Learning from Noisy Data Using Pretrained Vision-Language Representations

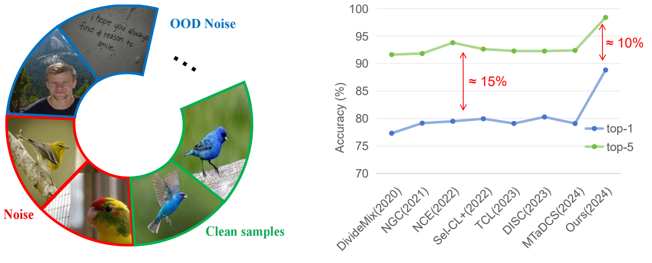
Yuqi Liao[1], Aodong Li, Yisha Chen[1], Qianfang Xu[1], Jiarui Xie[2], Anxin Li[2], Bo Xiao[1,*]
[1]School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China
[2]DOCOMO Beijing Communications Laboratories Co., Ltd., Beijing, China
{yuqii, yishachen, xuqianfang, xiaobo}@bupt.edu.cn,
{xiejr,liax}@docomolabs-beijing.com.cn

***Abstract*—Real-world image data often contains complex noise patterns that compromise the generalization capability of deep learning models. Existing methods for learning from noisy data typically rely on oversimplified assumptions about labeling noise, failing to capture the complexity of real-world noise. Our investigation reveals that real-world noise consists of both out-of-domain (OOD) and in-domain (ID) noise, frequently due to web crawler imprecision, lack of domain knowledge, and/or annotator oversight. Insufficient consideration of either part can negatively affect the performance of models. Furthermore, the prevalence of sub-class label errors in images with similar visual appearances presents a challenge to fine-grained classification tasks. In this paper, we propose a novel but simple denoising framework that leverages textual labels and pretrained vision-language models to mitigate both OOD noise and ID noise without relying on restrictive assumptions. Specifically, our approach comprises three sequential stages: 1) identification and exclusion of OOD samples using vision-language similarity distribution; 2) identification of a clean dataset through analysis of consistent labels in augmented images; and 3) training of the final model using both clean and noisy labels in a semi-supervised manner. Extensive experiments demonstrate the efficacy of our proposed method, which outperforms state-of-the-art approaches. We achieve an average performance improvement of 11% on datasets with synthetic noise and a notable 6% improvement on datasets with real-world noise. The source code and the Appendix are available at https://github.com/LiaoYuqi2000/Multimodal-LNL.**

***Index Terms*—Real-world noisy data, Vision-language representations, Fine-grained label errors**

I. Introduction

Most real-world datasets, exemplified by WebVision [1] and Food-101N [2], contain a significant level of noise [3]. This noise frequently stems from imperfections in the data collection process, including web crawler imprecision, annotators' lack of domain knowledge, and/or general oversight. Deep learning models, known for their strong expressivity, are often the preferred choice for capturing complex patterns in real-world data, especially in visual domains. But this same expressivity makes them vulnerable to noisy or erroneous data labels, as they tend to memorize the noise, thereby compromising their generalization performance [4]. Consequently, learning effective deep learning models from real world data necessitates robust training algorithms that can mitigate the impact of such noise.



(a) WebVision - Indigo Finch

(b) WebVision Validation Set Accuracy

Fig. 1. (a): Example clean and noisy images from the indigo finch category in the Webvision dataset; (b): The top-1 and top-5 accuracy of SOTA robust training methods on WebVision validation set.

Real-world noise consists of both out-of-domain (OOD) and in-domain (ID) noise. For example, in Fig. 1 (a), in a bird classification task, a picture of an irrelevant human is considered OOD noise, while a picture of another bird species mislabeled as indigo finch is referred to as ID noise. Both types of noise are harmful to deep learning models. The presence of OOD data during training may also lead the model to learn incorrect representations of in-distribution data and focus on spurious features. Hence adverse effects such as biased decision boundary, poor generalization, increased vulnerability to adversarial attacks, and degraded uncertainty estimation can occur. Existing methods on learning with noisy labels (LNL) mostly assume only ID noise is present and cannot effectively mitigate OOD noise [5]–[7].

