Screening, Rectifying, and Re-Screening: A Unified Framework for Tuning Vision-Language Models with Noisy Labels

Chaowei Fang[1] , **Hangfei Ma**[1] , **Zhihao Li**[1] , **De Cheng**[1*] , **Yue Zhang**[2] , **Guanbin Li**[3,4]
[1]Xidian University
[2]Xi'an Jiaotong University
[3]Sun Yat-Sen Univeristy
[4]Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou, 510006, China
{cwfang,dcheng}@xidian.edu.cn, {hfm, zhihaoli}@stu.xidian.edu.cn, liguanbin@mail.sysu.edu.cn

Abstract

Pre-trained vision-language models have shown remarkable potential for downstream tasks. However, their fine-tuning under noisy labels remains an open problem due to challenges like self-confirmation bias and the limitations of conventional small-loss criteria. In this paper, we propose a unified framework to address these issues, consisting of three key steps: **Screening**, **Rectifying**, and **Re-Screening**. First, a dual-level semantic matching mechanism is introduced to categorize samples into clean, ambiguous, and noisy samples by leveraging both macro-level and micro-level textual prompts. Second, we design tailored pseudo-labeling strategies to rectify noisy and ambiguous labels, enabling their effective incorporation into the training process. Finally, a re-screening step, utilizing cross-validation with an auxiliary vision-language model, mitigates self-confirmation bias and enhances the robustness of the framework. Extensive experiments across ten datasets demonstrate that the proposed method significantly outperforms existing approaches for tuning vision-language pre-trained models with noisy labels.

1 Introduction

In the field of image classification, label quality often depends on various factors, such as the complexity of distinguishing target objects and the expertise of annotators. In annotation systems like crowdsourcing, noisy labels are inevitable. Considering deep neural networks (DNNs) are highly susceptible to memorizing mislabeled training samples, addressing learning with noisy labels is essential for ensuring the robustness of learned DNNs. Recently, pre-trained vision-language models [Radford *et al.*, 2021] have achieved remarkable progress and are increasingly adopted for downstream image classification tasks. This paper specifically tackles the challenge of fine-tuning pre-trained vision-language models under the presence of noisy labels.

Learning DNNs with noisy labels presents two main challenges. First, it requires effectively identifying mislabeled samples to prevent DNNs from memorizing incorrect data. Research [Arpit *et al.*, 2017] reveals that DNNs tend to first memorize clean samples with dominant patterns before noisy samples with less representative patterns. Hence, the small loss criterion is widely used for separating clean samples from noisy samples. Methods such as co-training with two models [Li *et al.*, 2020; Lu *et al.*, 2021; Kim *et al.*, 2024] leverage this criterion to mutually identify noisy samples. The second challenge is extracting value from noisy samples. Existing methods typically assign pseudo-labels to noisy samples [Li *et al.*, 2020; Yao *et al.*, 2021] or design self-supervised constraints [Liu *et al.*, 2020].

Recently, several studies [Wu *et al.*, 2023; Guo and Gu, 2024; Feng *et al.*, 2024] have explored the challenge of tuning the vision-language models under noisy labels. For example, [Wu *et al.*, 2023] analyzes the impact of noisy labels on CLIP [Radford *et al.*, 2021], revealing that performance significantly deteriorates as noise increases. To address this, existing methods often rely on self-generated predictions to clean samples or assign pseudo labels, as seen in [Guo and Gu, 2024] and [Feng *et al.*, 2024]. However, these approaches suffer from self-confirmation bias, where prediction errors propagate and amplify during training. Besides, conventional small-loss criterion fails to reliably distinguish clean samples from noisy ones, especially in ambiguous cases.

To address these limitations, we propose a novel algorithm that consists of three main steps: screening observed labels, rectifying potential noisy labels, and re-screening rectified labels. First, we introduce a dual-level semantic matching criterion that combines macro-level and micro-level textual prompts to better differentiate clean samples from noisy ones in the training loss space. This criterion is applied for identifying whether a training sample is a clean, ambiguous, or noisy sample. Second, we design separate label rectification strategies for refining labels of noisy and ambiguous samples with pseudo labels, aiming to effectively incorporate them into training. Lastly, we mitigate self-confirmation bias by re-screening rectified labels using cross-validation with another vision-language model, BLIP [Li *et al.*, 2022]. Extensive experiments conducted on ten datasets, demonstrate that the proposed method achieves state-of-the-art performances in tuning vision-language models with noisy labels.

Key contributions of this paper are summarized as below:

*Corresponding author (De Cheng)

1) We propose a novel unified framework, comprising three steps: **Screening** to identify noisy and ambiguous samples, **Rectifying** to assign corrected labels, and **Re-Screening** to validate the rectified labels for improving model robustness.
2) A novel dual-level semantic matching method is introduced, using macro- and micro-level textual prompts to effectively distinguish noisy samples from clean ones.
3) Extensive experiments on ten datasets demonstrate that our method establishes a new state-of-the-art for tuning vision-language models with noisy labels.

2 Related Work

Learning with Noisy Labels. Noisy labels degrade robustness of DNNs due to their susceptibility to overfitting. To mitigate this, researchers have developed methods for noise detection and label cleaning. Leveraging the observation that DNNs learn clean samples faster than noisy ones [Arpit *et al.*, 2017], several methods (e.g., [Park *et al.*, 2023; Patel and Sastry, 2023; Kim *et al.*, 2024]) utilize the small-loss criterion to identify clean samples. While effective to some extent, these methods can misclassify noisy samples resembling clean ones. Approaches like [Guo and Gu, 2024; Feng *et al.*, 2024] employ label correction techniques after sample partitioning, but they are prone to self-confirmation bias due to their reliance on model predictions, potentially hindering performance.

The other type of methods uses training objectives based on label transfer matrices [Lin *et al.*, 2024; Nguyen *et al.*, 2024; Bae *et al.*, 2024], reweighted losses [Yao *et al.*, 2024], and self-supervised learning [Liu *et al.*, 2020; Zhu *et al.*, 2024]. However, these methods often exhibit limited effectiveness due to their reliance on specific noise patterns.

