Webly-Supervised Fine-Grained Recognition with Partial Label Learning

Yu-Yan Xu[1,2] , **Yang Shen**[1] , **Xiu-Shen Wei**[1,2*] and **Jian Yang**[1]

[1] Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Lab of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology

[2]State Key Laboratory of Integrated Services Networks, Xidian University

{xuyy, shenyang_98, weixs, csjyang}@njust.edu.com

Abstract

The task of webly-supervised fine-grained recognition is to boost recognition accuracy of classifying subordinate categories (*e.g.*, different bird species) by utilizing freely available but noisy web data. As the label noises significantly hurt the network training, it is desirable to distinguish and eliminate noisy images. In this paper, we propose two strategies, *i.e.*, open-set noise removal and closed-set noise correction, to both remove such two kinds of web noises w.r.t. fine-grained recognition. Specifically, for open-set noise removal, we utilize a pre-trained deep model to perform deep descriptor transformation to estimate the positive correlation between these web images, and detect the open-set noises based on the correlation values. Regarding closed-set noise correction, we develop a top-k recall optimization loss for firstly assigning a label set towards each web image to reduce the impact of hard label assignment for closed-set noises. Then, we further propose to correct the sample with its label set as the true single label from a partial label learning perspective. Experiments on several webly-supervised fine-grained benchmark datasets show that our method obviously outperforms other existing state-of-the-art methods.

1 Introduction

Deep Neural Networks have recently lead to significant progress in fine-grained tasks [Horn *et al.*, 2018]. However, constructing a fine-grained dataset can be an extremely difficult work since distinguishing subtle differences among fine-grained categories usually requires the domain-specific expert knowledge [Wei *et al.*, 2021b]. To further reduce the reliance on manual annotations as well as to learn more practical fine-grained models, training directly from web images is becoming increasingly popular [Yao *et al.*, 2021].

Approaches for learning models driven by web images, *aka* Webly-Supervised Learning (WSL), crawl freely available web images from public websites by using category names as queries [Yi and Wu, 2019]. But, whereas web images are cheaper and easier to be collected via image search, when applying the classifier learnt based only on web images to the test images, the performance will drop sharply due to the error-prone automatic tagging system or non-expert annotations [Niu *et al.*, 2018]. In addition, general dataset annotators can easily annotate coarse-grained images, but annotations of fine-grained images require the advanced expertise. Given all this, the task of webly-supervised fine-grained recognition is challenging and deserves in-depth studies.

There generally exists two kinds of label noise with regard to WSL tasks, *i.e.*, the closed-set label noise and the open-set label noise [Sun *et al.*, 2021a]. In fine-grained tasks, the open-set label noise usually caused by "cross-domain", which means those images are not of any categories in the same fine-grained domain. The closed-set label noise then refer to the portion of images that have the wrong labels within a fine-grained category, *e.g.*, a `least auklet` image but labeled with `parakeet auklet`. A simple and effective approach to deal with normal label noise is to perform samples selection that separates clean instances from noisy samples. These works [Han *et al.*, 2018] can not exploit the in-distribution noisy instances for representation learning and take the risk of discarding some clean images, of which is more serious in fine-grained problems. In fine-grained tasks, some researchers then propose a learning paradigm to train robust deep fine-grained models [Sun *et al.*, 2021b; Sun *et al.*, 2021a] from noisy web images. However, these works do not solve the problem of closed-set label noise.

To make full use of the noisy webly datasets, we first design the open-set label noise removal strategy to eliminate images that mistakenly treat some unknown classes as target classes, which helps preserve noisy-free images and closed-set images. A label with high relative confidence is not necessarily the ground-truth, especially in a noisy label environment. Therefore, compared with the information of the relative confidence, we prefer to utilize more easily reachable information in the noisy label learning for closed-set noise label correction. Partial label learning (PLL) [Cour *et al.*, 2011] deals with the problem where each training example is associated with a set of candidate labels including the ground-truth. Thus, as long as one of the label set is ground-truth, such samples can be used in PLL process. In this paper, we propose a strategy of progressively narrowing the range of

[*] Corresponding author. The first two authors contributed equally to this work.

correct categories within closed-set noisy images and leverage PLL to extract the correct categories from the sets of candidate categories. Concretely, we find the common patterns from webly fine-grained data and calculate the correlation between samples and common patterns. Then, we utilize correlation values to detect open-set label noise and remove it. For the correction of closed-set label noise, correction of closed-set noisy samples with hard labels is difficult because high-confidence category may not be ground-truth. Therefore, we utilize affordable information, *i.e.*, a label set containing the ground-truth. After that, we correct closed-set noisy label via the PLL. The correct labels are obtained using an encoding and decoding mechanism with error correction capability. This method prevents the prediction category from being misleading by fine-grained false positive labels that co-occurring with ground-truth in a label set.

To the best of our knowledge, this is the first work to leverage partial label learning to deal with webly-supervised fine-grained recognition. Our major contributions are three-fold:

- We propose an open-set label noise removal strategy and a closed-set label noise correction strategy to deal with the practical but challenging webly-supervised fine-grained recognition task.
- We particularly correct the closed-set label noise with label sets via partial label learning.
- We conduct comprehensive experiments on four webly-supervised fine-grained benchmark datasets, and our proposed method achieves superior recognition accuracy over competing solutions on these datasets.

