41]	91.06	86.35	78.87	71.12
Co-teaching [19]	91.22	87.28	84.33	78.72
GCE [67]	90.97	86.44	81.54	76.71
DAC [47]	90.94	86.16	80.88	74.80
DMI [58]	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**

We note, that the proposed *CLIPCleaner* can also be used in combination with other schemes. In table 5 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 [29]. 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

Table 4: Testing accuracy (%) on CIFAR-10 and CIFAR-100 with synthetic noise.

Dataset	CIFAR10					CIFAR100			
Noise type	Symmetric				Assymetric	Symmetric			
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+ [63]	89.5	85.7	67.4	47.9	-	65.6	51.8	27.9	13.7
F-correction [41]	86.8	79.8	63.3	42.9	87.2	61.5	46.6	19.9	10.2
PENCIL [62]	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 [29]	**96.1**	94.6	93.2	76.0	93.4	77.3	74.6	60.2	31.5
ELR+ [34]	95.8	94.8	93.3	78.7	93.0	77.6	73.6	60.8	33.4
MOIT [38]	93.1	90.0	79.0	69.6	92.0	73.0	64.6	46.5	36.0
SelCL+ [31]	95.5	93.9	89.2	81.9	93.4	76.5	72.4	59.6	48.8
TCL [22]	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**

Table 5: Testing accuracy (%) on Clothing1M.

CE	F-correction [41]	RRL [30]	C2D [68]	DivideMix [29]	ELR+ [34]	SSR+ [10]	TCL [22]	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 6: Testing accuracy (%) on WebVision.

Methods	WebVision		ILSVRC2012	
	Top1	Top5	Top1	Top5
Co-teaching [19]	63.5	85.20	61.48	84.70
DivideMix [29]	77.32	91.64	75.20	90.84
ELR+ [34]	77.78	91.68	70.29	89.76
NGC [54]	79.16	91.84	74.44	91.04
FaMUS [59]	79.4	92.8	77.0	92.8
RRL [30]	76.3	91.5	73.3	91.2
SelCL+ [31]	79.9	92.6	76.8	**93.0**
SSR+ [10]	80.9	92.8	75.8	91.8
TCL [22]	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

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

Method	Noise ratio			
	20%	40%	60%	80%
CE	47.36	42.70	37.30	29.76
Mixup [64]	49.10	46.40	40.58	33.58
DivideMix [29]	50.96	46.72	43.14	34.50
MentorMix [23]	51.02	47.14	43.80	33.46
FaMUS [59]	51.42	48.06	45.10	35.50
InstanceGM [16]	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**

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 Supplementary H.

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

Method	Accuracy
CE	79.4
SELFIE [44]	81.8
PLC [66]	83.4
NCT [6]	84.1
InstanceGM [16]	84.6
SSR+ [10]	88.5
Ours	**88.85±0.61**

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

[6] 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).

[7] 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).

[8] Chen Feng and Ioannis Patras. 2022. Adaptive Soft Contrastive Learning. In *2022 26th International Conference on Pattern Recognition (ICPR)*. 2721–2727. https://doi.org/10.1109/ICPR56361.2022.9956660

[9] Chen Feng and Ioannis Patras. 2023. MaskCon: Masked Contrastive Learning for Coarse-Labelled Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 19913–19922.

[10] 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.

[11] Chen Feng, Georgios Tzimiropoulos, and Ioannis Patras. 2024. NoiseBox: Towards More Efficient and Effective Learning with Noisy Labels. *IEEE Transactions on Circuits and Systems for Video Technology* (2024), 1–1. https://doi.org/10.1109/TCSVT.2024.3426994

[12] 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.

[13] 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).

[14] Zheng Gao, Chen Feng, and Ioannis Patras. 2024. Self-Supervised Representation Learning With Cross-Context Learning Between Global and Hypercolumn Features. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 1773–1783.

[15] Zheng Gao and Ioannis Patras. 2024. Self-Supervised Facial Representation Learning with Facial Region Awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2081–2092.

[16] 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.

[17] 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.

[18] Jacob Goldberger and Ehud Ben-Reuven. 2016. Training deep neural-networks using a noise adaptation layer. (2016).

[19] 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).

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9729–9738.

[21] 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).

[22] Zhizhong Huang, Junping Zhang, and Hongming Shan. 2023. Twin Contrastive Learning with Noisy Labels. In *CVPR*.

[23] 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.

[24] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. 2018. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*. PMLR, 2304–2313.

[25] 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.

[26] 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.

[27] 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).

[28] Taehyeon Kim, Jongwoo Ko, JinHwan Choi, Se-Young Yun, et al. 2021. Fine samples for learning with noisy labels. *Advances in Neural Information Processing Systems* 34 (2021), 24137–24149.

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

[30] Junnan Li, Caiming Xiong, and Steven Hoi. 2020. Learning from Noisy Data with Robust Representation Learning. (2020).

[31] 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.

[32] 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.

[33] 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).

[34] 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).

[35] 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.

[36] Eran Malach and Shai Shalev-Shwartz. 2017. Decoupling" when to update" from" how to update". *arXiv preprint arXiv:1706.02613* (2017).

[37] Ioannis Maniadis Metaxas, Georgios Tzimiropoulos, and Ioannis Patras. 2023. DivClust: Controlling Diversity in Deep Clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3418–3428.

[38] Diego Ortego, Eric Arazo, Paul Albert, Noel E 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.

[39] 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).

[40] 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.

[41] 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.

[42] 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.

[43] 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).

[44] 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.

[45] Zhonglin Sun, Chen Feng, Ioannis Patras, and Georgios Tzimiropoulos. 2024. LAFS: Landmark-based Facial Self-supervised Learning for Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1639–1649.

[46] 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.

[47] 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).

[48] 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.

[49] 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).

[50] 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.

[51] 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).

[52] 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.

[53] 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.

[54] 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).

[55] Xiaobo Xia, Bo Han, Nannan Wang, Jiankang 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.

[56] 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).

[57] 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.

[58] 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).

[59] 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.

[60] 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.

[61] 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).

[62] 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.

[63] 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.

[64] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412* (2017).

[65] 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.

[66] 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).

[67] 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).

[68] 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).

[69] 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.

[70] 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*.

[71] 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.

[72] 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.