Training strategies like curriculum learning (e.g., [Yu *et al.*, 2024]) and progressive label adjustment (e.g., [Zhang *et al.*, 2021], [Chen *et al.*, 2024]) mitigate noisy labels by gradually refining the learning process. While effective, these methods often entail increased computational overhead and necessitate careful hyperparameter tuning (e.g., epochs, curriculum structure). Improper optimization can lead to longer training times, higher resource consumption, and suboptimal performance.

Unlike existing methods, our method presents a three-step framework—screening, rectifying, and re-screening labels. We use a dual-level semantic matching criterion for improved label screening and distinct label rectification strategies for noisy and ambiguous samples. To mitigate self-confirmation bias, we employ a pre-trained vision-language model to cross-validate rectified labels, enhancing label correction accuracy and robustness compared to methods relying on a single model for both detection and rectification.

Prompt Tuning for Vision-Language Models. Vision-language pre-trained models have advanced significantly in recent years. CLIP [Radford *et al.*, 2021] employs prompt engineering to incorporate category-specific information into text inputs, allowing its pre-trained model to adapt to various tasks without further training. However, manual prompt design is time-consuming and requires expertise. To address this, CoOp [Zhou *et al.*, 2022b] introduces learnable prompts optimized for specific datasets, reducing manual effort and improving task adaptability. CoCoOp [Zhou *et al.*, 2022a] further enhances this by integrating image information into prompts using a lightweight network, enabling better context learning and generalization to unseen categories. BLIP [Li *et al.*, 2022] leverages weak supervision to generate high-quality visual information for task-specific purposes, improving performance. More recent works like ArGue [Tian *et al.*, 2024] align prompts with visual attributes from large language models and apply negative prompting for better out-of-distribution generalization, while AdvPT [Zhang *et al.*, 2024] uses learnable text prompts aligned with adversarial image embeddings to enhance robustness. While these methods advance prompt tuning, they primarily focus on static or model-based prompt designs. In contrast, our approach aims to adapt the pre-trained model to noisy label tasks. We combine macro- and micro-level textual prompts for label screening and rectification, and use BLIP to cross-validate rectified labels. This approach improves model robustness and adaptability, complementing traditional prompt tuning techniques.

3 Preliminary

Problem Definition. This paper aims to address the adaptation of the vision-language pre-trained model using training data with noisy labels. Formally, we denote the training dataset as $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x}_i$ denotes the i-th training sample, y_i denotes the observed label of $\mathbf{x}_i$, and N represents the total number of training images. Supposing the number of classes be C, we have $y_i \in \{1, \cdots, C\}$ which may be incorrect. The target is to adapt a pre-trained vision-language model which can tackle the image recognition task through matching images with class descriptions to the above dataset.

Prompt Tuning for Vision-Language Model. Following [Zhou *et al.*, 2022b], we use the vision-language pre-trained model CLIP consisting of an image encoder and a text encoder, as the image classification backbone model. We apply the prompt tuning algorithm to adapt the pre-trained text encoder to the target dataset. Given an input image $\mathbf{x}_i$, the image encoder transforms it into a feature vector $\mathbf{f}_i$. For purpose of tuning the text encoder, a set of learnable prompts are used to provide extra context of each class's textual prompt. We denote the learnable prompts of the j-th class as $\mathbf{t}_j$. Feeding the j-th class's text descriptions and the learnable prompts $\mathbf{t}_j$ into the text encoder, we can obtain the j-th class's embedding vector $\mathbf{o}_j$. Based on the image feature vector $\mathbf{f}_i$ and all classes' embedding vectors, we can infer the probability of $\mathbf{x}_i$ belonging to the j-th class as follows:

$$p(y = j \mid \mathbf{x}_i) = \frac{\exp\left(\cos\left(\mathbf{o}_j, \mathbf{f}_i\right)/\tau\right)}{\sum_{j'=1}^{C} \exp\left(\cos\left(\mathbf{o}_{j'}, \mathbf{f}_i\right)/\tau\right)}, \quad (1)$$

where τ is a hyperparameter set to 2. Through optimizing the learnable prompts only while freezing the image and text encoder, we can achieve the efficient adaptation of the pre-trained model into the target dataset. This can take advantage of the prior knowledge of the pre-trained model in extracting generalized vision and textual features. Due to few parameters required for optimization, only a small number of training data is required during the model adaptation process.

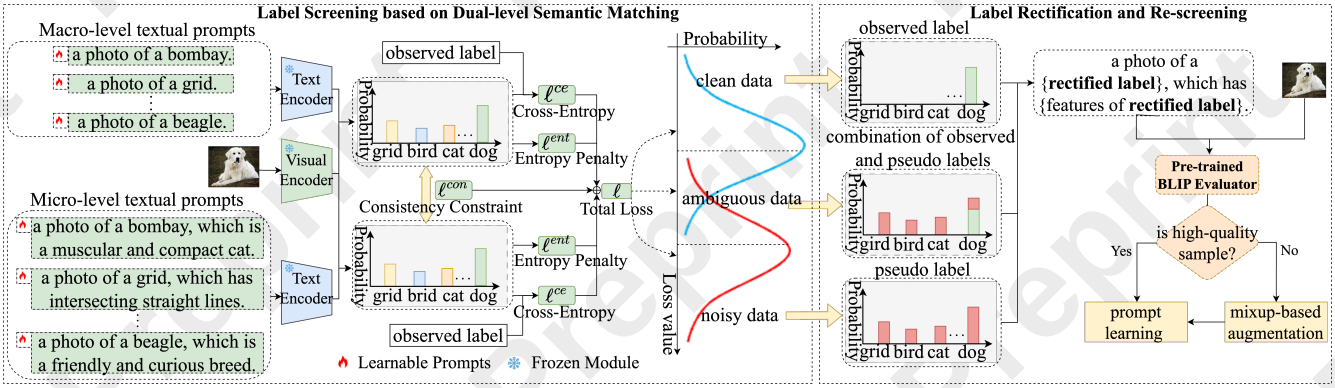


Figure 1: Overview of our framework, consisting of three steps: 1) label screening using dual-level semantic matching; 2) label rectification for noisy and ambiguous samples; and 3) label re-screening with BLIP. First, we screen labels based on the loss from predictions generated by macro- and micro-level textual prompts, partitioning samples into clean, ambiguous, and noisy subsets. Then, pseudo labels are applied to rectify labels in the ambiguous and noisy subsets. Finally, rectified labels are re-screened using BLIP.