2 Related Work

2.1 Fine-Grained Recognition

Fine-grained recognition is a fundamental research aspect of fine-grained image analysis [Wei *et al.*, 2021b], which focuses on distinguishing numerous visually similar subordinate categories that belong to the same basic category, *e.g.*, the fine distinction of animal species [Horn *et al.*, 2018], vehicle species [Maji *et al.*, 2013], etc. Existing fine-grained recognition methods can be roughly separated into three main paradigms, *i.e.*, 1) recognition by localization-classification sub-networks, 2) recognition by end-to-end feature encodings and 3) recognition with external information. Specifically, localization-classification sub-networks, *e.g.*, [Zhu *et al.*, 2020] was designed to obtain the discriminative semantic parts of fine-grained objects, and constructed a mid-level representation corresponding to semantic parts for final classification to bring accuracy improvement. Methods of end-to-end feature encodings attempted to learn a unified but discriminative feature representation to model subtle differences between fine-grained categories, *e.g.*, performing high-order feature interactions [Wei *et al.*, 2021a] and designing a loss function [Sun *et al.*, 2020a], etc. Recognition with external information tried to improve fine-grained recognition accuracy by leveraging the power of extra supervisions [Song *et al.*, 2020]. Recently, many trials would like to employ web data (webly-supervised images) to further improve the recognition accuracy from diverse perspectives, *e.g.*, developing robust loss functions [Hendrycks *et al.*, 2018] and estimating the noise transition matrix [Patrini *et al.*, 2017]. While, in this paper, we are the first to deal with webly-supervised fine-grained recognition with a partial label learning [Cour *et al.*, 2011] based method.

2.2 Webly-Supervised Learning

Training image recognition models with web images (*i.e.*, Webly-Supervised Learning, WSL) usually results in poor performance due to the presence of label noises and data bias [Sun *et al.*, 2020b; Zhang *et al.*, 2020]. The research efforts of WSL can be categorized into two groups, including loss correction [Yi and Wu, 2019] and sample selection [Han *et al.*, 2018; Yu *et al.*, 2019; Wei *et al.*, 2020]. In concretely, in the literature of loss correction, [Patrini *et al.*, 2017] introduced a loss function to estimate the noise transition matrix, and [Hendrycks *et al.*, 2018] designed the gold loss function to utilize trusted data. In the previous work of sample selection, [Yu *et al.*, 2019] proposed the strategy of "Update by Disagreement" while training two networks at the same time. JoCoR [Wei *et al.*, 2020] presented a joint loss and selected small-loss samples to update the parameters of two networks. Recently, some researchers employed web data in fine-grained recognition to improve the recognition accuracy, *e.g.*, using meta-dataset [Zhang *et al.*, 2020] and training two networks simultaneously [Sun *et al.*, 2021b].

2.3 Partial Label Learning

The aim of Partial Label Learning is to learn from each training example associated with a set of candidate labels, only one of which is valid for the training sample [Xu *et al.*, 2021]. There are two main streams in PLL methods including identifying the ground-truth label and estimating the candidate labels [Wang *et al.*, 2019]. Specifically, The main approach of identifying the ground-truth label is trying to disambiguate with the candidate labels via identification-based disambiguation [Lv *et al.*, 2020; Zhang and Yu, 2015] or averaging-based disambiguation [Cour *et al.*, 2011]. Furthermore, the identification-based PLL approaches regard the ground-truth label as latent variable and try to identify it and the average-based PLL methods treat all the candidate labels equally and average the modeling outputs as the prediction. While, compared with identifying the ground-truth, since the methods of estimating the candidate labels (*aka* confidence-based partial label learning) adopt the soft labeling information. These methods are lightly affected by false positive classes, and they can generally obtain superior recognition accuracy. In recent years, many works about partial label learning focused on coarse-grained datasets, and we discover the leveraging of PLL on webly-supervised fine-grained recognition.

3 Methodology

In this section, we introduce the overall framework, as well as elaborating the key strategies and its corresponding modules.

3.1 Overview

In general, our method is composed of two crucial strategies, *i.e.*, the open-set label noise removal strategy and the

