# JoAPR: Cleaning the Lens of Prompt Learning for Vision-Language Models

Yuncheng Guo[1], Xiaodong Gu[2]*
Department of Electronic Engineering, Fudan University, Shanghai 200438, China
[1]23210720033@m.fudan.edu.cn, [2]xdgu@fudan.edu.cn

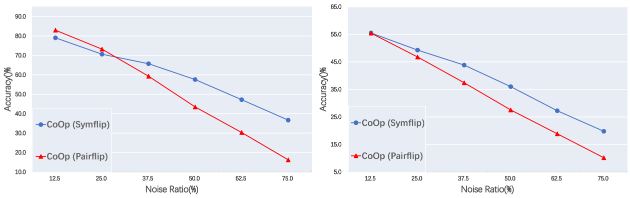
## Abstract

*Leveraging few-shot datasets in prompt learning for Vision-Language Models eliminates the need for manual prompt engineering while highlighting the necessity of accurate annotations for the labels. However, high-level or complex label noise challenges prompt learning for Vision-Language Models. Aiming at this issue, we propose a new framework for improving its robustness. Specifically, we introduce the* ***Jo****int* ***A****daptive* ***P****artitioning for Label* ***R****efurbishment (****JoAPR****), a structured framework encompassing two key steps. 1) Data Partitioning, where we differentiate between clean and noisy data using joint adaptive thresholds. 2) Label Refurbishment, where we correct the labels based on the partition outcomes before retraining the network. Our comprehensive experiments confirm that JoAPR substantially enhances the robustness of prompt learning for Vision-Language Models against label noise, offering a promising direction for future research.*

## 1. Introduction

Vision-Language Pre-Trained Models(VL-PTMs) [18, 33, 48] have garnered significant attention in the field of computer vision due to their image understanding and processing capabilities. CLIP [33], as a representative model in VL-PTMs, excels at comprehending the intricate relationship between images and text, going beyond traditional classification tasks. It leverages a vast dataset of over 400 million image-text pairs for pre-training. The pre-trained text encoder and image encoder are instrumental in obtaining embeddings for text and images, enabling the fusion of textual and visual information in downstream tasks. The abundance of data equips CLIP with remarkable transfer learning capability.

Specifically, CLIP formulates a description text for each class, referred to as a prompt (e.g. "A photo of a $[CLASS]$"), which is then fed into the pre-trained text encoder to generate corresponding embeddings. Simultaneously, the image to be recognized is processed by the pre-trained image encoder to extract image features. The similarity between these text and image embeddings is calculated to achieve the task of image recognition. Recently, Zhou *et al*. [55] shed light on one of the primary challenges in developing large visual models: the design of appropriate prompts. They discover that even minor changes to the prompt significantly impact the model's performance. Manual prompt design, however, is time-consuming and finding the best prompt is impractical. To address this, they introduce prompt learning into VL-PTMs, known as CoOp. This model requires only a few-shot dataset to learn suitable prompts and has substantially improved the transferability of large vision-language models to downstream tasks.



(a) Caltech101

(b) DTD

Figure 1. Accuracy vs. Noise Ratio on two datasets. The graph depicts the impact of synthetic label noise (Symflip and Pairflip) on the CoOp's accuracy. CoOp maintains a certain degree of performance at lower noise ratios, indicating some resilience to mild noise conditions. However, with an increase in the noise ratio, there is a clear downward trend in model accuracy for both types of noise, with Pairflip leading to a more pronounced degradation in performance. This finding highlights the limited robustness of prompt learning for VL-PTMs.

Despite the enhanced transfer learning capability of prompt learning for VL-PTMs, a critical issue in its development has been overlooked by many researchers: how to handle label noise in the few-shot dataset. Despite easier acquisition of massive datasets today, accurate annotation remains costly, often leading to label noise that impairs model performance. Prompt learning for VL-PTMs is no exception, motivating us to explore ways to make it more robust. Recently, Wu *et al*. [41] observe the presence

*corresponding author

of label noise in prompt learning. However, their research primarily focuses on why prompt learning for VL-PTMs is more robust compared to traditional transfer learning methods, falling short in investigating the effects of higher noise levels and more complex noise on prompt learning for VL-PTMs. While prompt learning for VL-PTMs is indeed more robust than traditional transfer learning methods, this robustness is far from sufficient. As depicted in Fig. 1, the model's performance significantly degrades under conditions of high label noise and complex label noise. In this paper, drawing inspiration from sample-selection methods in Learning with Label Noise (LNL), we propose JoAPR, which leverages the network's *memorization effects* to handle noisy data. Considering CLIP's inherent prior knowledge and powerful zero-shot learning ability, we adopt a strategy utilizing label refurbishment to combat label noise. Due to the small number of samples in few-shot dataset, it is a difficult problem to accurately divide the clean data and noise data. JoARP employs a two-component Gaussian Mixture Model(GMM) to model the loss values of the data and utilizes joint adaptive thresholds to distinguish clean data from noisy data. Subsequently, we refurbish the labels of the data and retrain the model on the revised dataset.

Our contributions can be summarized as

(1) We unveil the inadequacy of prompt learning for VL-PTMs in coping with higher noise ratios or more complex noise.

(2) We introduce JoAPR, the first systematic solution to tackle label noise in prompt learning for VL-PTMs, markedly boosting their robustness. Our model enables prompt learning for VL-PTMs to sustain exceptional performance, even under harsh conditions with extremely high noise levels or complex noise patterns.

(3) We design joint adaptive thresholds for clean and noisy data, effectively removing the need for manual hyperparameter tuning and significantly enhancing classification accuracy. Simultaneously, we address performance degradation due to misclassified noisy data by implementing a strategy to control the probability of clean label refurbishment. This dual approach enables our model to excel in few-shot learning scenarios.

(4) We conduct extensive experiments on ten datasets with varying noise ratios and different types of label noise to validate and illustrate the robust performance of our model. Additionally, numerous supplementary experiments are conducted to further show JoAPR.