In-domain (ID) mislabeling noise in real-world data is particularly prevalent in fine-grained classification datasets. While it's unlikely to mislabel pictures of birds as ”planes”, it's quite possible to mix up great white sharks with tiger sharks. Objects of fine-grained classes often exist in similar environments, resulting in images with highly similar visual appearances. Correctly recognizing these fine-grained classes with deep neural network classifiers is already challenging. This ID confusion, exhibiting as mislabeling noise, is further incorporated into datasets through imprecise web crawlers and human annotators lacking domain knowledge, worsening the situation. This challenge of robust training is evident in the predictions of DivideMix [8] in Fig. 2. Although the resulting model can broadly classify the given two test images as turtles and sharks, it fails to identify them specifically as a terrapin or

[*] Corresponding author.

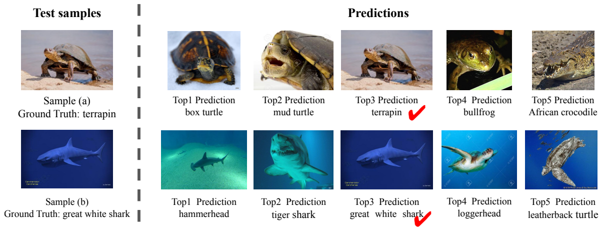


Fig. 2. DivideMix's prediction results for samples (a) and (b). Left: Test images; Right: Predicted categories from top-1 to top-5, each accompanied by an example image. Correct predictions are indicated by a red check mark.

a great white shark. This inefficacy is further illustrated by the significant performance gap between top-1 and top-5 accuracy among many popular robust training algorithms (Fig. 1 (b)).

In this paper, we propose a new robust training framework that utilizes an additional modality–label texts–and robust image representations to address the challenges in learning from real-world noisy datasets. This textual modality provides rich semantic information about image contents, complementing the pixel-level data. By checking the consistency between label texts and image contents, we establish an effective denoising method. The second component of our method, robust image representations, captures accurate visual concepts, providing distinguishable features for fine-grained classes and facilitating the learning process of downstream tasks. To achieve this, we leverage the vision-language model CLIP [9], which is pretrained on large-scale self-supervised datasets, to acquire both the textual modality and robust image representations.

Specifically, our proposed framework consists of three key stages. In Stage I, OOD samples are identified and removed based on the vision-language similarity distribution. In Stage II, we distinguish clean and noisy labels by evaluating the consistency between multimodal model predictions and provided labels. In Stage III, the model is trained using both clean labels and corrected noisy labels. Extensive ablation experiments are conducted to validate the effectiveness of each stage.

The main contributions are summarized as follows.

- We propose a simple framework to address OOD data and fine-grained noise simultaneously without relying on restrictive assumptions.
- Our model integrates text information to enhance fine-grained feature discrimination.
- Extensive experiments demonstrate that our method significantly outperforms state-of-the-art (SOTA) methods across various scenarios, achieving an average improvement of 11% on two synthetic noise datasets and 6% on three real-world noise datasets.

II. Related Works

Numerous valuable studies have emerged in LNL fields, mainly focusing on noise transition matrix, robust loss function, label correction and sample selection. These methods perform well on synthetic noisy datasets. For example, ProMix [10] achieves 93.4% accuracy on CIFAR-10 even with 90% synthetic noise. However, they struggle to achieve significant improvements on real-world noisy datasets. For instance, the annual top-1 accuracy growth on the WebVision [1] validation set has remained under 1%, as shown in Fig. 1 (b). This is mainly because most existing methods are designed for synthetic noise, which oversimplifies real-world scenarios by focusing only on simple ID noise while ignoring the OOD and fine-grained noise in real-world data. The following sections will introduce and analyze the limitations of each method.

Noise transition matrix. Noise transition matrix T represents the probability of noise transfer across different classes or instances. Estimating T is challenging due to its high degree of freedom. To address this, researchers usually simplify T using various assumptions. For example, Liu et al. [11] uses anchor points, and MEIDTM [7] makes geometric assumptions about T. These assumptions inevitably introduce approximation error. Additionally, these methods are incapable of handling OOD data.

Robust loss function. The robust loss function is designed to reduce the impact of noisy samples. CTRR [12] modifies contrastive learning loss from the perspective of gradients, and NCR [13] introduces neighbor consistency-based regularization. Although these methods offer theoretical guarantees, their reliance on specific scenarios or data distributions restricts the practical effectiveness.

Label correction. The purpose of label correction is to estimate the true labels for all training samples. PENCIL [14] updates learnable corrected labels through back-propagation. Joint-optimization [15] uses a regularization term to reduce incorrect corrections. However, these methods may mistakenly correct accurate labels, leading to error accumulation, and the labels of OOD samples cannot be correctly modified.