4 Method

4.1 Overview of Proposed Framework

As shown in Fig. 1, we propose a framework leveraging pre-trained vision-language models to tackle noisy label learning. It consists of three steps: 1) screening observed labels to identify potential mislabels, 2) rectifying these labels for effective optimization, and 3) re-screening to mitigate self-confirmation bias.

In the screening step, we use a dual-level semantic matching mechanism with macro-level and micro-level prompts to classify samples as clean, ambiguous, or noisy ones. For rectification, noisy samples are assigned pseudo-labels, while ambiguous ones receive a combination of observed and pseudo labels. Finally, we use BLIP to re-screen and validate rectified labels. High-quality samples are directly used for prompt learning, while the rest are augmented via data mixing to minimize the influence of error labels.

4.2 Label Screening Based on Dual-level Semantic Matching

Dual-level Semantic Matching. During training, DNN-based models tend to memorize clean samples first, followed by noisy samples, which is why the small-loss criterion is commonly used to differentiate between the two. However, for CLIP prompt learning, using a single textual prompt struggles to effectively separate clean and noisy samples in the loss space. Macro-level prompts, based on class names, ensure inter-class separability but are prone to overfitting noisy samples due to the simplicity of the image-language matching. On the other hand, micro-level prompts, which describe fine-grained attributes, provide more reliable matching, reducing overfitting to noisy samples but may cause underfitting on clean samples. As a result, relying on just one type of prompt for image-class matching leads to significant overlap between clean and noisy samples. To address this, we propose a dual-level semantic matching mechanism, combining both macro-level and micro-level prompts as a label screening criterion. This approach facilitates the identification of clean, ambiguous, and noisy samples.

1) *Macro-level Textual Prompt.* We follow CLIP's original design, using a class name template to generate macro-level textual prompts. For class j, the macro-level textual prompt T_j^{mac} is defined as:

$$T_j^{\text{mac}} = \text{'a photo of a } \{\text{class } j\}.\text{'}$$

By feeding T_j^{mac} and learnable prompts $\mathbf{t}_j^{\text{mac}}$ into the text encoder, we obtain the embedding vector $\mathbf{o}_j^{\text{mac}}$ for class j. According to Eq. 1, these class embeddings $\{\mathbf{o}_j^{\text{mac}}\}_{j=1}^{C}$ are then used to compute the probabilities $\mathbf{p}_i^{\text{mac}} \in [0,1]^C$, representing the likelihood of sample $\mathbf{x}_i$ belonging to each class.

2) *Micro-level Textual Prompt.* Unlike macro-level prompts, micro-level prompts introduce class-specific features to enhance the alignment between image and text embeddings. Following [Feng *et al.*, 2024], we generate these prompts by incorporating detailed features such as shapes, textures, colors, and other unique attributes of each class. For class j, the micro-level textual prompt T_j^{mic} is expressed as:

$$T_j^{\text{mic}} = \text{'a photo of a } \{\text{class } j\}, \text{ which is/has } \{\text{features of class } j\}.\text{'}$$

The text encoder processes T_j^{mic} and $\mathbf{t}_j^{\text{mic}}$ to produce the micro-level embedding $\mathbf{o}_j^{\text{mic}}$. Using these embeddings $\{\mathbf{o}_j^{\text{mic}}\}_{j=1}^{C}$, we can predict the class probabilities $\mathbf{p}_i^{\text{mic}} \in [0,1]^C$, which represent the likelihood of $\mathbf{x}_i$ belonging to each class.

3) *Label Screening Criterion.* Based on the class probabilities estimated from the macro-level and micro-level textual prompts, we define a label screening criterion with a sample-wise loss function composed of three key terms: (1) a conventional cross-entropy loss, (2) a consistency constraint between the two types of prompt-induced predictions, and (3) an entropy penalty term. These terms are designed to enhance model robustness, especially in the presence of noisy labels.

The cross-entropy loss evaluates the discrepancy between the predicted class probabilities and the observed label. Specifically, for sample $\mathbf{x}_i$ with observed label $\mathbf{y}_i$ (one-hot

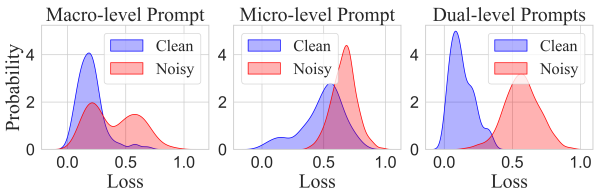


Figure 2: The loss distributions of clean and noisy samples using macro-level (left), micro-level (middle), and dual-level (right) textual prompts on DTD dataset under 75% Pairflip label noise.

encoded vector of y_i), the loss is computed as:

$$\ell^{\text{ce}}(\mathbf{x}_i, \mathbf{y}_i) = -\sum_{j=1}^{C} \left[y_{i,j} \log(p_{i,j}^{\text{mac}}) + y_{i,j} \log(p_{i,j}^{\text{mic}}) \right], \quad (2)$$

where $y_{i,j}$ represents the j-th element of $\mathbf{y}_i$, and $p_{i,j}^{\text{mac}}$ and $p_{i,j}^{\text{mic}}$ are the j-th elements of $\mathbf{p}_i^{\text{mac}}$ and $\mathbf{p}_i^{\text{mic}}$, respectively.

To address the issue of noisy labels, we impose a consistency constraint between the predictions generated by macro-level and micro-level prompts. This is achieved by calculating the Jensen-Shannon divergence between the two probability distributions $\mathbf{p}_i^{\text{mac}}$ and $\mathbf{p}_i^{\text{mic}}$. The consistency loss for sample $\mathbf{x}_i$ is given by:

$$\ell^{\text{con}}(\mathbf{x}_i) = \sum_{j=1}^{C} \left[p_{i,j}^{\text{mac}} \log \left(\frac{p_{i,j}^{\text{mac}}}{p_{i,j}^{\text{mic}}} \right) + p_{i,j}^{\text{mic}} \log \left(\frac{p_{i,j}^{\text{mic}}}{p_{i,j}^{\text{mac}}} \right) \right]. \quad (3)$$

This term helps mitigate the impact of noisy labels by promoting agreement between the predictions from both levels, effectively reducing the overfitting to noisy samples. By aligning the macro-level and micro-level predictions, the model is less likely to overfit to mislabeled data, resulting in improved generalization.