## 2. Related work

**Prompt Learning for VL-PTMs** Regarding the textual description input for the text encoder, known as the prompt, CLIP [33] initially employs the manually defined format "A photo of a $[CLASS]$". However, Zhou *et al.* [55] point out that even slight changes in wording could have a profound impact on performance. Manually discovering the optimal prompt proves to be a non-trivial task. To address this challenge, they pioneer the introduction of prompt learning to the realm of VL-PTMs and devise the CoOp model. CoOp utilizes a few-shot labeled dataset to train the prompt learning module, enabling the automatic identification of the most effective prompt. CoOp maintains the frozen text encoder and image encoder of CLIP, significantly enhancing CLIP's adaptability to downstream tasks with remarkable efficiency. Building upon the foundations laid by CoOp, CoCoOp [54] extends its capabilities by addressing the limitation of CoOp's context generalization. CoCoOp introduces a lightweight network to incorporate information from visual cues into the prompt, enhancing context learning and allowing for broader generalization across unseen categories within the same dataset. KAPT [19] introduces the incorporation of external knowledge into prompt learning. It leverages accurate descriptions of concepts and their contextual relationships to further enhance the generalization of prompt learning.

**Learning with Noise Labels** Deep Neural Networks (DNNs) trained with extensive data have showcased their formidable capabilities across various fields. The quality of data has emerged as a cornerstone for DNNs performance. While large-scale datasets are readily available, the manual labeling cost associated with them can be prohibitively high. Using mislabeled data can lead DNNs to overfit to noisy labels, significantly compromising model performance. Consequently, addressing label noise and ensuring robust model performance in its presence have become pivotal areas of research. To tackle this challenge, researchers have proposed a series of approaches, including the following. (1) Incorporating robust network structures specially designed to adapt to noise [6, 13, 21, 45, 46]. (2) Employing robust regularization tools to combat label noise [12, 17, 26, 40, 43, 51]. (3) Designing robust loss functions capable of tolerating noise present in the dataset [10, 11, 24, 25, 39]. (4) Correcting the loss through the utilization of an estimation matrix [4, 16, 31, 42, 47]. (5) Leveraging sample selection or meta learning techniques to distinguish noisy data within the dataset and rectify erroneous labels [23, 30, 34, 36, 37, 49, 53].

**Label Noise in Prompt Learning for VL-PTMs** Recently, Wu *et al.* [41] have investigated the impact of label noise on prompt learning for VL-PTMs. Their primary focus, however, is on understanding why prompt learning for VL-PTMs is more robust to noisy labels compared to traditional transfer learning approaches, like model fine-tuning and linear probes. While they emphasize prompt learning's increased robustness, it is still unclear whether this level of robustness is sufficient for effectively handling label noise.

Additionally, Wu *et al*. have not explored the effects of higher noise ratios or more intricate forms of label noise on prompt learning for VL-PTMs. As illustrated in Fig. 1, it becomes evident that the performance of prompt learning for VL-PTMs is significantly compromised in scenarios characterized by high levels of noise or complex noise patterns that are more challenging to manage.

## 3. Preliminary

**Noise generation** Before introducing our method, let's establish some fundamental definitions. We define the clean dataset as $D : \{(X_i, Y_i)\}_{i=1}^{N}$, where $X_i$ represents the input image, and $Y_i \in (1, 2, ..., L)$ represents the corresponding label. Here, $(0, 1)^L$ denotes the one-hot vector representation of $Y_i$. The noisy dataset, affected by label noise, is denoted as $\overline{D} : \{(X_i, \overline{Y}_i)\}_{i=1}^{N}$. Furthermore, we introduce the concept of an artificially generated noise transition matrix, defined as $T_{ij}(X) = P(\overline{Y} = j|Y = i, X)$. This matrix represents the probability that a label $Y = i$ is incorrectly labeled as $\overline{Y} = j$. We artificially generate two types of label noise. The first type is Symflip noise, where noisy labels are randomly drawn from other categories in the dataset. The second type is Pairflip noise, where noisy labels are exclusively selected from labels adjacent to the current label. This latter type of noise is more challenging to address and can significantly impact model performance, serving as a more stringent test of model robustness.

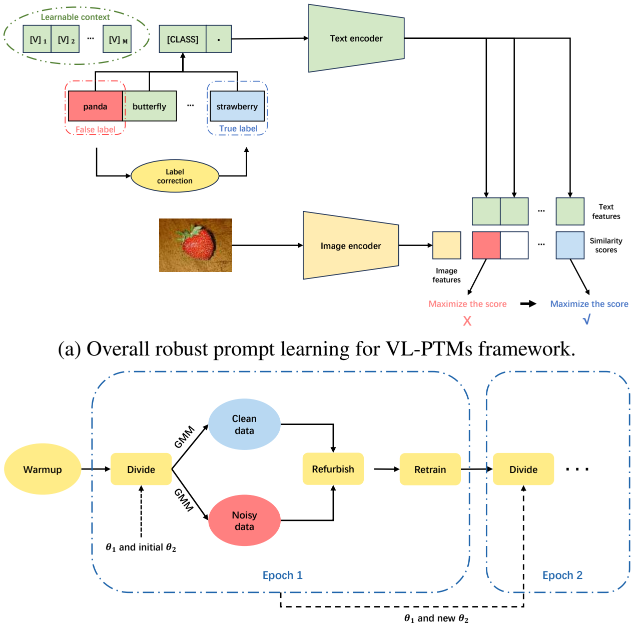
**Context Optimization (CoOp)** Instead of using the conventional "a photo of a $[CLASS]$" prompt, CoOp introduces $M$ learnable context vectors denoted as $t = [V]_1[V]_2...[V]_M[CLASS]$, where each $[V]_M(m \in \{1, ..., M\})$ is a vector with the same dimension as word embeddings, and $[CLASS]$ represents the word embedding(s) for the class name. With $g(\cdot)$ as the text encoder, the prediction probability is calculated as follows.

$$p(Y = i|\boldsymbol{X}) = \frac{\exp(\cos(g(\boldsymbol{t}_i), \boldsymbol{X})/\tau)}{\sum_{j=1}^{L}\exp(\cos(g(\boldsymbol{t}_j), \boldsymbol{X})/\tau)}. \quad (1)$$

The training process utilizes cross-entropy loss as the optimization objective and keeps the base model of CLIP frozen throughout.

## 4. Methodology

In this section, we delve into the specifics of our methodology, with the JoAPR architecture thoroughly delineated in Figs. 2a and 2b and **algorithm provided in the Supplementary Material**. The *memorization effect* observed in DNNs is crucial for sample selection, which describes DNNs initial learning of simpler patterns, preceding their advancement to memorize and overfit noisy data. Consequently, clean data generally shows a lower loss value than noisy data, a finding corroborated by [1, 2, 35]. Drawing on the insights from [1, 23], our strategy utilizes the small-loss rule to distinguish clean data from noisy data. We specifically adopt the Expectation-Maximization (EM) algorithm to model the network output loss distribution via a two-component GMM. Throughout each training epoch, we employ joint adaptive thresholds to differentiate clean data from noisy data. Following this separation, all labels undergo a process of label refurbishment, after which the data is retrained to improve the model's accuracy.