Sample selection. Sample selection methods utilize both clean and relabeled noisy data, often outperforming other approaches. However, they also exhibit certain limitations. UNICON [16] assumes a uniform noise level across categories, which is unrealistic. PNP [17] employs a noise predictor for OOD detection, but its reliance on unsupervised information during training constrains the detection performance. SNSCL [18] applies supervised contrastive learning to distinguish fine-grained features, but achieves only marginal improvements. Recent works, such as DynaCor [19] and MTaDCS [20], do not take realistic OOD noise or fine-grained noise into account.

Our approach belongs to sample selection. Unlike previous methods, we leverage pretrained multimodal features to address both OOD and fine-grained noise without assumptions.

III. Method

A. Overview

We address a practical problem of learning a classification model from real-world noisy data. The noisy dataset is denoted as $D = \{(x_i, y_i, t_i)\}_{i=1}^{N}$, where x_i is the i-th image sample, y_i is its scalar label, t_i is the corresponding text label and N represents the dataset size. As discussed in I, real-world noise typically includes OOD and fine-grained ID noise. Our

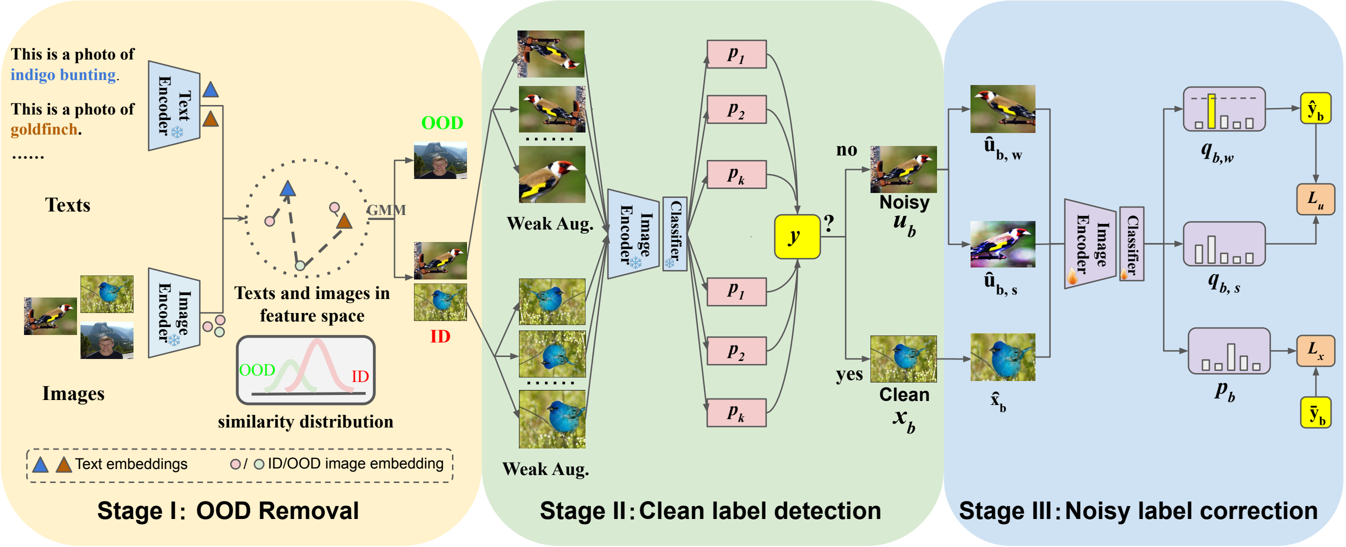


Fig. 3. Overview of the entire framework. The framework consists of three stages: OOD removal, clean label detection and noisy label correction.

method utilizes the text information t_i to enhance fine-grained feature discrimination.

The overall framework is shown in Fig. 3. The entire model consists of a text encoder, an image encoder, and a classifier. The text and image encoders are initialized with the pre-trained parameters of CLIP. Inspired by [21], we use the label text embeddings extracted from the text encoder to initialize the classifier weights. Specific process is provided in Appendix A. This practice integrates the textual information into the model while accelerating convergence.

Our method comprises three stages: OOD removal, clean label detection, and noisy label correction. The first two stages are offline, with the image encoder, text encoder, and classifier frozen. In Stage III, the image encoder and classifier are trained jointly. Typically, we use CLIP's strong zero-shot capability for OOD removal and clean label detection. If CLIP under-performs on a dataset, we fine-tune the model on this noisy dataset for a few epochs to align the text and image embeddings. The fine-tuned model is then used across all stages. To prevent overfitting to noisy labels, we limit the number of epochs for fine-tuning.