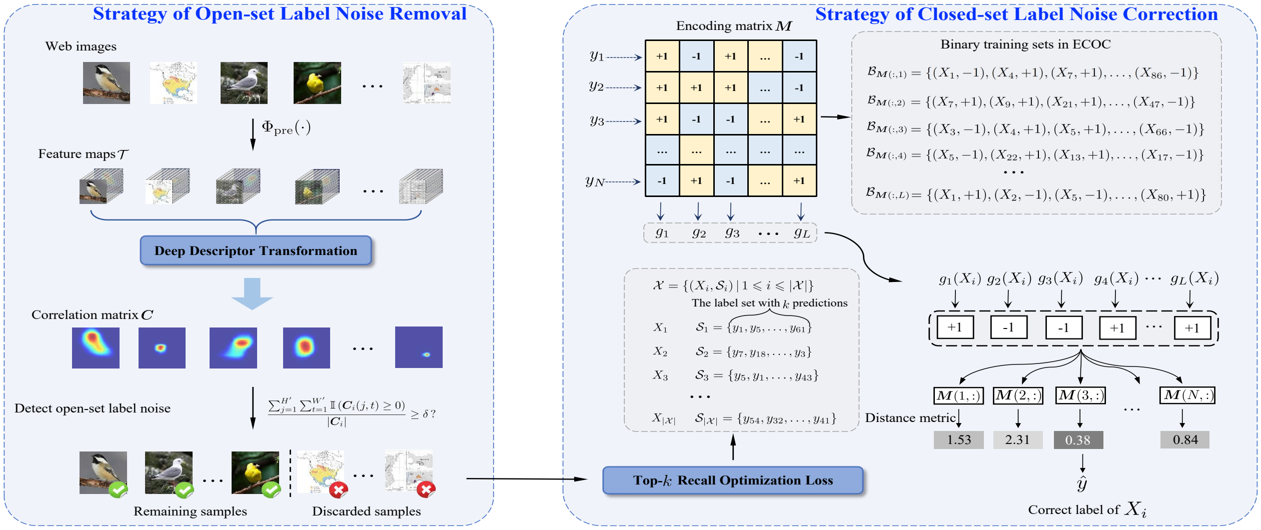
Strategy of Open-set Label Noise Removal
Web images
$\Phi_{\text{pre}}(\cdot)$
Feature maps $\mathcal{T}$
Deep Descriptor Transformation
Correlation matrix C
Detect open-set label noise
$\frac{\sum_{j=1}^{H'}\sum_{t=1}^{W'}\mathbb{I}(C_i(j,t)\geq 0)}{|C_i|}\geq\delta$?
Remaining samples
Discarded samples
Strategy of Closed-set Label Noise Correction
Encoding matrix M
y_1
y_2
y_3
y_N
+1 -1 +1 ... -1
+1 +1 +1 ... -1
+1 -1 -1 ... +1
... ...
-1 +1 -1 ... +1
Binary training sets in ECOC
$\mathcal{B}_{M(:,1)}=\{(X_1,-1),(X_4,+1),(X_7,+1),\dots,(X_{86},-1)\}$
$\mathcal{B}_{M(:,2)}=\{(X_7,+1),(X_9,+1),(X_{21},+1),\dots,(X_{47},-1)\}$
$\mathcal{B}_{M(:,3)}=\{(X_3,-1),(X_4,+1),(X_5,+1),\dots,(X_{66},-1)\}$
$\mathcal{B}_{M(:,4)}=\{(X_5,-1),(X_{22},+1),(X_{13},+1),\dots,(X_{17},-1)\}$
$\mathcal{B}_{M(:,L)}=\{(X_1,+1),(X_2,-1),(X_5,-1),\dots,(X_{80},+1)\}$
g_1 g_2 g_3 $\cdots$ g_L
$\mathcal{X}=\{(X_i,\mathcal{S}_i)\,|\,1\leqslant i\leqslant|\mathcal{X}|\}$
The label set with k predictions
X_1 $\mathcal{S}_1=\{y_1,y_5,\dots,y_{61}\}$
X_2 $\mathcal{S}_2=\{y_7,y_{18},\dots,y_3\}$
X_3 $\mathcal{S}_3=\{y_5,y_1,\dots,y_{43}\}$
$X_{|\mathcal{X}|}$ $\mathcal{S}_{|\mathcal{X}|}=\{y_{54},y_{32},\dots,y_{41}\}$
Top-k Recall Optimization Loss
$g_1(X_i)$ $g_2(X_i)$ $g_3(X_i)$ $g_4(X_i)$ $\cdots$ $g_L(X_i)$
+1 -1 -1 +1 ··· +1
$M(1,:)$ $M(2,:)$ $M(3,:)$ ··· $M(N,:)$
Distance metric
1.53
2.31
0.38
0.84
$\hat{y}$
Correct label of X_i

Figure 1: Pipeline of our proposed method, which consists of two strategies. The first strategy erases open-set noisy images and obtain an image space $\mathcal{X}$. The second strategy is composed of two components, *i.e.*, 1) performing the top-k recall optimization loss on remaining images to gain the label sets containing the ground-truth; 2) utilizing the distance between closed-set noisy images and each row of the encoding matrix $\boldsymbol{M}$ to obtain the prediction category for each closed-set noisy image. Finally, the samples in image space $\mathcal{X}$ will be put into the network for training.

closed-set label noise correction strategy. For the open-set label noise removal strategy, we utilize the common patterns obtained from noisy-free samples to detect noisy samples. If the sample is not related to the common pattern to some extent, it would be detected as an open-set label noise and removed. For the closed-set label noise correction strategy, the top-k recall optimization loss optimizes the model by pushing samples from ground-truth labels forward. With such a loss, the top-k label set will contain the ground-truth label whenever possible, and we can deal with the closed-set label noise from partial label learning perspective. When correcting labels, we convert the multi-class label correction problem into multiple binary classification problems. After that, the outputs of multiple binary classifiers are combined into a codeword and predicted. Therefore, closed-set noise samples get corrected labels by encoding and decoding mechanisms with error correction capabilities. The pipeline of our proposed method is shown in Figure 1.