(a) Overall robust prompt learning for VL-PTMs framework.

(b) The pipeline of JoAPR.

Figure 2. (a) shows how our approach makes prompt learning more robust. (b) illustrates the iterative process of JoAPR, starting with a warmup phase followed by epochs that separate clean data from noisy data for label refurbishment and model retraining.

### 4.1. Warmup

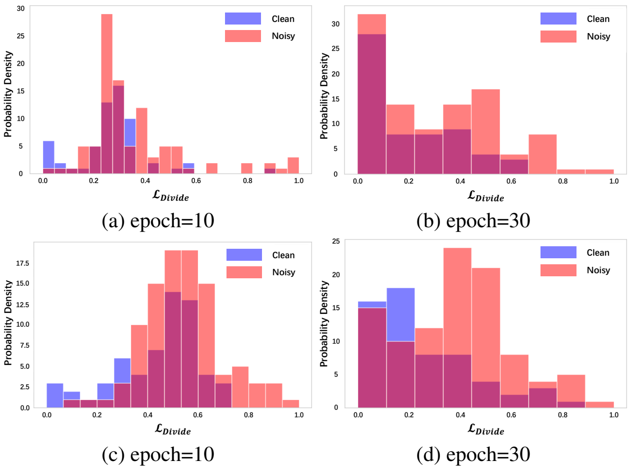
Warmup techniques are extensively employed in DNNs to expedite model convergence and mitigate oscillation. However, within the context of sample selection, Warmup may prematurely adapt models to noisy samples, complicating the divide and label refurbishment process. Addressing this, [23] implements an additional penalty loss to encourage more uniform predictions, thus reducing the risk of overfitting to noisy data. In the Warmup phase, the objective function comprises the sum of cross-entropy loss and a confidence penalty $-\mathcal{H}$ as defined in [32].

$$\begin{cases} \mathcal{L}_{Warmup} = \mathcal{L}_{CE} + \alpha_1(-\mathcal{H}) \\ \mathcal{H} = -\sum_{i=1}^{N} P_{model}(X_i) log(P_{model}(X_i)) \end{cases} \quad (2)$$

Here, $\alpha_1$ is a hyperparameter controlling the penalty's intensity, and $P_{model}(\cdot)$ represents the model's output. Nonetheless, our findings indicate that employing a confidence penalty can introduce additional challenges. When using cross-entropy for segregating clean and noisy data in datasets with limited samples and high noise levels, the penalty term may lead to overly uniform predictions. This can result in a discrete or overly overlapping $\mathcal{L}_{Divide}$ distribution between clean and noisy samples, as illustrated in Figs. 3a and 3b, thereby increasing data division errors. To mitigate this and encourage a more distinct aggregation of loss values for clean and noisy data, we incorporate a compensation term alongside cross-entropy in evaluating each sample's loss value for the data partitioning as

$$\mathcal{L}_{Divide} = \mathcal{L}_{CE} + \mathcal{H} \tag{3}$$

The penalty and compensation terms differ only in sign. As evidenced in Figs. 3c and 3d, introducing the compensation term clusters the loss value distribution more tightly, and reducing overlap between clean and noisy data in high-noise, few-shot datasets with limited samples. In practice, though, in few-shot datasets with a larger sample size, the tendency of the compensation term to centralize $\mathcal{L}_{Divide}$ values may increase this overlap, thus affecting data partitioning. **For a detailed description of this phenomenon, please refer to the Supplementary Material**. We also propose the JoAPR* framework, which omits compensation terms. Notably, due to the smaller size of the few-shot dataset compared to the regular dataset, attempts to reduce overlap are only partially successful. As depicted in Figs. 3c and 3d, significant overlap persists even with the compensation term. Rest assured, this issue will be further addressed in Sec. 4.3.



(a) epoch=10

(b) epoch=30

(c) epoch=10

(d) epoch=30

Figure 3. 62.5% of Pairflip is added to the EuroSAT. Blue represents clean data and red represents noisy data. (a) and (b) show the probability density distribution of cross-entropy loss values at epoch=10 and epoch=30 respectively. (c) and (d) show the probability density distribution of loss values after adding the compensation term at epoch=10 and epoch=30 respectively.

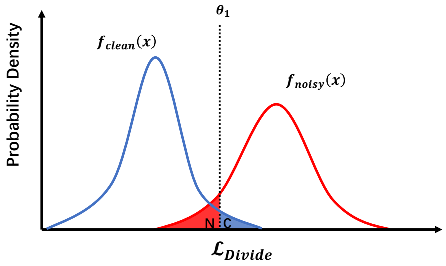


Figure 4. A schematic representation of Gaussian functions fitted to clean and noisy data separately.

## 4.2. Divide

This subsection discusses utilizing a GMM with two components to fit the $\mathcal{L}_{Divide}$. Traditional methods, which involve setting the partitioning threshold as a hyperparameter and manually adjusting it for data division, can be time-consuming and challenging in finding the optimal threshold. In contrast, our approach employs two adaptive thresholds, aiming to streamline the process and enhance partitioning accuracy. For enhanced convergence stability, we calculate the average $\mathcal{L}_{Divide}$ over the last five epochs.

$$\mathcal{L}_{Divide} = \frac{\sum_{i=t-4}^{t} \mathcal{L}_{Divide}^{i}}{5} \tag{4}$$

### 4.2.1 Threshold $\theta_1$

We fit the loss with GMM each epoch. The Gaussian functions for clean data $f_{clean}(x) = \frac{1}{\sigma_c\sqrt{2\pi}}e^{-\frac{(x-\mu_c)^2}{2\sigma_c^2}}$ and noisy data $f_{noisy}(x) = \frac{1}{\sigma_n\sqrt{2\pi}}e^{-\frac{(x-\mu_n)^2}{2\sigma_n^2}}$ are defined respectively. Two Gaussian functions are illustrated in Fig. 4. To identify the optimal threshold, we seek $\theta_1$ that minimizes the shaded area's overlap as

$$\min \quad F(\theta_1) = \int_{\theta_1}^{+\infty} f_{clean}(x)\,dx + \int_{-\infty}^{\theta_1} f_{noisy}(x)\,dx \tag{5}$$
$$\text{s.t.} \quad \theta_1 > 0$$

The optimal $\theta_1$ is derived using differentiation, as shown in Eq. (6).