The entropy penalty term encourages more tight predictions by penalizing overly smooth outputs. It is designed to reduce the overlap between clean and noisy samples. The entropy penalty for sample $\mathbf{x}_i$ is calculated as:

$$\ell^{\text{ent}}(\mathbf{x}_i) = -\sum_{j=1}^{C} \left[p_{i,j}^{\text{mac}} \log(p_{i,j}^{\text{mac}}) + p_{i,j}^{\text{mic}} \log(p_{i,j}^{\text{mic}}) \right]. \quad (4)$$

The overall loss function is a weighted sum of the three loss terms:

$$\ell(\mathbf{x}_i, \mathbf{y}_i) = \ell^{\text{ce}}(\mathbf{x}_i, \mathbf{y}_i) + \lambda \ell^{\text{con}}(\mathbf{x}_i) + \beta \ell^{\text{ent}}(\mathbf{x}_i), \quad (5)$$

where λ and β are hyperparameters.

Fig. 2 illustrates the loss distributions of clean and noisy samples using macro-level, micro-level, and dual-level textual prompts on the DTD dataset with 75% Pairflip label noise. The loss distributions of clean and noisy samples overlap significantly when using either macro-level or micro-level prompts alone. In contrast, the dual-level textual prompts effectively reduce the overlap between clean and noisy samples, highlighting the advantage of combining macro-level and micro-level prompts for more accurate label screening.

Tri-Segment Sample Screening. Previous approaches [Guo and Gu, 2024; Arazo *et al.*, 2019] typically use small-loss mechanisms to separate clean and noisy samples. However, noise complexity often leads to misclassification of clean samples with losses resembling noisy ones, or vice versa. To address this, we propose a tri-segment screening strategy that categorizes samples into clean, ambiguous, noisy classes, accounting for overlap between clean and noisy data.

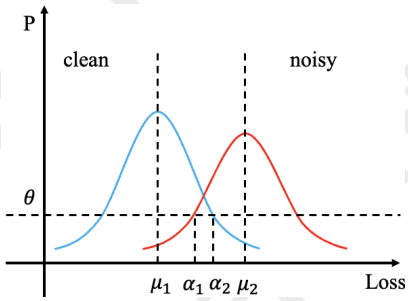


Figure 3: The semantic representation quantifying the overlap between the clean and noisy data distributions.

We model sample-wise losses with a two-component Gaussian Mixture Model (GMM), fitting the model every two epochs. The first and second Gaussian components represent the clean and noisy data loss distributions, respectively:

$$f_c(l) = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(l-\mu_1)^2}{2\sigma_1^2}}, \quad (6)$$

$$f_n(l) = \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{(l-\mu_2)^2}{2\sigma_2^2}}, \quad (7)$$

where $f_c(l)$ and $f_n(l)$ are the probability density functions of clean and noisy data, respectively. μ_k and σ_k denote the mean and variance of the k-th Gaussian component, respectively.

Given the overlap between these distributions as illustrated in Fig. 3, we define the ambiguous region for samples with loss values in this range. For a confidence level θ, we determine the ambiguous region boundaries α_1 and α_2 as follows:

$$\begin{cases} f_c(\alpha_1) = \theta, & \alpha_1 > \mu_1 \\ f_n(\alpha_2) = \theta, & \alpha_2 < \mu_2 \end{cases}. \quad (8)$$

Solving these equations gives thresholds α_1 and α_2:

$$\begin{cases} \alpha_1 = \mu_1 + \sqrt{-2\sigma_1^2 \ln(\theta\sigma_1\sqrt{2\pi})} \\ \alpha_2 = \mu_2 - \sqrt{-2\sigma_2^2 \ln(\theta\sigma_2\sqrt{2\pi})} \end{cases}. \quad (9)$$

The lower and upper boundaries of the ambiguous region are $\eta_l = \min(\alpha_1, \alpha_2)$ and $\eta_u = \max(\alpha_1, \alpha_2)$, respectively. A sample $\mathbf{x}_i$ is classified as follows:

- If $\ell(\mathbf{x}_i, \mathbf{y}_i) < \eta_l$, $\mathbf{x}_i$ is clean;
- If $\eta_l \leq \ell(\mathbf{x}_i, \mathbf{y}_i) < \eta_u$, $\mathbf{x}_i$ is ambiguous;
- If $\ell(\mathbf{x}_i, \mathbf{y}_i) \geq \eta_u$, $\mathbf{x}_i$ is noisy.

This label screening strategy partitions the dataset into clean, ambiguous, and noisy subsets, guiding further data utilization and label rectification for training.

4.3 Label Rectification and Re-screening

Label Rectification. To ensure sufficient training data, we introduce label rectification strategies that update each training sample's label using pseudo labels generated from model predictions. The pseudo label of sample $\mathbf{x}_i$, denoted as $\hat{\mathbf{y}}_i$, is the average of predicted probabilities from macro-level and micro-level prompts, i.e., $\hat{\mathbf{y}}_i = (\mathbf{p}_i^{\text{mac}} + \mathbf{p}_i^{\text{mic}})/2$.

The updated label $\tilde{\mathbf{y}}_i$ is determined by the following rules:

- If $\mathbf{x}_i$ is clean, $\tilde{\mathbf{y}}_i = \mathbf{y}_i$.
- If $\mathbf{x}_i$ is ambiguous, we combine the observed label $\mathbf{y}_i$ and pseudo label $\hat{\mathbf{y}}_i$ with confidence weight w_i:

$$\tilde{\mathbf{y}}_i = w_i \mathbf{y}_i + (1 - w_i)\hat{\mathbf{y}}_i. \tag{10}$$

  w_i is determined by the posterior probability that $\mathbf{x}_i$ belongs to the clean set.
- If $\mathbf{x}_i$ is noisy, the observed label is likely incorrect, so the pseudo label $\hat{\mathbf{y}}_i$ is used: $\tilde{\mathbf{y}}_i = \hat{\mathbf{y}}_i$.