3.2 Strategy of Open-set Label Noise Removal

Considering that samples from the same category should have similar patterns but samples mislabeled into the category do not [Bouveyron and Girard, 2009], we infer samples from open-set noisy labels have entirely dissimilar patterns to all categories. Additionally, fine-grained data contains small inter-category variations [Wei *et al.*, 2021b] which means there exist common patterns among fine-grained categories. Therefore, inspired by Deep Descriptor Transformation (DDT) [Wei *et al.*, 2017], our strategy of open-set label noise removal is to find common patterns from webly fine-grained data. When a sample is irrelevant with the common patterns to some extent, we detect it as an open-set label noise.

In concretely, assuming the label space to be $\mathcal{Y}$ and the image space to be $\mathcal{I}$, for each set $\mathcal{I}'=\{I_1,I_2,I_3,\dots,I_n\}\in\mathcal{I}$ which contains n images, we feed them to a pre-trained CNN model Φ_{pre} and obtain the corresponding feature map: $\boldsymbol{t}_i=\Phi_{\text{pre}}(I_i)\in\mathbb{R}^{H\times W\times d}$, where H, W, d present the height, width and depth of $\boldsymbol{t}_i$. After that, we put all the feature maps together to derive a feature set $\mathcal{T}\in\mathbb{R}^{n\times H\times W\times d}$. We get the common pattern in $\mathcal{T}$ by applying Principal Component Analysis (PCA) [Wold *et al.*, 1987] along the depth dimension. Therefore, we get the eigenvector $\boldsymbol{p}\in\mathbb{R}^d$ corresponding to the largest eigenvalue as common pattern detector after the PCA process. Then, each spatial location of the given feature maps are channel-wise weighted and summarized to get the indicator matrixes $\mathcal{H}$. To put it more precisely, the indicator matrix $\boldsymbol{H}_i$ in $\mathcal{H}$ corresponding with the i-th feature map $\boldsymbol{t}_i$ is formulated as: $\boldsymbol{H}_i=\boldsymbol{t}_i\cdot\boldsymbol{p}$. Furthermore, we require indicator matrixes with the same size as input samples to reflect the correlation between each pixel and the common pattern. Thus, we obtain the $\boldsymbol{C}_i\in\mathbb{R}^{H'\times W'}$ by upsampling the indicator matrix $\boldsymbol{H}_i\in\mathbb{R}^{H\times W}$ according to the input size. $\boldsymbol{C}_i$ contains positive and negative values which can reflect the positive and negative correlations of these deep descriptors. Because $\boldsymbol{p}$ is obtained in $\mathcal{I}'$, the positive correlation could indicate the common pattern through n images. Thus, we set a threshold δ about the correlation value to detect each image with:

$$\frac{\sum_{j=1}^{H'}\sum_{t=1}^{W'}\mathbb{I}\left(\boldsymbol{C}_i(j,t)\geq 0\right)}{|\boldsymbol{C}_i|}\geq\delta\,, \tag{1}$$

where $\mathbb{I}(\cdot)$ represents the indicator function. If a sample does not satisfy Equation (1), it will be regarded as an open-set label noise and removed from the noisy image space. Finally, we can get a sample space $\mathcal{X}$ with our proposed open-set label noise removal strategy.

Why obtained the common patterns are effective in open-set label noise removal? Actually, in the setting of WSL or learning from noisy labels [Sun *et al.*, 2021b], regardless of whether the proportion of unknown classes is high or low, the classes in them are *quite* cluttered. Meanwhile, according to the definition of noise in WSL (*i.e.*, the items or observations which raise suspicious by differing significantly from the majority of the data), a single class of the unknown classes cannot be the dominant class. That is to say that, the number of a certain class's (*e.g.*, "chart") images will not be more than the number of images of the target class; otherwise this "chart" class will become the target class. Therefore, the common patterns obtained must come from clean samples and using the common patterns to detect the relevance of the samples in noisy datasets will show good performance. Thus, the common patterns in our method can effectively work.

3.3 Strategy of Closed-set Label Noise Correction

Ideal for dealing with the closed-set label noise problem is to give each sample in $\mathcal{X}$ a more confident label. Therefore, we provide the following two components in this strategy.

Top-k Recall Optimization Loss

Mislabeled images in fine-grained categories have more similar deep descriptors, which makes it more difficult to give them correct labels. Thus, we leverage the partial label learning to solve the closed-set label noise problem associated with fine-grained image classification. For closed-set noisy samples which are difficult to predict the label, we tend to utilize more easily reachable information in noisy label learning, *i.e.*, ranking the ground-truth in a label set with k candidate labels. Thus, as long as one of the top-k is ground-truth, such samples can be used in the partial label learning. Therefore, we propose a loss which drives the model to reach a high recall rate. The loss is designed to push samples whose class is the same as the query's class forward as much as possible, which is implemented by penalizing those misplaced samples.