### 4.2.2 Threshold $\theta_2$

Employing the EM algorithm, each sample is fitted to a two-component GMM to obtain the posterior probability $p(g_{clean}|\mathcal{L})$. Here, $g_{clean}$ denotes the Gaussian component

$$\theta_1 = \frac{-2\mu_n\sigma_c^2 + 2\mu_c\sigma_n^2 + \sqrt{(2\mu_n\sigma_c^2 - 2\mu_c\sigma_n^2)^2 - 4(\sigma_n^2 - \sigma_c^2)[(\mu_c\sigma_n)^2 - (\mu_n\sigma_c)^2 + 2(\sigma_c\sigma_n)^2 \ln\frac{\sigma_c}{\sigma_n}]}}{2\sigma_n^2 - 2\sigma_c^2} \tag{6}$$

with a smaller mean (representing clean samples), and $\mathcal{L}$ is short for $\mathcal{L}_{Divide}$. A sample is deemed clean if its posterior probability exceeds a certain threshold. We dynamically adjust this threshold based on the previous epoch's partitioning results, rather than manually tuning it as a hyperparameter. Specifically, it is defined as

$$\theta_2^t = \frac{N - card(\{(X_i, \overline{Y}_i) \mid p(g_{clean}|\mathcal{L}) > \theta_2^{t-1}\})}{N} \tag{7}$$

where $N$ is the total sample count, and $card(\cdot)$ represents the size of set. This approach is akin to a positive feedback process: if fewer clean samples are identified in the last epoch, the current epoch's noise ratio is considered higher, and $\theta_2$ is increased to make clean sample identification stricter, and vice versa. The initial value of $\theta_2$ is set to 0.5.

### 4.2.3 Joint Adaptive Partitioning

Sec. 5.7 confirms that using both $\theta_1$ and $\theta_2$ improves partitioning accuracy. Using these thresholds, we divide the original dataset $\overline{D}$ into clean $D_c$ and noisy $D_n$ subsets.

$$D_c : \{(X_i^c, Y_i^c) \mid \mathcal{L} < \theta_1 \ or \ \ p(g_{clean}|\mathcal{L}) > \theta_2\}_{i=1}^{N_c}$$
$$D_n : \{(X_i^n) \mid \mathcal{L} > \theta_1 \ and \ \ p(g_{clean}|\mathcal{L}) < \theta_2\}_{i=1}^{N_n}$$

## 4.3. Refurbish

Following the joint adaptive partitioning, we obtain the clean dataset $D_c$ and the noisy dataset $D_n$. Note that we discard the original labels of the noisy data. Initially, we apply $K$ data augmentations to these datasets that align with CoOp.

$$\begin{cases} (X_{i,a}^c)_{a=1}^K = Augment(X_i^c) \\ (X_{i,a}^n)_{a=1}^K = Augment(X_i^n) \end{cases} \tag{8}$$

Subsequently, we refurbish the labels of these datasets as

$$\begin{cases} \widehat{Y}_i^c = p_i Y_i^c + (1 - p_i) P_{model,i}^c \\ \widehat{Y}_i^n = P_{model,i}^n \\ P_{model,i}^c = \dfrac{\sum_{a=1}^k P_{model}(X_{i,a}^c)}{K} \\ P_{model,i}^n = \dfrac{\sum_{a=1}^k P_{model}(X_{i,a}^n)}{K} \end{cases} \tag{9}$$

where $P_{model}(\cdot)$ is the model's predicted output, and $p_i$ is the probability that the current data belongs to the clean label(or alternatively referred to as the label's confidence). We refurbish clean labels, recognizing that, despite the more accurate division achieved through joint adaptive partitioning, misclassification of noisy labels as clean is inevitable, which can be observed in Fig. 3. Therefore, the clean label is refurbished with the probability $p_i$, which is determined by

$$p_i = p(g_{clean}|\mathcal{L}) \cdot p_{softmax}(Y_i^c) \tag{10}$$

In this equation, $p(g_{clean}|\mathcal{L})$ is the posterior probability of belonging to the clean dataset, as derived from the EM algorithm, while $p_{softmax}(Y_i^c)$ is the softmax layer's predicted probability for the model output at the label $Y_i^c$. Early in training, as the model is yet to fully fit the samples, $p_{softmax}(Y_i^c)$ is often significantly less than 1. Notably, for noisy labels mistakenly categorized as clean, this value is lower compared to true clean samples, as noise samples are more challenging to fit. Consequently, even with a high $p(g_{clean}|\mathcal{L})$, the overall $p_i$ value will be low, leading to a preference for the model's predicted label over the original label during label refurbishment. As training progresses, $p_{softmax}(Y_i^c)$ for genuine clean samples approaches 1, while remaining low for misclassified noisy samples. This disparity ensures that misclassified samples continue to rely more on the model's predicted label, thereby mitigating robustness reduction caused by the misclassification of clean and noisy data. Note that for the initial epoch, we set $p_i = p(g_{clean}|\mathcal{L})$ to stabilize the model. The refurbished dataset is represented as follows.

$$\widehat{D}_c : \{(X_{i,a}^c, \widehat{Y}_i^c);\ i \in (1, ..., N_c),\ a \in (1, ..., K)\}$$
$$\widehat{D}_n : \{(X_{i,a}^n, \widehat{Y}_i^n);\ i \in (1, ..., N_n),\ a \in (1, ..., K)\}$$
$$\widehat{D} = Concat(\widehat{D}_c\ , \widehat{D}_n)$$

## 4.4. Retrain

Following the label refurbishment, we acquire the dataset $\widehat{D}$. Before retraining, $\widehat{D}$ undergoes data augmentation using MixMatch [3], a technique that fuses pseudo-labels with Mixup [50] and is prevalent in semi-supervised learning. This method of integrating semi-supervised learning techniques into label noise management, pioneered by [23], has been adapted in our approach as well. Let $\widehat{Y}$ represent a label from the dataset $\widehat{D}$, initially sharpened to bring the generated pseudo-label closer to the true one-hot vector.

$$\begin{cases} \widetilde{Y} = Sharpen(\widehat{Y}, T) \\ Sharpen(\widehat{y}_i, T) := \dfrac{\widehat{y}_i^T}{\sum_{j=1}^L \widehat{y}_j^T} \end{cases} \tag{11}$$