Label Re-screening. Using pseudo labels for rectification can introduce self-confirmation bias due to the lack of reliable quality evaluation. To mitigate this, we use the pre-trained BLIP model for cross-validation. For each sample $\mathbf{x}_i$, we generate a text prompt $\widetilde{T}_i$ by incorporating the class with the highest value in $\tilde{\mathbf{y}}_i$ into the micro-level prompt template. We then use BLIP to compute the similarity between $\mathbf{x}_i$ and $\widetilde{T}_i$, denoted as s_i. The scores of all samples are fitted with a two-component GMM, and the posterior probability of s_i belonging to the component with the larger mean is calculated, resulting in a quality score q_i. This score is used to separate the training samples into high-quality and low-quality sets:

$$\begin{cases} \widetilde{\mathcal{D}}_h = \{(\mathbf{x}_i, \tilde{\mathbf{y}}_i) \mid q_i > \kappa\}, \\ \widetilde{\mathcal{D}}_l = \{(\mathbf{x}_i, \tilde{\mathbf{y}}_i) \mid q_i \leq \kappa\} \end{cases}$$

where κ is a threshold, typically set to 0.5.

4.4 Training Objective

Considering the unreliability of samples in $\widetilde{\mathcal{D}}_l$, we apply the mixup operation [Berthelot *et al.*, 2019] to augment them, yielding an updated dataset $\widetilde{\mathcal{D}}'_l$. This reduces the model's reliance on individual noisy labels and enhances data diversity, preventing overfitting. The final training objective is computed by accumulating losses over the union of $\widetilde{\mathcal{D}}_h$ and $\widetilde{\mathcal{D}}'_l$:

$$L = - \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in \widetilde{\mathcal{D}}_h \cup \widetilde{\mathcal{D}}'_l} \ell(\mathbf{x}_i, \mathbf{y}_i).$$

This objective is used to optimize the learnable prompts, mitigating the influence of noisy samples while making full use of all training samples.

5 Experiments

5.1 Datasets

We evaluate our method on ten datasets, including Flowers102 [Nilsback and Zisserman, 2008], EuroSAT [Helber *et al.*, 2019], StanfordCars [Krause *et al.*, 2013], OxfordPets [Parkhi *et al.*, 2012], DTD [Cimpoi *et al.*, 2014], Caltech101 [Fei-Fei *et al.*, 2004], UCF101 [Soomro, 2012], Food101 [Bossard *et al.*, 2014], ImageNet [Deng *et al.*, 2009] and SUN397 [Xiao *et al.*, 2010]. Following JoAPR [Guo and Gu, 2024], we adopt the same noise-labeled ways to generate our noisy dataset. One is Symflip noise, which is generated by randomly drawing labels from other categories in the dataset to replace the true labels. The other is more challenging Pairflip noise, which is generated by exclusively selecting labels that are adjacent to the true label to replace it. Following CoOp [Zhou *et al.*, 2022b], we sample a 16-shot training set from each dataset and employ the original test set for evaluation. To assess robustness, we test the method under noise rates ranging from 12.5% to 75%.

5.2 Implementation Details

Following the previous works [Zhou *et al.*, 2022b; Guo and Gu, 2024], we adopt the CLIP model with ResNet-50 [He *et al.*, 2016] as visual encoder and Transformer [Vaswani *et al.*, 2017] as text encoder. The number of tokens in learnable prompts is set to 16. We adopt SGD optimizer to train our model with an initialization learning rate of 0.002 and apply cosine annealing strategy. The maximum number of training epochs is set to 200, except for ImageNet where it is set to 50. To mitigate gradient explosion in the early stages of training, we adopt a warm-up strategy and fix the learning rate to 1×10^{-5} in first training epoch. The hyper-parameters λ, β and θ, are set to 0.5, 0.001, and 0.01, respectively.

5.3 Comparison with State-of-the-art Methods

Table 1 presents a comprehensive comparison between our method and state-of-the-art (SoTA) approaches including CoOp [Zhou *et al.*, 2022b], Robust UPL [Wu *et al.*, 2023], and JoAPR/JoAPR* [Guo and Gu, 2024], across ten datasets under various noise levels. Our method achieves the highest accuracy in most scenarios and the highest average accuracy across all noise levels. It can be seen that CoOp demonstrates robustness at lower noise levels, but its performance deteriorates significantly as the noise ratio increases. Robust UPL [Wu *et al.*, 2023] rely on pseudo-labels to alleviate the impact of noisy data on model performance. As the noise rate increases, the confidence in pseudo-labels decreases, which negatively impacts model performance. JoAPR improves upon CoOp by using data partitioning and label refurbishment but still struggles in high-noise scenarios. For instance, under 62.5% Symflip noise on the DTD dataset, our method outperforms JoAPR by 8.73%. On the StanfordCars dataset with 75% Pairflip noise, our method surpasses JoAPR* by 10.86%. Additionally, our method demonstrates strong robustness to high noise levels across all datasets.