Formally, in the label space $\mathcal{Y} = \{y_1, y_2, \ldots, y_N\}$, define the sample space $\mathcal{X}$ as $\mathcal{X} = \{\mathcal{X}_{y_1}, \mathcal{X}_{y_2}, \ldots, \mathcal{X}_{y_N}\}$, where $\mathcal{X}_{y_i}$ represents a collection of instances belong to the i-th category y_i. During the training data selecting stage, we randomly select C categories to generate a mini-batch $\mathcal{A}$. For each selected category y_i, we get n^* samples in $\mathcal{X}_{y_i}$. We can obtain embedding features $\mathcal{F} = \{\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_a\}$ based on $\mathcal{A} = \{A_1, A_2, \ldots, A_a\} \in \mathcal{X}$ where $a = n^* \times C$. More specifically, we gain the embedding feature $\boldsymbol{f}_i$ of its input image A_i in $\mathcal{A}$ via a backbone CNN model Φ_{CNN} by $\boldsymbol{f}_i = \Phi_{\text{CNN}}(A_i) \in \mathbb{R}^c$, where c is the length of the embedding feature $\boldsymbol{f}_i$. The similarity matrix $\boldsymbol{s} \in \mathbb{R}^{a \times a}$ is generated by cosine similarity based on embedding features to measure the distance among $\mathcal{A}$. In details, the similarity of the i-th query image and the j-th support image is calculated as $s_{i,j} = \frac{\boldsymbol{f}_i^\top \boldsymbol{f}_j}{\|\boldsymbol{f}_i\|\|\boldsymbol{f}_j\|}$.

We can define the set $\mathcal{K}$ as the group of top-k images sorted by the similarity of each query image and other $\boldsymbol{s}_{q,:}$, where $\boldsymbol{s}_{q,:} \in \boldsymbol{s}$ and $s_{q,q} \notin \boldsymbol{s}_{q,:}$, *i.e.*, $\mathcal{K} = \left\{A_j \in \mathcal{A} : s_{q,j} \geq s_{q,[k]}, q \neq j\right\}$, $[k]$ denotes the k-th largest element. One possible situation is that images in the same label as the query image might not exist in set $\mathcal{K}$. A straightforward approach is to use these images to replace those existing images in $\mathcal{K}$. Images belong to the same label with the query image but not in set $\mathcal{K}$ are called positive images, while negative images are samples in $\mathcal{K}$ but have different label with the query image. To achieve this operation, these positive images excluded from $\mathcal{K}$ constitute $\mathcal{P}$, while these negative images in $\mathcal{K}$ constitute $\mathcal{N}$. We denote $\mathcal{P}$ as $\mathcal{P} = \{A_j \in \mathcal{A} \setminus \mathcal{K} : y_j = y_q\}$, where $\mathcal{A} \setminus \mathcal{K}$ means the relative complement set of $\mathcal{K}$ in $\mathcal{A}$ and y_q is a label of the query image. Denote $|\mathcal{P}|$ as the number of positive images ranked out $\mathcal{K}$. We set n^* for each class within a mini-batch less than k, then we can get the negative set $\mathcal{N} = \left\{A_j \in \mathcal{K} : y_j \neq y_q \; and \; s_{q,j} \geq \boldsymbol{s}^n_{q,|\mathcal{P}|}\right\}$, where $\boldsymbol{s}^n$ is a matrix containing only the similarity scores of negative images. Thus, the loss function is defined as $\mathcal{L}_{topk} = \sum_{q=1}^{a} \left(\sum_{A_i \in \mathcal{N}} s_{q,i} - \sum_{A_j \in \mathcal{P}} s_{q,j} \right)$.

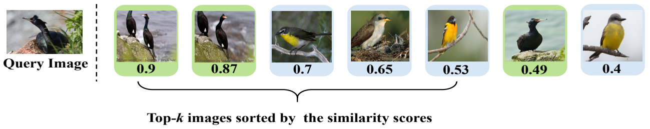

Figure 2: A bird example of loss calculation. Here $k = 5$, and the number of support images is 3. We reorder the support images of $\boldsymbol{s}$. The green samples are positive and the blue samples are negative. The $\mathcal{P}$ consists of all positive images out of top-k, $\sum_{A_j \in \mathcal{P}} s_{q,j} = 0.49$, and $|\mathcal{P}| = 1$, due to the number of support samples with same label is less than k, the last two negative images are not considered as misplaced. So the $\sum_{A_i \in \mathcal{N}} s_{q,i} = 0.7$, and $\mathcal{L}_{topk} = 0.7 - 0.49$.

A bird example is illustrated in Figure 2. Optimization of the model using the loss function can include the correct labels in the top-k classes with high confidence. Then, the top-k predicted labeles with high confidence are used as label sets $\mathcal{S}$ for corresponding samples.

Error-Correcting Output Codes

After obtaining as many label sets containing correct labels as possible, we need to determine a unique label for each sample. Error-Correcting Output Codes (ECOC) is a binary decomposition of multi-classification problem based on an encoding-decoding procedure. We utilize the improved ECOC to find correct labels, this method avoids being misled by the fine-grained false positive labels co-occurring with ground-truth label when correcting samples.