B. Stage I: OOD Removal

OOD data can hurt machine learning model performance because they may guide the model to learn incorrect representations. We propose to detect OOD samples by detecting low density regions of feature similarities between images and texts. Given a dataset with C classes, x_i denotes an image and t_j denotes the class name corresponding to label j. The CLIP model encodes x_i and t_j using its image encoder p_{image} and text encoder p_{text}, represented as $I(x_i) = p_{image}(x_i)$ and $T(t_j) = p_{text}(prompt + t_j)$, where the adopted prompt is "This is a photo of". The similarity between image x_i and texts in this dataset is computed by:

$$sim_i = \max_{j \leq C} \frac{I(x_i) \cdot T(t_j)}{\|I(x_i)\| \cdot \|T(t_j)\|} \tag{1}$$

Similar to [6], [8], we use a two-component Gaussian Mixture Model (GMM) to fit the similarity distribution. For each sample x_i, we employ the Expectation-Maximization (EM) algorithm to estimate the posterior probability $p(g|sim_i)$ of it being classified as ID, where g denotes the Gaussian component with the larger mean (i.e., the component with the larger similarity). Sample x_i is classified as OOD and discarded to enhance dataset quality if $p(g|sim_i)$ is below the threshold th_o. To minimize misclassification of ID samples and retain potentially useful data, th_o is usually set low.

C. Stage II: Clean Label Detection

After removing OOD data, we will collect a sample of clean data with high confidence as our seed training set. We detect clean data by assessing the consistency between the provided labels and predictions of different augmentations. Note we employ a classifier initialized from text embeddings and keep it frozen at this stage. For clean data, the model's predictions across different augmentations should be consistent and align with the provided label. We quantify this consistency by calculating the percentage of predictions that are consistent with the given label among K augmentations. Specifically, the consistency score is defined as:

$$score(x_i) = \frac{\sum_{k=1}^{K} \mathbb{I}[p_{i,k} = y_i]}{K} \tag{2}$$

$$\begin{cases} x_i \in clean, & score(x_i) \geq th \\ x_i \in noisy, & score(x_i) < th \end{cases} \tag{3}$$

where $p_{i,k}$ denotes the scalar prediction result for the k-th data augmentation of sample x_i, and y_i is the original scalar label.

A higher score indicates the sample is more likely clean. We then use a large threshold th to separate ID samples into clean and noisy subsets, ensuring high precision for clean samples. The noisy subsets are reused in Stage III.

D. Stage III: Noisy Label Correction

To fully leverage both clean and noisy data, we adopt robust semi-supervised learning, treating clean samples as labeled data and noisy samples as unlabeled data. In the first few epochs, we warm up the training with clean samples to help the model achieve a good initial classification performance. Subsequently, similar as [22], we train on clean samples using standard cross-entropy loss and learn from noisy samples with high-confidence pseudo-labels. Specifically, let $X = \{(x_b, y_b)\}_{b=1}^{B_x}$ be a batch of B_x clean samples. We perform a weak augmentation using crop-and-flip for each sample x_b, resulting in $\hat{x}_b$. Given a model with parameters θ, the model's output is denoted by $f(\hat{x}_b; \theta)$. The softmax function is denoted by $\sigma(x)$, where x is the input vector. Additionally, we denote the cross-entropy between two probability distributions p and q as $\mathcal{H}(p, q)$. The loss of the clean samples can be computed by:

$$L_x = \frac{1}{B_x} \sum_{b=1}^{B_x} \mathcal{H}(\sigma(f(\hat{x}_b; \theta)), \mathbf{e}_{[y_b]}) \tag{4}$$

where $\mathbf{e}_{[y_b]}$ is the one-hot vector with one at the y_b-th position. As for noisy samples, let $U = \{u_b\}_{b=1}^{B_u}$ be a batch of B_u unlabled examples. We conduct a weak augmentation using crop-and-flip and a strong augmentation following RandAugment [23]($n = 2, m = 10$) for each sample u_b, resulting in $\hat{u}_{b,w}$ and $\hat{u}_{b,s}$ respectively. We first generate a pseudo-label $\hat{y}_b$ with the model output $q_{b,w} = f(\hat{u}_{b,w}; \theta)$ and temperature coefficient T:

$$\hat{y}_b = \arg\max \sigma\left(\frac{q_{b,w}}{T}\right) \tag{5}$$

A pseudo-label $\hat{y}_b$ is considered to have high confidence if the maximum value of $\sigma(\frac{q_{b,w}}{T})$ exceeds a predefined threshold τ. We then convert high-confidence pseudo-label into a one-hot format and use it to supervise the sample $\hat{u}_{b,s}$. The model output for $\hat{u}_{b,s}$ is $q_{b,s} = f(\hat{u}_{b,s}; \theta)$. The loss for noisy samples is defined as:

$$L_u = \frac{1}{B_u} \sum_{b=1}^{B_u} \mathbb{I}\{\max(\sigma(\frac{q_{b,w}}{T})) \geq \tau\} \mathcal{H}(\sigma(q_{b,s}), \mathbf{e}_{[\hat{y}_b]}) \tag{6}$$