5.4 Ablation Study

Effectiveness of Key Components. To better determine the contributions of key components in our method, we conducted a series of ablation experiments on the DTD and OxfordPets datasets, as shown in Table 2. The pre-trained vision-language model CoOp is used as the baseline (referred to as No.1). No.2, No.3, and No.4 represent the application of macro-level semantic prompt (MaLS), micro-level semantic prompt (MiLS), and dual-level semantic prompts (DLSM) to

Dataset	Method	Symflip						Pairflip					
		12.5%	25%	37.5%	50%	62.5%	75%	12.5%	25%	37.5%	50%	62.5%	75%
Flowers102	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
	Robust UPL	86.61	83.33	77.31	70.45	61.74	50.23	85.83	81.50	75.97	67.71	60.03	49.65
	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
	Ours	**90.22**	**90.01**	**86.11**	**84.98**	**82.14**	**75.52**	**89.85**	**89.20**	**85.02**	**77.95**	**74.38**	**67.40**
EuroSAT	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
	Robust UPL	77.46	71.87	64.05	57.13	46.11	32.36	76.78	70.71	63.51	52.14	41.70	27.11
	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
	Ours	**83.02**	**83.10**	**79.56**	**74.89**	**71.68**	**60.02**	**83.22**	**79.83**	**73.33**	**63.93**	**58.94**	**41.23**
StanfordCars	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
	Robust UPL	65.23	60.66	55.43	49.30	43.42	30.53	64.35	59.75	53.70	47.16	42.75	28.84
	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
	Ours	**70.13**	**68.60**	**66.57**	**64.18**	**60.30**	**57.05**	**68.92**	**67.04**	**64.21**	**58.56**	**55.55**	**54.69**
OxfordPets	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
	Robust UPL	83.79	82.37	79.25	79.69	67.90	46.29	83.13	81.77	79.18	77.95	65.17	44.43
	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
	Ours	**90.32**	**89.89**	**89.48**	**89.56**	**88.66**	**88.23**	**90.95**	**89.34**	**88.99**	**88.72**	**88.44**	**87.93**
DTD	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
	Robust UPL	58.69	55.61	49.00	43.14	35.76	28.13	58.63	54.96	48.47	42.26	32.47	26.18
	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
	Ours	**65.19**	**64.30**	**62.41**	**60.28**	**59.40**	**57.98**	**65.48**	**64.48**	**62.29**	**59.22**	**56.09**	**51.24**
Caltech101	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
	Robust UPL	83.08	81.77	81.29	76.52	71.08	65.00	81.64	81.22	80.77	75.17	65.19	60.12
	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
	Ours	**91.16**	**91.60**	**91.44**	**91.39**	**90.83**	**89.70**	**91.72**	**91.08**	**91.32**	**90.99**	**89.78**	**88.88**
UCF101	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
	Robust UPL	69.91	67.35	64.24	59.98	54.69	45.52	68.56	65.37	63.67	57.19	52.85	43.46
	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
	Ours	**76.24**	**76.13**	**75.63**	**73.73**	**71.05**	**69.57**	**77.16**	**73.88**	**70.84**	**69.57**	**65.53**	**63.42**
Food101	CoOp	72.83	69.43	66.57	63.33	57.37	46.67	72.00	65.30	56.00	43.30	26.90	12.87
	Robust UPL	74.09	71.33	70.04	64.99	61.03	46.43	73.26	70.71	69.06	62.96	60.52	44.90
	CoOp+JoAPR	75.27	75.30	75.03	74.90	75.33	75.00	75.13	75.33	75.03	75.30	75.23	75.20
	CoOp+JoAPR*	75.50	75.10	74.90	75.07	75.07	75.00	74.77	75.10	75.03	75.45	75.03	74.75
	Ours	**79.40**	**79.46**	**79.07**	**79.08**	**78.57**	**78.70**	**79.38**	**79.18**	**78.86**	**78.89**	**78.32**	**76.79**
ImageNet	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
	Robust UPL	61.97	60.91	60.10	57.48	55.58	51.16	61.37	58.76	54.25	49.03	41.27	37.58
	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
	Ours	61.27	**61.36**	**60.83**	**60.64**	**60.42**	**60.23**	61.43	**61.16**	**60.99**	**60.96**	**60.63**	**60.43**
SUN397	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
	Robust UPL	66.81	67.05	64.16	61.05	57.92	48.74	66.08	66.78	63.88	59.95	54.36	48.68
	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
	Ours	**67.78**	**68.40**	**67.62**	**67.53**	**64.92**	**63.13**	**67.39**	**66.92**	**65.68**	**63.92**	**63.13**	**62.39**
Average	CoOp	71.28	65.89	60.47	54.33	45.46	33.73	71.34	62.94	52.07	39.16	25.84	14.03
	Robust UPL	72.76	70.23	66.49	61.97	55.52	44.44	71.96	69.15	65.25	59.15	51.63	41.10
	CoOp+JoAPR	74.69	74.49	73.38	71.79	69.04	64.79	74.39	73.71	71.85	67.48	64.00	56.33
	CoOp+JoAPR*	74.55	74.62	73.58	71.36	68.09	62.88	74.31	74.21	71.98	68.83	63.25	56.14
	Ours	**77.47**	**77.29**	**75.87**	**74.63**	**72.80**	**70.01**	**77.55**	**76.21**	**74.15**	**71.27**	**69.08**	**65.44**

Table 1: Comparisons with existing SoTA methods across ten datasets. The highest accuracy achieved for each setting is highlighted in bold.

the baseline, respectively. Here, all training samples and their observed labels are directly used without processing. Comparing No.2, No.3, and No.4 highlights the effectiveness of using dual-level semantic matching for label screening. No.5 and No.6 demonstrate the importance of rectifying ambiguous and noisy labels selected out by our method, while No.7 and No.8 show that label re-screening further improves robustness. These contributions enhance our model's ability to handle various noise more effectively.

Comparison with Existing Two-Segment Label Screening and Rectification Methods. To evaluate the effectiveness of our proposed method, we compare it with two commonly used two-segment methods that categorize samples into clean and noisy categories: 1) keeping clean labels and refurbishing noisy labels as in Eq. (10) [Sohn *et al.*, 2020; Feng *et al.*, 2024], referred to as KCL-RNL; 2) refurbishing clean labels as in Eq. (10) and replacing noisy labels with pseudo labels [Li *et al.*, 2020; Guo and Gu, 2024], referred to as RCL-RNL. As shown in Fig. 4, our method achieves superior performance across noise ratios on two datasets. This validates the advantages of tri-segment partitioning for precise label rectification and enhanced robustness through differentiated sample handling.

Confidence Weight Analysis. We compared our adaptive setting scheme of the confidence weight (w_i) with fixed values (0.3, 0.5, 0.7). As shown in Table 3, the adaptive scheme consistently outperforms fixed weights by better handling uncertainty in ambiguous samples.