In the encoding stage, an encoding matrix $\boldsymbol{M} \in \{+1, -1\}^{N \times L}$is produced to support the learning process, where N represents the number of categories, and L represents the number of binary classifiers. More specifically, let $\boldsymbol{v} = [v_1, v_2, \ldots, v_N]^\top \in \{+1, -1\}^N$ denote the N-bits column coding which divide the label space into positive half $\mathcal{Y}_v^+ = \{y_j | v_j = +1, 1 \leqslant j \leqslant N\}$ and negative half $\mathcal{Y}_v^- = \{y_j | v_j = -1, 1 \leqslant j \leqslant N\}$.

Given a training sample $(X_m, \mathcal{S}_m)$, $X_m \in \mathcal{X}$, we regard $\mathcal{S}_m$ as an entirety to help build a binary classifier. Furthermore, the image X_m is used as a positive or negative sample only when the whole label set $\mathcal{S}_m$ fall into $\mathcal{Y}_v^+$ or $\mathcal{Y}_v^-$. We set some conditions on each column coding to ensure that each binary classifier includes enough positive and negative samples. Thus, binary training sets $\mathcal{B}$ can be generated for binary classifier training, where $\mathcal{B} = \{\mathcal{B}_{\boldsymbol{M}(:,1)}, \mathcal{B}_{\boldsymbol{M}(:,2)}, \ldots, \mathcal{B}_{\boldsymbol{M}(:,L)}\}$, $\mathcal{B}_{\boldsymbol{M}(:,l)} = \{(X_m, +1/-1) \,|\, 1 \leqslant m \leqslant |\mathcal{X}|\}$.

In the decoding stage, the crucial process is using binary classifiers to make predictions on closed-set label noise. Generally, given a test sample, the original method is to generate L-bits codeword by concatenating the outputs of the L binary classifiers. The Hamming distance is calculated between the codeword and each row of $\boldsymbol{M}$, the class label corresponding to the closest distance is returned as the prediction. For the performance of each classifier is unstable and in order to get more accurate results, we choose the loss-weighted decoding. At the same time, due to the similarity of fine-grained categories within a label set, the performance of the binary classifiers will also be affected by similar fine-grained categories, in which condition we add a fine-grained connected set restriction to affect the decoding phase.

In particular, for each category, we construct the connected set $\mathcal{E}_y$. Then, the j-th connected set $\mathcal{E}_{y_j}$ can be written as $\mathcal{E}_{y_j} = \{\mathcal{E}_{y_j} \cup \mathcal{S}_m : y_j \in \mathcal{S}_m, 1 \leqslant m \leqslant |\mathcal{X}|\}$. With the assist of $\mathcal{E}_y$, we set a performance matrix $\boldsymbol{G}^{N\times L}$ to represent the capability of classifiers. The performance of the t-th classifier g_t on the j-th category is calculated as follows:

$$\boldsymbol{G}(j,t) = \min_{z\in\mathcal{E}_{y_j}} \Big(\frac{1}{|\mathcal{Q}_z|} \sum_{(X_m,\mathcal{S}_m)\in\mathcal{Q}_z} \mathbb{I}\left(g_t(X_m) = \boldsymbol{M}(z,t)\right)\Big)\,, \tag{2}$$

where $\mathcal{Q}_z = \{(X_m, \mathcal{S}_m) | y_z \in \mathcal{S}_m, 1 \leqslant m \leqslant |\mathcal{X}|\}$, and $\mathbb{I}(\cdot)$ represents the indicator function. In order to get the relative performance of g over each category, we normalized the performance matrix $\boldsymbol{G}$ by row $\boldsymbol{G}^*(j,t) = \frac{\boldsymbol{G}(j,t)}{\sum_{r=1}^{L}\boldsymbol{G}(j,r)}$, where $j \in \mathbb{N}_+^N$ and $t \in \mathbb{N}_+^L$. Given a closed-set noisy image X_{cs}, the label prediction can be obtained via

$$\mathop{\arg\min}_{y_j(1\leqslant j\leqslant N)} \sum_{t=1}^{L} \boldsymbol{G}^*(j,t)\exp(-g_t(X_{cs})\boldsymbol{M}(j,t))\,. \tag{3}$$

Finally, we obtain the correct labels $\hat{y}$ of the closed-set noisy images by Equation (3) and send the rest of samples to the backbone network for re-training.

4 Experiments

In this section, we evaluate our proposed method on four real-world noisy datasets and compare our method with other state-of-the-art models. Meanwhile, we also conduct a series of ablation studies to estimate the importance of each component.

4.1 Datasets and Implementation Details

Datasets. *Web-Aircraft* [Sun *et al.*, 2020b] is a fine-grained aircraft dataset. It contains 13,503 training images and 3,333 test images, belonging to 100 different aircraft models. *Web-Bird* [Sun *et al.*, 2020b] is a fine-grained bird dataset. There are 18,388 noisy training instances and 5,794 clean test instances, belonging to 200 bird species. *Web-Car* [Sun *et al.*, 2020b] is a car dataset containing 196 different classes. The training set is composed of 21,448 samples and the test set consists of 8,041 samples. The *WebFG-496* is obtained directly by merging above three datasets. *WebiNat-5089* [Sun *et al.*, 2021b] contains 5,089 fine-grained categories and consists of 1,184,520 training images, which is the largest webly supervised fine-grained dataset. For *WebiNat-5089*, the validation set of *iNat2017* [Horn *et al.*, 2018] is utilized as the test set.