Lastly, our overall training objective function is:

$$L = L_x + \lambda_u L_u \tag{7}$$

where λ_u is used to control the strength of the noisy loss.

To further enhance the model's generalization, we employ a weight-space ensemble, combining the trained parameters with the initial model parameters using a linear weighted approach.

IV. Experiments

In this section, we validate our method's effectiveness on two synthetic and three real-world noisy datasets. All experiments are implemented on two GeForce RTX3090 GPU and PyTorch 1.11.0.

TABLE I
Comparison with SOTA methods on CIFAR-10 and CIFAR-100 with symmetric and asymmetric noise from different levels.

Dataset	CIFAR-10					CIFAR-100			
Methods/Noise ratio	Sym.				Asym.	Sym.			
	20%	50%	80%	90%	40%	20%	50%	80%	90%
MOIT [25]	94.1	91.1	75.8	70.1	93.2	75.9	70.1	51.4	24.5
UNICON [16]	96.0	95.6	93.9	90.8	94.1	78.9	77.6	63.9	44.8
DivideMix [8]	96.1	94.6	93.2	76.0	93.4	77.3	74.6	60.2	31.5
CTRR [12]	93.1	-	83.7	81.7	89.0	70.1	-	43.7	-
ProMix [10]	97.7	97.4	95.5	93.4	96.6	82.6	80.1	69.4	42.9
MTaDCS [20]	95.9	95.4	92.7	89.6	94.2	77.1	74.1	65.7	61.4
CLIP (zero-shot) [9]	97.9					85.8			
Ours	**99.3**	**99.3**	**99.1**	**98.9**	**99.1**	**92.3**	**92.0**	**91.2**	**90.2**
Ours + Ensemble	**99.4**	**99.4**	**99.2**	**99.0**	**99.3**	**92.5**	**92.1**	**91.3**	**90.3**

A. Evaluation on Synthetic Noisy Datasets

The synthetic noisy dataset is created by corrupting clean labels, enabling control over the noise rate and type. We use CIFAR-10 [24] and CIFAR-100 [24] as the clean datasets.

Datasets. CIFAR-10 and CIFAR-100 each contain 50K training images and 10K test images of size 32 × 32. We employ two types of noise: symmetric and asymmetric. Symmetric noise uniformly flips labels to other categories. Asymmetric noise flips labels based on specified similar class-pairs (e.g., BIRD→PLANE, DOG↔CAT).

Training details. In this section, we only perform Stage II and III because synthetic noise does not include OOD. The hyperparameters are detailed in Appendix B.

Quantitative results. Table. I shows the performance on CIFAR-10/100 dataset with different types and levels of synthetic label noise. We select the epoch with the best validation performance for testing. Although our approach is designed for real-world noise, it is also applicable and useful for synthetic noise. As can be seen, our method significantly outperforms other methods across all noise settings, especially in high noise ratios. And the model ensemble further improves the accuracy.

B. Evaluation on Real-world Noisy Datasets

The real-world noise, arising from crawlers and annotation errors, is a practical issue to be addressed. We validate our method on WebVision [1], Animal-10N [26] and Food-101N [2].

Datasets. WebVision contains more than 2.4 million images crawled from Google and Flickr using 1,000 classes in ImageNet ILSVRC12 as query words. The estimated label noise rate is 20%. Following [8], we also train on the first 50 classes of the Google image subset, and report the top-l and top-5 accuracy on both Webvision and ImageNet ILSVRC12 validation set. Animal-10N has 10 animal classes with confusing appearances, containing 50k training and 5k testing images. Noise labels are caused by human-labeled mistakes and the noise level is around 8%. Only ID noise is observed in it. Food-101N, a food dataset classified into 101 classes, consists of 310k training images collected from web. Both ID and OOD noise exist in it and the noise ratio is around 20%. The model is trained on Food-101N and evaluated on Food-101 with 25k test images [5].