Here, $\widehat{y}_i$ is the $i$-th element of the one-hot vector form of $\widehat{Y}$, $T$, the temperature constant set to 2, and $L$ is the total number of categories in $\widehat{D}$. The resulting dataset after sharpening is denoted as $\widetilde{D}$. Next, we select a pair of samples ($\widetilde{X}_1$, $\widetilde{X}_2$) where $\widetilde{X}_1$ is from $\widetilde{D}$ and $\widetilde{X}_2$ from $Shuffle(\widetilde{D})$, along with their corresponding label pair ($\widetilde{Y}_1$, $\widetilde{Y}_2$). Then, we apply the Mixup as

$$\lambda \sim Beta(\beta, \beta), \tag{12}$$

$$\lambda^{'} = \max(\lambda, 1 - \lambda), \tag{13}$$

$$\widetilde{X}^{'} = \lambda^{'} \widetilde{X}_1 + (1 - \lambda^{'}) \widetilde{X}_2, \tag{14}$$

$$\widetilde{Y}^{'} = \lambda^{'} \widetilde{Y}_1 + (1 - \lambda^{'}) \widetilde{Y}_2, \tag{15}$$

where $Beta(\cdot)$ denotes the Beta distribution and $\beta$ is set to 0.001. Eq. (13) ensures $(\widetilde{X}^{'}, \widetilde{Y}^{'})$ is closer to $(\widetilde{X}_1, \widetilde{Y}_1)$. Mixup results in the dataset $\widetilde{D}^{'}$. [1] highlights that under high noise levels, network predictions tend to guide most samples towards the same class to minimize loss. To counter this, a regularization term used in [36] is added to encourage the model to make more uniform predictions as

$$\begin{cases} \mathcal{L}_{Retrain} = \mathcal{L}_{CE} + \alpha_2 \mathcal{L}_R \\ \mathcal{L}_R = \sum_{L} \frac{1}{L} \log(\frac{1}{L} \frac{\sum_{\widetilde{X}^{'} \in \widetilde{D}^{'}} P_{model}(\widetilde{X}^{'})}{KN}) \end{cases} \tag{16}$$

where $\alpha_2$ serves as a hyperparameter managing the regularization strength, $K$ represents the number of data augmentations, fixed at 1, and $N$ is the number of samples in the original dataset $\overline{D}$. The objective is now to maximize $P(\widetilde{Y}^{'}|X)$ instead of $P(\overline{Y}|X)$ via $\mathcal{L}_{Retrain}$.

## 5. Experiments

We conduct extensive experiments to ascertain the efficacy of our proposed model under complex and high-noise conditions. The following sections and **Supplementary Material** detail these experiments.

### 5.1. Datasets and Baseline

We employ ten diverse datasets covering generic objects classification, texture classification, fine-grained classification, scene recognition, action recognition, and satellite imagery recognition. Among these, Food101N [22] is included as a real-world noisy dataset. Given that there is no previous approach to tackle label noise in prompt learning for VL-PTMs, our baseline is CoOp, the most influential work in the domain. Following CoOp's methodology, we sample a 16-shot training set from each dataset, employing the original test set for evaluation. We introduce Symflip and Pairflip noise at varying intensities—12.5% to 75.0%—into these datasets. The results reported are averages over three experimental runs, with the highest accuracies highlighted in bold for emphasis.

### 5.2. Training Details

In our experiments, we follow the same setup as CoOp to keep the comparison fair. We utilize the same SGD optimizer, initiating with a learning rate of 0.002 and employing cosine annealing. The maximum epoch is set to 200, except for ImageNet, where it is 50. Our model backbone aligns with CoOp, employing the CLIP model with a ResNet-50[14] as visual encoder and a 63M parameter text Transformer[38] as text encoder. We utilize 16 context tokens shared across all categories and place the class token at the end. The hyperparameter $\alpha_2$ is set to 0.5 for Caltech101 and 1.0 for other datasets. **Warmup epochs and $\alpha_1$ settings are available in the Supplementary Material**.

### 5.3. Comparison with CoOp

Tabs. 1 and 2 illustrates the performance of the original CoOp and CoOp using JoAPR/JoAPR* across ten datasets. With compensation term added, JoAPR surpasses JoAPR* on EuroSAT, but on other datasets, JoAPR and JoAPR* each show strengths in different scenarios. The CoOp enhanced with JoAPR/JoAPR* shows greater robustness across all datasets, except at a 12.5% noise ratio on ImageNet. CoOp demonstrates robustness at lower noise levels and improves as the sample size increases, evident in datasets like SUN397 and ImageNet. This phenomenon suggests that a larger dataset size makes CoOp less prone to overfitting noisy labels. However, as the noise ratio increases, CoOp's performance is significantly impacted. In scenarios with 75% Pairflip noise, CoOp achieves less than 20% accuracy on most datasets, highlighting the substantial effect of high or complex noise on its performance. Surprisingly, CoOp with JoAPR/JoAPR* performs remarkably well under extreme noise conditions. This is particularly evident in datasets like Caltech101 and OxfordPets, where performance is only minimally affected even at a challenging 75% Pairflip noise. Notably, on the extensive ImageNet dataset, CoOp with JoAPR/JoAPR* excels regardless of the noise level, showcasing its resilience in any noisy situation. This observation holds true for other datasets as well.

### 5.4. Comparison with Robust Loss

Numerous robust loss functions have been proposed. In investigating why prompt learning exhibits greater robustness than traditional fine-tuning, Wu *et al.* [41] show that

Table 1. Comparison with CoOp on Food101N.

| Method | Accuracy |
|---|---|
| CoOp | 69.50 |
| CoOp+JoAPR | 72.57 |
| CoOp+JoAPR* | **73.87** |

Table 2. Comparison with CoOp on nine datasets.