Implementation Details. We set the threshold $\delta = 0.2$. For the sampling strategy of top-k recall optimization loss, we set $k = 5$ and $n^* = 4$. For the number of binary classifiers in ECOC, we set $L = 128$. The rule for dividing label space is used only when the number of positive and negative samples is greater than 25. Furthermore, the total number of available samples is greater than 10% of that corresponding with input samples. We adopt the ResNet-50 [He *et al.*, 2016] as our backbone. We use the stochastic gradient descent optimizer with the momentum set as 0.9. The batch size is 32 for per GPU and the epoch number is set as 110. The initial learning rate is 5×10^{-3} while the weight decay is 2×10^{-5}. The warmup stage lasts for 10 epochs. We conduct all experiments on three GeForce RTX 3060 GPUs.

4.2 Comparison Results

In experiments, we adopt three real-world web-image-based noisy datasets to validate the effectiveness and superiority of our method and compare our proposed method with the following state-of-the-art models.

Table 1 presents the classification accuracy on three aforementioned benchmark fine-grained datasets. As shown in the table, our proposed method significantly and consistently outperforms the other baseline methods on these datasets. In particular, compared with the state-of-the-art method Co-LDL, our method achieves 2.89%, 1.70% and 2.36% improvements on *Web-Aircraft*, *Web-Bird* and *Web-Car*. Moreover, our proposed method also obtains superior result with 2.32% improvement on overall average accuracy. These observations validate the effectiveness of the proposed method, as well as its promising practicality in real applications of webly-supervised fine-grained recognition. For the comparison on *WebiNat-5089* contains more than 1.1 million training images, we utilize three benchmarks and a competitive method, *i.e.*, VGG-16 [Simonyan and Zisserman, 2014], GoogLeNet [Szegedy *et al.*, 2015], ResNet-50 [He *et al.*, 2016] and Peer-learning [Sun *et al.*, 2021b]. As evaluated in Table 2, our approach gains the highest accuracy among these methods.

4.3 Ablation Studies

In this section, we demonstrate the effectiveness of crucial components, *i.e.*, open-set label noise removal (cf. Section 3.2) and closed-set label noise correction (cf. Section 3.3). Since the top-k recall optimization loss (cf. Section 3.3) is the basis for constructing the label set to ECOC (cf. Section 3.3), so we regard these as an entirety for ablation studies. In ablation studies, we apply these components incrementally on a vanilla backbone (*i.e.*, ResNet-50) as the baseline. As evaluated in Table 3, by stacking these two components one by one, the recognition results are steadily im-

Methods	Type	Publications	Backbone	WebFG-496			
				Web-Aircraft	*Web-Bird*	*Web-Car*	Average
Decoupling [Eran and Shai, 2017]*	WS	NeurIPS 2017	ResNet50	75.91	71.61	79.41	75.64
Co-teaching [Han *et al.*, 2018]*	WS	NeurIPS 2018	ResNet50	79.54	76.68	84.95	80.39
Co-teaching+ [Yu *et al.*, 2019]*	WS	ICML 2019	ResNet50	74.80	70.12	76.77	73.90
PENCIL [Yi and Wu, 2019]*	WS	CVPR 2019	ResNet50	78.82	75.09	81.68	78.53
JoCoR [Wei *et al.*, 2020]*	WS	CVPR 2020	ResNet50	80.11	79.19	85.10	81.47
AFM [Peng *et al.*, 2020]*	WS	ECCV 2020	ResNet50	81.04	76.35	83.48	80.29
Self-adaptive [Huang *et al.*, 2020]*	WS	NeurIPS 2020	ResNet50	77.92	78.49	78.19	78.20
Jo-SRC [Yao *et al.*, 2021]	WS	CVPR 2021	ResNet50	82.28	80.41	86.59	83.10
Peer-learning [Sun *et al.*, 2021b]	WSFG	ICCV 2021	B-CNN (VGG-16)	74.38	76.48	78.52	76.46
Co-LDL [Sun *et al.*, 2021a]*	WSFG	TMM 2021	ResNet50	83.83	81.02	89.17	84.67
Our method	WSFG	This paper	ResNet50	**86.72**	**82.72**	**91.53**	**86.99**

- "WS" is the abbreviation of "Webly Supervised", and "WSFG" is the abbreviation of "Webly Supervised Fine-Grained".
- The results of methods marked with "$*$" are from [Sun *et al.*, 2021a].

Table 1: The comparison with state-of-the-art approaches in test accuracy $(\%)$ on real-world noisy datasets.

	Method	Backbone	*WebiNat-5089*
Benchmarks	VGG-16	–	44.77
	GoogLeNet	–	39.71
	ResNet-50	–	48.23
Webly	Peer-learning	ResNet-50	54.56
	Our method	ResNet-50	**57.59**

Table 2: The comparison of classification accuracy $(\%)$ of four models and our proposed method on the *WebiNat-5089* dataset.