TABLE II
ACCURACY(%) ON WEBVISION AND ILSVRC12 VALIDATION SETS. THE MODEL IS TRAINED ON WEBVISION-50.

Methods	WebVision		ILSVRC12	
	top-1	top-5	top-1	top-5
DivideMix (ECCV, 2020) [8]	77.32	91.64	75.20	90.84
NGC (ICCV, 2021) [27]	79.16	91.84	74.44	91.04
NCE (ECCV, 2022) [28]	79.5	93.8	76.3	94.1
Sel-CL+ (CVPR, 2022) [29]	79.96	92.64	76.84	93.04
TCL (CVPR, 2023) [6]	79.1	92.3	75.4	92.4
DISC (CVPR, 2023) [5]	80.28	92.28	77.44	92.28
MTaDCS (ECCV, 2024) [20]	79.1	92.4	76.2	92.6
CLIP (zero-shot) [9]	85.88	98.16	86.04	98.80
Ours	**88.52**	**98.16**	**87.16**	**98.36**
Ours + Ensemble	**88.84**	**98.40**	**88.12**	**98.84**

TABLE III
COMPARISON WITH THE SOTA METHODS ON ANIMAL-10N.

Methods	Acc
SELFIE (ICML, 2019) [26]	81.8
Co-learning (ACM MM, 2021) [30]	82.95
SPR (CVPR, 2022) [31]	86.8
CTRR (CVPR, 2022) [12]	86.71
DISC (CVPR, 2023) [5]	87.1
ProMix (IJCAI, 2023) [10]	89.98
CLIP (zero-shot) [9]	73.84
Ours	**94.18**
Ours + Ensemble	**94.70**

TABLE IV
COMPARISON WITH THE SOTA METHODS ON FOOD-101N.

Methods	Acc
DivideMix (ECCV, 2020) [8]	85.88
Co-learning (ACM MM, 2021) [30]	87.57
Jo-SRC (CVPR, 2021) [32]	86.66
PNP-soft (CVPR, 2022) [17]	87,50
DivideMix+SNSCL (CVPR, 2023) [18]	86.40
DISC (CVPR, 2023) [5]	89.02
CLIP (zero-shot) [9]	93.71
Ours	**94.50**
Ours + Ensemble	**94.78**

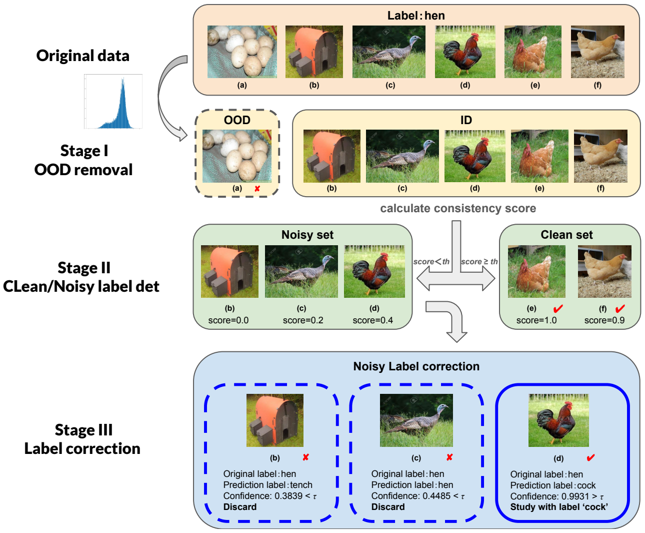


Fig. 4. Visualization of the model processing pipeline. The example images are sourced from the category “hen” of the WebVision dataset.

Training details. A comprehensive three-stage process is applied to WebVision and Food-101N based on the noise types in the datasets, while for Animal-10N, only Stage II and III are used. Additionally, CLIP exhibits poor zero-shot performance on Animal-10N, so we first warm-up the model for 5 epochs on the entire noisy dataset, as described in Section III-A. The hyperparameter details are shown in Appendix B.

Quantitative results. The results for WebVision, Animal-10N and Food-101N datasets are presented in Table. II, Table. III, and Table. IV, respectively. Our method improves test accuracy by 4% to 10%, demonstrating its effectiveness in handling real-world noise. Model ensemble further boosts performance by 0.3% to 2%. Additionally, we reduce the gap between top-1 and top-5 accuracy on Webvision and ILSVRC12 from 15% to 10%, indicating the positive role of text features in capturing fine-grained distinctions.