| Dataset | Noise Type | Symflip | | | | | | Pairflip | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Method\Noise Ratio | 12.5% | 25.0% | 37.5% | 50.0% | 62.5% | 75.0% | 12.5% | 25.0% | 37.5% | 50.0% | 62.5% | 75.0% |
| ImageNet [8] | CoOp | **62.47** | 61.23 | 60.17 | 58.53 | 55.03 | 50.47 | **62.17** | 59.13 | 53.97 | 45.47 | 34.33 | 20.33 |
| | **CoOp+JoAPR** | 60.87 | 61.07 | **60.70** | 60.30 | 58.77 | **55.60** | 60.00 | 59.97 | 59.23 | 57.90 | 56.43 | **53.67** |
| | **CoOp+JoAPR*** | 61.23 | **61.30** | **60.70** | **60.33** | **59.00** | 55.13 | 60.73 | **60.67** | **60.07** | **58.53** | **56.67** | 53.53 |
| SUN397 [44] | CoOp | 66.30 | 63.37 | 60.07 | 56.63 | 50.73 | 40.27 | 64.73 | 57.17 | 47.53 | 34.90 | 21.30 | 10.37 |
| | **CoOp+JoAPR** | 67.23 | **68.13** | **67.33** | **67.03** | **64.77** | **60.90** | **66.97** | 66.30 | 64.90 | 61.50 | 55.90 | 48.83 |
| | **CoOp+JoAPR*** | **67.47** | 67.47 | 67.07 | 66.70 | 63.87 | 58.03 | 66.80 | **66.77** | **65.23** | **63.10** | **57.87** | **51.10** |
| Caltech101 [9] | CoOp | 79.03 | 70.60 | 65.70 | 57.57 | 47.20 | 36.67 | 82.97 | 73.20 | 59.27 | 43.47 | 30.30 | 16.23 |
| | **CoOp+JoAPR** | 88.50 | 89.07 | 88.47 | **89.03** | **87.67** | 84.87 | 88.80 | 89.17 | 88.80 | 88.27 | **86.17** | **84.43** |
| | **CoOp+JoAPR*** | **89.20** | **89.30** | **89.60** | 88.83 | 87.10 | **85.20** | **89.47** | **89.47** | **89.57** | **89.13** | 86.07 | 84.27 |
| Flowers102 [27] | CoOp | 86.13 | 81.07 | 74.93 | 68.47 | 55.50 | 39.37 | 86.47 | 76.43 | 63.07 | 45.20 | 27.10 | 12.40 |
| | **CoOp+JoAPR** | **90.13** | 88.13 | 84.47 | 82.13 | 75.60 | 75.13 | **89.80** | 88.83 | **84.73** | 73.27 | 71.03 | **62.87** |
| | **CoOp+JoAPR*** | 88.50 | **88.33** | **85.93** | **82.70** | **77.33** | **75.50** | 89.67 | **89.17** | 84.63 | **76.47** | **73.80** | 58.87 |
| StanfordCars [20] | CoOp | 66.37 | 59.00 | 54.23 | 47.70 | 36.93 | 24.70 | 65.67 | 57.03 | 46.47 | 33.10 | 20.70 | 11.30 |
| | **CoOp+JoAPR** | **68.60** | **67.63** | 65.77 | **63.53** | **58.97** | 51.53 | 67.00 | 64.47 | 61.20 | 54.87 | 47.20 | 36.57 |
| | **CoOp+JoAPR*** | 68.33 | 67.57 | **66.23** | 63.00 | 58.57 | **51.80** | **67.67** | **65.53** | **63.43** | **58.50** | **52.33** | **43.83** |
| OxfordPets [29] | CoOp | 77.67 | 69.23 | 58.73 | 48.37 | 35.37 | 22.37 | 76.40 | 65.70 | 51.87 | 37.00 | 25.90 | 14.17 |
| | **CoOp+JoAPR** | 85.20 | 85.40 | **85.27** | 85.67 | **85.30** | **83.77** | 85.97 | 86.93 | 86.07 | 85.87 | 82.77 | 76.93 |
| | **CoOp+JoAPR*** | **85.93** | **86.13** | 85.17 | **86.27** | 84.53 | 83.10 | **87.13** | **87.50** | **87.37** | **86.53** | **85.33** | **81.17** |
| UCF101 [28] | CoOp | 68.73 | 64.43 | 58.37 | 51.83 | 43.67 | 30.30 | 68.83 | 61.27 | 49.37 | 38.80 | 24.63 | 13.73 |
| | **CoOp+JoAPR** | **73.90** | 73.17 | **72.77** | 70.00 | **67.10** | **65.40** | 72.93 | **72.43** | **70.43** | 66.27 | 61.80 | 52.77 |
| | **CoOp+JoAPR*** | 73.37 | **73.83** | 71.40 | **70.30** | 66.83 | 63.80 | **73.03** | 72.40 | 69.77 | **69.10** | **63.40** | **56.23** |
| EuroSAT [15] | CoOp | 77.77 | 71.27 | 62.13 | 54.90 | 45.53 | 26.73 | 78.77 | 67.37 | 55.73 | 42.83 | 28.33 | 18.70 |
| | **CoOp+JoAPR** | 78.33 | 79.37 | **78.33** | **72.23** | **66.20** | **49.37** | **80.00** | **78.57** | **73.03** | **63.03** | **58.47** | **39.47** |
| | **CoOp+JoAPR*** | **79.30** | **80.53** | 78.07 | 67.33 | 59.20 | 34.45 | 78.23 | 78.50 | 69.43 | 58.23 | 40.85 | 25.90 |
| DTD [7] | CoOp | 55.50 | 49.27 | 43.83 | 36.00 | 27.23 | 19.77 | 55.43 | 46.77 | 37.40 | 27.53 | 18.87 | 10.17 |
| | **CoOp+JoAPR** | **58.83** | **57.67** | 55.70 | **53.07** | **50.67** | 46.30 | **57.33** | 55.13 | 55.03 | 48.53 | **45.00** | **32.53** |
| | **CoOp+JoAPR*** | 56.63 | 56.63 | **56.77** | **53.07** | 49.40 | **46.83** | 55.60 | **57.03** | **55.30** | **53.27** | 41.17 | 31.70 |

Generalized Cross Entropy (GCE) [52] may further enhance this robustness. Our investigation evaluates how CoOp performs when integrated with GCE in the presence of label noise. As presented in Fig. 5, CoOp utilizing GCE instead of CE demonstrates improved robustness. Nonetheless, when compared to our JoAPR/JoAPR* approach, GCE's robustness falls short in high-noise scenarios or with complex noise type. Notably, on the EuroSAT and DTD datasets with 75% Pairflip noise, CoOp augmented with GCE underperforms even the baseline CoOp that uses CE.