Robust Loss. Following [Wu *et al.*, 2023], we evaluate CoOp with generalized cross-entropy (GCE) [Zhang and Sabuncu, 2018] on two noisy datasets. As shown in Fig. 4, while GCE improves CoOp's performance over cross-entropy, our method achieves better results in most cases.

Hyper-parameters Analysis. We analyze the impact of con-

No.	MaLS	MiLS	DLSM	TSS-LR	LRS	DTD						OxfordPets					
						Symflip			Pairflip			Symflip			Pairflip		
						12.5%	50%	75%	12.5	50%	75%	12.5%	50%	75%	12.5	50%	75%
1						55.50	36.00	19.77	55.43	27.53	10.17	77.67	48.37	22.37	76.40	37.00	14.17
2	✓					60.17	47.77	33.94	59.01	40.20	27.84	87.54	80.10	60.83	86.14	75.42	60.89
3		✓				60.13	47.54	32.16	60.12	39.96	25.11	87.36	80.03	58.27	86.89	74.22	60.53
4			✓			60.52	48.46	35.82	60.76	41.31	30.73	87.68	80.57	64.98	87.03	77.35	61.98
5	✓			✓		62.41	55.38	51.48	61.52	51.48	46.34	89.64	88.42	87.24	89.78	87.60	86.59
6			✓	✓		62.77	57.57	51.89	64.54	52.07	49.29	89.75	89.07	88.06	90.08	88.09	86.92
7	✓			✓	✓	63.12	58.04	52.25	63.00	54.43	46.57	90.13	88.74	87.68	90.02	87.98	86.73
8			✓	✓	✓	**65.19**	**60.28**	**57.98**	**65.48**	**59.22**	**51.24**	**90.32**	**89.56**	**88.23**	**90.95**	**88.72**	**87.93**

Table 2: Ablation study for key components on DTD and OxfordPets datasets.

w_i	Symflip			Pairflip		
	12.5%	50%	75%	12.5%	50%	75%
0.3	64.83	59.1	49.71	65.31	53.78	35.64
0.5	65.15	57.74	50.24	65.04	52.72	37.88
0.7	64.95	57.68	50.41	65.08	52.48	38.83
Ours	**65.19**	**60.28**	**57.98**	**65.48**	**59.22**	**51.24**

Table 3: Accuracy on DTD using fixed and adaptive confidence weight (w_i) schemes.

Dataset	Method	Symflip 100%	Pairflip 100%
Caltech101	CoOp	1.3	0.6
	CoOp+JoAPR	81.50	84.90
	CoOp+JoAPR*	84.80	84.10
	Ours	**87.71**	**87.63**
OxfordPets	CoOp	4.7	1.4
	CoOp+JoAPR	72.60	76.90
	CoOp+JoAPR*	82.40	70.40
	Ours	**85.45**	**85.04**

Table 4: Accuracy on datasets with 100% label noise.

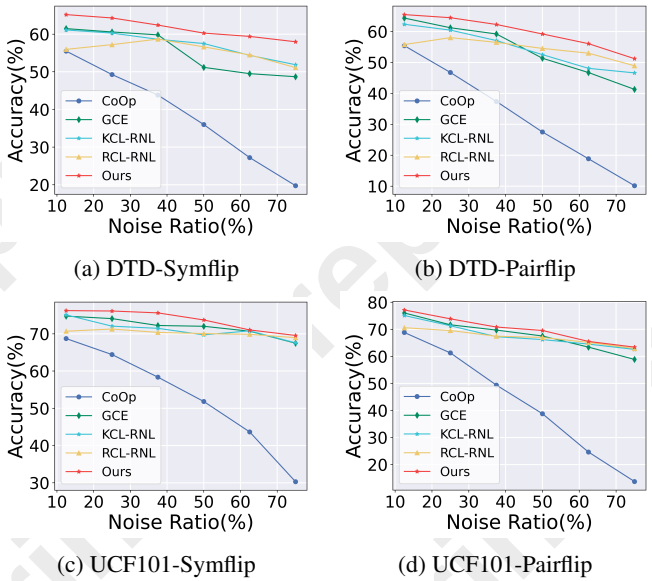
fidence threshold θ on model performance (Fig. 5). The model performs best at $\theta = 0.01$, with performance degrading as θ increases due to misclassification of clean samples.

5.5 Extreme Noisy Analysis

To assess the robustness of our method, we conducted a challenging experiment where we introduced 100% noise into the Caltech101 and OxfordPets datasets, meaning all training data were deliberately mislabeled. As shown in Table 4, our method demonstrates exceptional performance in both Symflip and Pairflip noise scenarios, significantly outperforming JoAPR. These results highlight the effectiveness and robustness of our approach under extreme noise conditions.

6 Conclusion

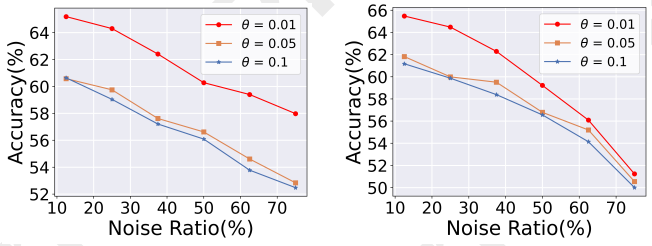
We propose a framework for tuning vision-language models with noisy labels through screening, rectifying, and re-screening strategies. Our dual-level semantic matching mechanism partitions samples into clean, ambiguous, and noisy samples, while the label rectification steps assign pseudo labels to ambiguous and noisy samples, and the re-screening step reduces self-confirmation bias through cross-validation.



(a) DTD-Symflip (b) DTD-Pairflip

(c) UCF101-Symflip (d) UCF101-Pairflip

Figure 4: Performance of CoOp variants on two datasets. By default, cross-entropy is used for training losses. 'GCE' uses GCE for loss calculation. 'KCL-RNL' and 'RCL-RNL' use two-segmented based strategies to screening and rectifying observed labels.



(a) DTD-Symflip(θ) (b) DTD-Pairflip(θ)

Figure 5: Hyper-parameters analysis on DTD dataset.

Extensive experiments show superior performance over prior methods. However, challenges remain under high noise rates and fine-grained settings, which we aim to address through noise-robust loss functions and curriculum learning.