Qualitative results. We select six images labeled as “hen” from WebVision to visualize the effect of each stage, as shown in Fig. 4. Firstly, we utilize a two-component GMM to partition sample (a) into the OOD subset and the remaining samples into the ID subset. The OOD sample (a) is then discarded. Secondly, we compute consistency scores for ID samples using (2). Samples with lower scores, namely (b), (c) and (d), are categorized as noisy, while those with higher scores, namely (e) and (f), are classified as clean. Thirdly, we assign pseudo labels to noisy samples using (5). Due to the low confidence of pseudo labels for samples (b) and (c), they are not used in this iteration. Thus, the undetected OOD sample (b) from Stage I does not affect the training. Finally, samples (d), (e) with original label “hen” and sample (f) with pseudo-label “cock” are employed for model training.

C. Ablation Study and Discussions

In this section, the model consists of an image encoder and a classifier, and the “baseline” refers to training with the entire noisy dataset. We progressively integrate the three stages and model ensemble into the baseline model to assess their effectiveness. The experiments are conducted on real-world noisy datasets and the result is presented in Table. V.

The effect of OOD removal. We plot the histogram of image-text similarities for WebVision, shown in the top-left corner of Fig. 4. The horizontal axis represents the normalized similarity and the vertical axis represents the number of samples. The distribution approximately conforms a two-component Gaussian mixture. Separating these components helps distinguish between OOD and ID data. We train the model using all data and ID data separately. The performance improves after removing OOD (r2 vs. r3 in Table. V), demonstrating the effectiveness of the OOD detection module.

The effect of clean sample detection. We remove noisy samples from the ID subset and train the model only using clean samples. Compared to training on the entire ID subset, we observe a notable enhancement (see Table. V, r3 vs. r4). This indicates our method effectively filters out noisy samples and reduces the noise ratio in the training dataset.

.

TABLE V
ABLATION EXPERIMENT RESULTS ON WEBVISION, ANIMAL-10N, FOOD-101N. ME: MODEL ENSEMBLE. NUM: NUMBER OF TRAINING SAMPLES.

Index	Components							WebVision		ILSVRC12		Animal-10N		Food-101N	
	Zero-shot	Baseline	Stage I	Stage II	Stage III	ME	Num	top-1	top-5	top-1	top5	Num	Acc(best/last)	Num	Acc(best/last)
r1	✓						-	85.88	98.16	86.04	98.80	-	73.84	-	93.71
r2		✓					65944	86.52	97.04	83.08	97.20	50000	93.56 / 92.52	310009	92.17 / 89.95
r3			✓				53654	87.08	98.12	84.28	98.00	-	-	295575	92.79 / 90.38
r4			✓	✓			39247	87.72	98.12	86.36	98.36	40050	93.74 / 93.74	220361	94.38 / 94.38
r5			✓	✓	✓		53654	88.52	98.16	87.16	98.36	50000	94.18 / 94.14	295575	94.50 / 94.40
r6			✓	✓	✓	✓	53654	88.84	98.40	88.12	98.84	50000	94.70	295575	94.78

The effect of label correction. Label correction assigns high-confidence pseudo-labels to noisy samples, allowing the model to learn from both clean and noisy data. The increase in accuracy shown in Table. V (r4 vs. r5) suggests that the noisy labels have been effectively rectified.

Detailed analyses of Stage I and II, as well as the backbone's impact, are presented in Appendix C to E. Hyperparameter sensitivity is covered in Appendix F, and the method's limitations are discussed in Appendix G.

V. CONCLUSION

In this paper, we identify that the performance of existing methods is limited because they neglect the OOD and fine-grained noise present in real-world scenarios. To address these issues, we propose a novel but simple framework that leverages pretrained vision-language representations to handle OOD and fine-grained noise without assumptions. Extensive experiments verify its effectiveness.

VI. ACKNOWLEDGE

This research was supported by the National Natural Science Fund of China(Grant Nos.62076031 and 62076036).

REFERENCES

[1] W Li, L Wang, W Li, E Agustsson, and L Van Gool, "Webvision database: Visual learning and understanding from web data. arxiv 2017," *arXiv preprint arXiv:1708.02862*.

[2] Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang, "Cleannet: Transfer learning for scalable image classifier training with label noise," in *CVPR*, 2018, pp. 5447–5456.

[3] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee, "Learning from noisy labels with deep neural networks: A survey," *IEEE transactions on neural networks and learning systems*, vol. 34, no. 11, pp. 8135–8153, 2022.

[4] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[5] Yifan Li, Hu Han, Shiguang Shan, and Xilin Chen, "Disc: Learning from noisy labels via dynamic instance-specific selection and correction," in *CVPR*, 2023, pp. 24070–24079.