## 5.5. Comparison with Sample Selection

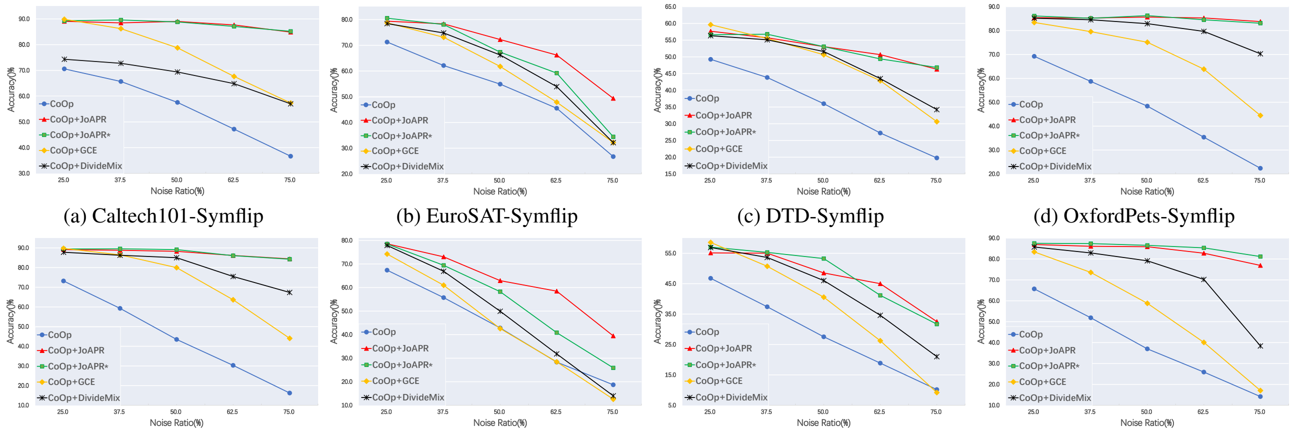
Considering that the idea of JoAPR originates from sample selection in LNL, DivideMix [23], which addresses label noise in a semi-supervised manner and excels in this domain, is chosen for comparison with JoAPR. The results are depicted in Fig. 5. Analyzing the outcomes, we observe that CoOp combined with DivideMix is more robust than CoOp with GCE in most cases. However, DivideMix performs considerably worse than JoAPR/JoAPR*. It should be noted that on the Caltech101 dataset injected with Symflip noise, DivideMix even performs worse than CoOp with GCE. This outcome underscores that traditional methods for handling label noise are not particularly effective in enhancing the robustness of prompt learning with few-shot datasets and further validates the efficacy of JoAPR/JoAPR* in mitigating the impact of label noise.

## 5.6. The Generalization of JoAPR

To assess the generalization of our proposed method, we extend our evaluation to CoOp's subsequent iteration, CoCoOp [54]. CoCoOp exhibits superior generalization to unseen classes within the same dataset compared to CoOp. Consequently, when comparing with CoCoOp, our focus remains on testing classes not present in the training set. The results are presented in Tab. 3. From the table, it is evident that the impact of noise on CoCoOp does not exhibit a negative correlation as observed in CoOp. This discrepancy arises because, even if CoCoOp overfits to noisy labels, its performance is unaffected when tested on new and unknown categories. However, it's crucial to note that noise labels themselves introduce incorrect knowledge learned by CoCoOp, leading to a notable impact, especially with the addition of Pairflip noise. Incorporating JoAPR/JoAPR*, CoCoOp excels across all scenarios, echoing the robust performance observed in CoOp.

Table 3. Comparison with CoCoOp on EuroSAT.

| Noise Type | Method\Noise Ratio | 12.5% | 25.0% | 37.5% | 50.0% | 62.5% | 75.0% |
|---|---|---|---|---|---|---|---|
| Symflip | CoCoOp | 61.30 | 55.83 | 46.77 | 60.40 | 43.20 | 32.77 |
| | CoCoOp+JoAPR | **62.20** | **61.90** | 58.43 | **63.93** | **59.37** | **63.90** |
| | CoCoOp+JoAPR* | 60.27 | 59.30 | **58.70** | 59.93 | 58.83 | 60.07 |
| Pairflip | CoCoOp | 56.90 | 54.43 | 54.03 | 32.47 | 34.67 | 24.97 |
| | CoCoOp+JoAPR | **62.43** | **61.77** | **61.47** | **61.53** | 60.30 | **59.67** |
| | CoCoOp+JoAPR* | 56.50 | 59.07 | 59.60 | 61.43 | **60.90** | 52.73 |



(a) Caltech101-Symflip (b) EuroSAT-Symflip (c) DTD-Symflip (d) OxfordPets-Symflip

(e) Caltech101-Pairflip (f) EuroSAT-Pairflip (g) DTD-Pairflip (h) OxfordPets-Pairflip

Figure 5. Performance of CoOp using GCE, DivideMix, CE and JoAPR/JoAPR* on four datasets respectively.

## 5.7. Ablation Study

Ablation studies on EuroSAT (Tab. 4) confirm the significance of each constituent in JoAPR's framework. Component omission led to performance deterioration, emphasizing their collective contribution to the model's overall efficacy. (1) Omitting Warmup hinders quick, stable model convergence. Warmup's confidence penalty term prevents overfitting to noisy labels, so its removal disrupts subsequent stages. (2) Using either $\theta_1$ or $\theta_2$ alone reduces the accuracy of the partitioning, leading to diminished model performance. (3) $p_{softmax}$ prevents early-stage overfitting to misclassified noise labels. (4) Label refurbishment, our method's ultimate goal, involves correcting mislabeled labels. As shown in our ablation study, model performance degrades more significantly when noisy labels are unrefurbished compared to clean labels, aligning with our expectations. (5) Mixmatch serves dual purposes: data augmentation and entropy reduction in refurbished pseudo-labels through Sharpen, making them more akin to one-hot labels.

## 5.8. Further Analysis

In order to see how powerful JoAPR really is, we conduct a highly challenging experiment introducing 100% noise to both the Caltech101 and OxfordPets, meaning that all training data is mislabeled. We conduct three rounds of experiments and select the best-performing one. The results, as depicted in Tab. 5, are remarkably surprising! With the assistance of JoAPR, CoOp demonstrates exceptional performance in both Symflip and Pairflip scenarios. In contrast, CoOp without JoAPR assistance becomes virtually non-functional. This suggests that for prompt learning, meticulously annotated datasets may no longer be a requirement. Consequently, this will significantly reduces the cost and time involved in dataset production or collection.