Acknowledgements

This work is supported in part by the National Key R&D Program of China (No. 2024YFB3908503, 2024YFB3908500, 2022ZD0120100), in part by the National Natural Science Foundation of China (No. 62376206, 62176198, 62322608), in part by Guangdong Basic and Applied Basic Research Foundation (No. 2024A1515010255), and in part by Fundamental Research Funds for the Central Universities (No. QTZX25083).

References

[Arazo *et al.*, 2019] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *Proceedings of the International Conference on Machine Learning*, pages 312–321. PMLR, 2019.

[Arpit *et al.*, 2017] Devansh Arpit, Stanisław Jastrzundefinedbski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *Proceedings of the International Conference on Machine Learning*, page 233–242, 2017.

[Bae *et al.*, 2024] HeeSun Bae, Seungjae Shin, Byeonghu Na, and Il chul Moon. Dirichlet-based per-sample weighting by transition matrix for noisy label learning. In *Proceedings of the International Conference on Learning Representations*, 2024.

[Berthelot *et al.*, 2019] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Proceedings of the Advances in Neural Information Processing Systems*, 32, 2019.

[Bossard *et al.*, 2014] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Proceedings of the European Conference on Computer Vision*, pages 446–461, 2014.

[Chen *et al.*, 2024] Shengyuan Chen, Qinggang Zhang, Junnan Dong, Wen Hua, Qing Li, and Xiao Huang. Entity alignment with noisy annotations from large language models. In *Proceedings of the Advances in Neural Information Processing Systems*, 2024.

[Cimpoi *et al.*, 2014] 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.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[Fei-Fei *et al.*, 2004] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 178–178, 2004.

[Feng *et al.*, 2024] Chen Feng, Georgios Tzimiropoulos, and Ioannis Patras. Clipcleaner: Cleaning noisy labels with clip. In *Proceedings of the ACM International Conference on Multimedia*, pages 876–885, 2024.

[Guo and Gu, 2024] Yuncheng Guo and Xiaodong Gu. Joapr: Cleaning the lens of prompt learning for vision-language models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 28695–28705, June 2024.

[He *et al.*, 2016] 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.

[Helber *et al.*, 2019] 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.

[Kim *et al.*, 2024] Jang-Hyun Kim, Sangdoo Yun, and Hyun Oh Song. Neural relation graph: a unified framework for identifying label noise and outlier data. *Proceedings of the Advances in Neural Information Processing Systems*, 36, 2024.

[Krause *et al.*, 2013] 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.

[Li *et al.*, 2020] Junnan Li, Richard Socher, and Steven C. H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *Proceedings of the International Conference on Learning Representations*, 2020.

[Li *et al.*, 2022] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning*, pages 12888–12900, 2022.

[Lin *et al.*, 2024] Yexiong Lin, Yu Yao, and Tongliang Liu. Learning the latent causal structure for modeling label noise. In *Proceedings of the Advances in Neural Information Processing Systems*, 2024.

[Liu *et al.*, 2020] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Proceedings of the Advances in Neural Information Processing Systems*, 33:20331–20342, 2020.

[Lu *et al.*, 2021] Yangdi Lu, Yang Bo, and Wenbo He. Comatching: Combating noisy labels by augmentation anchoring. *arXiv:2103.12814*, 2021.

[Nguyen *et al.*, 2024] Tri Nguyen, Shahana Ibrahim, and Xiao Fu. Noisy label learning with instance-dependent outliers: Identifiability via crowd wisdom. In *Proceedings of the Advances in Neural Information Processing Systems*, 2024.

[Nilsback and Zisserman, 2008] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, 2008.

[Park *et al.*, 2023] Dongmin Park, Seola Choi, Doyoung Kim, Hwanjun Song, and Jae-Gil Lee. Robust data pruning under label noise via maximizing re-labeling accuracy. *Proceedings of the Advances in Neural Information Processing Systems*, 36:74501–74514, 2023.

[Parkhi *et al.*, 2012] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505, 2012.

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

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763, 2021.

[Sohn *et al.*, 2020] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Proceedings of the Advances in Neural Information Processing Systems*, 33:596–608, 2020.

[Soomro, 2012] K Soomro. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv:1212.0402*, 2012.

[Tian *et al.*, 2024] Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Argue: Attribute-guided prompt tuning for vision-language models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 28578–28587, 2024.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Proceedings of the Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[Wu *et al.*, 2023] 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 International Conference on Computer Vision*, pages 15488–15497, 2023.

[Xiao *et al.*, 2010] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, 2010.

[Yao *et al.*, 2021] Yazhou Yao, Zeren Sun, Chuanyi Zhang, Fumin Shen, Qi Wu, Jian Zhang, and Zhenmin Tang. Jo-src: A contrastive approach for combating noisy labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5192–5201, 2021.

[Yao *et al.*, 2024] Lujian Yao, Haitao Zhao, Zhongze Wang, Kaijie Zhao, and Jingchao Peng. CoSW: Conditional sample weighting for smoke segmentation with label noise. In *Proceedings of the Advances in Neural Information Processing Systems*, 2024.

[Yu *et al.*, 2024] Yeonguk Yu, Minhwan Ko, Sungho Shin, Kangmin Kim, and Kyoobin Lee. Curriculum fine-tuning of vision foundation model for medical image classification under label noise. In *Proceedings of the Advances in Neural Information Processing Systems*, 2024.

[Zhang and Sabuncu, 2018] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Proceedings of the Advances in Neural Information Processing Systems*, 31, 2018.

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

[Zhang *et al.*, 2024] Jiaming Zhang, Xingjun Ma, Xin Wang, Lingyu Qiu, Jiaqi Wang, Yu-Gang Jiang, and Jitao Sang. Adversarial prompt tuning for vision-language models. In *Proceedings of the European Conference on Computer Vision*, pages 56–72. Springer, 2024.

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

[Zhou *et al.*, 2022b] 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.

[Zhu *et al.*, 2024] Yilun Zhu, Jianxin Zhang, Aditya Gangrade, and Clayton Scott. Label noise: Ignorance is bliss. In *Proceedings of the Advances in Neural Information Processing Systems*, 2024.