[6] Zhizhong Huang, Junping Zhang, and Hongming Shan, "Twin contrastive learning with noisy labels," in *CVPR*, 2023, pp. 11661–11670.

[7] De Cheng et al., "Instance-dependent label-noise learning with manifold-regularized transition matrix estimation," in *CVPR*, 2022, pp. 16630–16639.

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

[9] Alec Radford et al., "Learning transferable visual models from natural language supervision," in *ICML*. PMLR, 2021, pp. 8748–8763.

[10] Haobo Wang, Ruixuan Xiao, Yiwen Dong, Lei Feng, and Junbo Zhao, "Promix: Combating label noise via maximizing clean sample utility," *arXiv preprint arXiv:2207.10276*, 2022.

[11] Tongliang Liu and Dacheng Tao, "Classification with noisy labels by importance reweighting," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 38, no. 3, pp. 447–461, 2015.

[12] Li Yi, Sheng Liu, Qi She, A Ian McLeod, and Boyu Wang, "On learning contrastive representations for learning with noisy labels," in *CVPR*, 2022, pp. 16682–16691.

[13] Ahmet Iscen, Jack Valmadre, Anurag Arnab, and Cordelia Schmid, "Learning with neighbor consistency for noisy labels," in *CVPR*, 2022, pp. 4672–4681.

[14] Kun Yi and Jianxin Wu, "Probabilistic end-to-end noise correction for learning with noisy labels," in *CVPR*, 2019, pp. 7017–7025.

[15] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa, "Joint optimization framework for learning with noisy labels," in *CVPR*, 2018, pp. 5552–5560.

[16] Nazmul Karim, Mamshad Nayeem Rizve, Nazanin Rahnavard, Ajmal Mian, and Mubarak Shah, "Unicon: Combating label noise through uniform selection and contrastive learning," in *CVPR*, 2022, pp. 9676–9686.

[17] Zeren Sun et al., "Pnp: Robust learning from noisy labels by probabilistic noise prediction," in *CVPR*, 2022, pp. 5311–5320.

[18] Qi Wei et al., "Fine-grained classification with noisy labels," in *CVPR*, 2023, pp. 11651–11660.

[19] Suyeon Kim et al., "Learning discriminative dynamics with label corruption for noisy label detection," in *CVPR*, 2024, pp. 22477–22487.

[20] Qingzheng Huang et al., "Mtadcs: Moving trace and feature density-based confidence sample selection under label noise," in *ECCV*. Springer, 2024, pp. 178–195.

[21] Mitchell Wortsman et al., "Robust fine-tuning of zero-shot models," in *CVPR*, 2022, pp. 7959–7971.

[22] Kihyuk Sohn et al., "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in neural information processing systems*, vol. 33, pp. 596–608, 2020.

[23] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *CVPR workshops*, 2020, pp. 702–703.

[24] Alex Krizhevsky, Geoffrey Hinton, et al., "Learning multiple layers of features from tiny images," 2009.

[25] Diego Ortego, Eric Arazo, Paul Albert, Noel E O'Connor, and Kevin McGuinness, "Multi-objective interpolation training for robustness to label noise," in *CVPR*, 2021, pp. 6606–6615.

[26] Hwanjun Song, Minseok Kim, and Jae-Gil Lee, "Selfie: Refurbishing unclean samples for robust deep learning," in *ICML*. PMLR, 2019, pp. 5907–5915.

[27] Zhi-Fan Wu et al., "Ngc: A unified framework for learning with open-world noisy data," in *ICCV*, 2021, pp. 62–71.

[28] Jichang Li, Guanbin Li, Feng Liu, and Yizhou Yu, "Neighborhood collective estimation for noisy label identification and correction," in *ECCV*. Springer, 2022, pp. 128–145.

[29] Shikun Li, Xiaobo Xia, Shiming Ge, and Tongliang Liu, "Selective-supervised contrastive learning with noisy labels," in *CVPR*, 2022, pp. 316–325.

[30] Cheng Tan, Jun Xia, Lirong Wu, and Stan Z Li, "Co-learning: Learning from noisy labels with self-supervision," in *ACM MM*, 2021, pp. 1405–1413.

[31] Yikai Wang, Xinwei Sun, and Yanwei Fu, "Scalable penalized regression for noise detection in learning with noisy labels," in *CVPR*, 2022, pp. 346–355.

[32] Yazhou Yao et al., "Jo-src: A contrastive approach for combating noisy labels," in *CVPR*, 2021, pp. 5192–5201.