Table 4. Ablation studies on EuroSAT.

| Noise Type | Symflip | | | Pairflip | | |
|---|---|---|---|---|---|---|
| Method\Noise Ratio | 25.0% | 50.0% | 75.0% | 25.0% | 50.0% | 75.0% |
| JoAPR | **79.37** | **72.23** | **49.37** | **78.57** | **63.03** | **39.47** |
| JoAPR w/o Warmup | 74.00 | 48.77 | 33.20 | 73.60 | 61.83 | 27.53 |
| JoAPR w/o $\theta_1$ | 52.43 | 61.23 | 21.23 | 78.30 | 61.37 | 20.07 |
| JoAPR w/o $\theta_2$ | 75.07 | 63.13 | 43.47 | 75.87 | 60.50 | 35.57 |
| JoAPR w/o $p_{softmax}$ | 79.15 | 65.17 | 33.07 | 76.10 | 54.50 | 23.03 |
| JoAPR w/o clean refurbishment | 78.37 | 67.57 | 37.00 | 76.13 | 55.13 | 19.13 |
| JoAPR w/o noise refurbishment | 69.40 | 56.07 | 25.80 | 66.87 | 41.47 | 19.00 |
| JoAPR w/o MixMatch | 58.70 | 43.30 | 39.37 | 66.07 | 44.73 | 10.53 |

Table 5. Test results on datasets with all mislabeled labels.

| Dataset | Method\Noise Type | Symflip | Pairflip |
|---|---|---|---|
| Caltech101 | CoOp | 1.30 | 0.60 |
| | CoOp+JoAPR | 81.50 | **84.90** |
| | CoOp+JoAPR* | **84.80** | 84.10 |
| OxfordPets | CoOp | 4.70 | 1.40 |
| | CoOp+JoAPR | **72.60** | **76.90** |
| | CoOp+JoAPR* | 72.40 | 70.40 |

# 6. Conclusion

In this study, we uncover the vulnerability of prompt learning for VL-PTMs to label noise and propose a new method that uses joint adaptive thresholds to effectively separate clean from noisy data, aiding in label correction. This is, to our understanding, the first in-depth approach specifically designed to address label noise in VL-PTMs' prompt learning. Our broad experimental testing verifies that this approach considerably enhances the models' robustness to both simulated and authentic noise scenarios.

# Acknowledgement

This work was supported in part by National Natural Science Foundation of China under Grant 62176062.

## References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *International conference on machine learning*, pages 312–321. PMLR, 2019. 3, 6

[2] Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017. 3

[3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019. 5

[4] Haw-Shiuan Chang, Erik Learned-Miller, and Andrew McCallum. Active bias: Training more accurate neural networks by emphasizing high variance samples. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[5] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Plot: Prompt learning with optimal transport for vision-language models. *arXiv preprint arXiv:2210.01253*, 2022. 2

[6] Xinlei Chen and Abhinav Gupta. Webly supervised learning of convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1431–1439, 2015. 2

[7] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 7

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 7

[9] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 7

[10] Lei Feng, Senlin Shu, Zhuoyi Lin, Fengmao Lv, Li Li, and Bo An. Can cross entropy loss be robust to label noise? In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 2206–2212, 2021. 2

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

[12] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2

[13] Bo Han, Jiangchao Yao, Gang Niu, Mingyuan Zhou, Ivor Tsang, Ya Zhang, and Masashi Sugiyama. Masking: A new perspective of noisy supervision. *Advances in neural information processing systems*, 31, 2018. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[15] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 7

[16] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. *Advances in neural information processing systems*, 31, 2018. 2

[17] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International conference on machine learning*, pages 2712–2721. PMLR, 2019. 2

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1

[19] Baoshuo Kan, Teng Wang, Wenpeng Lu, Xiantong Zhen, Weili Guan, and Feng Zheng. Knowledge-aware prompt tuning for generalizable vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15670–15680, 2023. 2

[20] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 7

[21] Kimin Lee, Sukmin Yun, Kibok Lee, Honglak Lee, Bo Li, and Jinwoo Shin. Robust inference via generative classifiers for handling noisy labels. In *International conference on machine learning*, pages 3763–3772. PMLR, 2019. 2

[22] Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5447–5456, 2018. 6

[23] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020. 2, 3, 5, 7

[24] Yueming Lyu and Ivor W Tsang. Curriculum loss: Robust learning and generalization against label corruption. *arXiv preprint arXiv:1905.10045*, 2019. 2

[25] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *International conference on machine learning*, pages 6543–6553. PMLR, 2020. 2

[26] Aditya Krishna Menon, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Can gradient clipping mitigate label noise? In *International Conference on Learning Representations*, 2019. 2

[27] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008*

*Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 7

[28] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 7

[29] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 7

[30] Deep Patel and PS Sastry. Adaptive sample selection for robust learning under label noise. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3932–3942, 2023. 2

[31] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1944–1952, 2017. 2

[32] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017. 3

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2

[34] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. Selfie: Refurbishing unclean samples for robust deep learning. In *International Conference on Machine Learning*, pages 5907–5915. PMLR, 2019. 2

[35] Hwanjun Song, Minseok Kim, Dongmin Park, and Jae-Gil Lee. How does early stopping help generalization against label noise? *arXiv preprint arXiv:1911.08059*, 2019. 3

[36] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5552–5560, 2018. 2, 6

[37] Yuanpeng Tu, Boshen Zhang, Yuxi Li, Liang Liu, Jian Li, Yabiao Wang, Chengjie Wang, and Cai Rong Zhao. Learning from noisy labels with decoupled meta label purifier. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19934–19943, 2023. 2

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 6

[39] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 322–330, 2019. 2

[40] Hongxin Wei, Lue Tao, Renchunzi Xie, and Bo An. Open-set label noise can improve robustness against inherent label noise. *Advances in Neural Information Processing Systems*, 34:7978–7992, 2021. 2

[41] Cheng-En Wu, Yu Tian, Haichao Yu, Heng Wang, Pedro Morgado, Yu Hen Hu, and Linjie Yang. Why is prompt tuning for vision-language models robust to noisy labels? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15488–15497, 2023. 1, 2, 6

[42] Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? *Advances in neural information processing systems*, 32, 2019. 2

[43] Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *International conference on learning representations*, 2020. 2

[44] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 7

[45] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699, 2015. 2

[46] Jiangchao Yao, Jiajie Wang, Ivor W Tsang, Ya Zhang, Jun Sun, Chengqi Zhang, and Rui Zhang. Deep learning from noisy image labels with quality embedding. *IEEE Transactions on Image Processing*, 28(4):1909–1922, 2018. 2

[47] Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. Dual t: Reducing estimation error for transition matrix in label-noise learning. *Advances in neural information processing systems*, 33: 7260–7271, 2020. 2

[48] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 1

[49] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *International Conference on Machine Learning*, pages 7164–7173. PMLR, 2019. 2

[50] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 5

[51] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 2

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

[53] Guoqing Zheng, Ahmed Hassan Awadallah, and Susan Dumais. Meta label correction for noisy label learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11053–11061, 2021. 2

[54] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 2, 7

[55] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1, 2