Method	Accuracy $(\%)$
Baseline	82.04
+ Open-set Label Noise Removal	85.85
+ Closed-set Label Noise Correction (H)	86.41
+ Closed-set Label Noise Correction (L)	86.65
+ Closed-set Label Noise Correction (CL)	**86.72**

Table 3: Ablation results of components in our proposed method on *Web-Aircraft*. Note that, "H" presents the Hamming distance during the prediction phase at the decoding stage, "L" is the loss-weighted decoding and "CL" is the loss-weighted decoding with fine-grained connected set restriction.

proved, which justifies the effectiveness of our proposed components. Furthermore, in order to show the impact of fine-grained categories and the unstable performance of the binary classifier, we compared the results of Hamming distance decoding, loss-weighted decoding and loss-weighted decoding with fine-grained connected set restriction. Experiments show that the loss-weighted decoding with fine-grained connected set restriction not only alleviates the negative impact of unstable performance of binary classifiers, but also mitigates undesirable influence of fine-grained categories.

The number of binary classifiers is closely related to the error correction capability. The more binary classifiers, the better the error correction capability. Figure 3 shows the performance with different number of binary classifiers. It can be seen that the test accuracy almost saturates when L reaches 128. In terms of accuracy and effectiveness, we choose 128 as the number of binary classifiers.

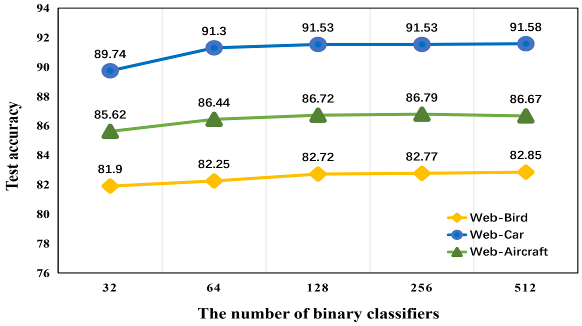

Figure 3: Test accuracy $(\%)$ with different numbers of binary classifiers on three real-world noisy datasets.

5 Conclusion

In this paper, we presented two strategies for dealing with label noise problem in webly-supervised fine-grained recognition tasks, *i.e.*, the open-set noise removal and the closed-set noise correction strategy. Particularly, the first strategy helps erase images that are completely out of the label space to obtain noisy-free images and closed-set noisy images. The second stage adopt PLL together with the proposed top-k recall optimization loss to locate and correct closed-set noisy images in fine-grained datasets. Experiments on four noisy datasets demonstrate that our approach achieve the state-of-the-art performance. Additionally, the rapid growth of multimedia data in the web brings us more useful information but also contains more noise that needs to be screened, which motivates us to study fine-grained cross-modal recognition based on our proposed method as the future work.

Acknowledgements

This work was supported by National Key R&D Program of China (2021YFA1001100), Natural Science Foundation of Jiangsu Province of China under Grant (BK20210340), the Fundamental Research Funds for the Central Universities (No. 30920041111), and Beijing Academy of Artificial Intelligence (BAAI). We gratefully acknowledge the support of MindSpore, CANN (Compute Architecture for Neural Networks) and Ascend AI Processor used for this research.

References

[Bouveyron and Girard, 2009] Charles Bouveyron and Stéphane Girard. Robust supervised classification with mixture models: Learning from data with uncertain labels. *PR*, 42(11):2649–2658, 2009.

[Cour *et al.*, 2011] Timothee Cour, Ben Sapp, and Ben Taskar. Learning from partial labels. *JMLR*, 12(42):1501–1536, 2011.

[Eran and Shai, 2017] Malach Eran and Shalev-Shwartz Shai. Decoupling "when to update" from "how to update". In *NeurIPS*, pages 960–970, 2017.

[Han *et al.*, 2018] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, pages 8527–8537, 2018.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Hendrycks *et al.*, 2018] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. In *NeurIPS*, pages 10456–10465, 2018.

[Horn *et al.*, 2018] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *CVPR*, pages 8769–8778, 2018.

[Huang *et al.*, 2020] Lang Huang, Chao Zhang, and Hongyang Zhang. Self-adaptive training: Beyond empirical risk minimization. In *NeurIPS*, pages 960–970, 2020.

[Lv *et al.*, 2020] Jiaqi Lv, Miao Xu, Lei Feng, Gang Niu, Xin Geng, and Masashi Sugiyama. Progressive identification of true labels for partial-label learning. In *ICML*, pages 6500–6510, 2020.

[Maji *et al.*, 2013] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[Niu *et al.*, 2018] Li Niu, Ashok Veeraraghavan, and Ashu Sabharwal. Webly supervised learning meets zero-shot learning: A hybrid approach for fine-grained classification. In *CVPR*, pages 7171–7180, 2018.

[Patrini *et al.*, 2017] 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 *CVPR*, pages 1944–1952, 2017.

[Peng *et al.*, 2020] Xiaojiang Peng, Kai Wang, Zhaoyang Zeng, Qing Li, Jianfei Yang, and Yu Qiao. Suppressing mislabeled data via grouping and self-attention. In *ECCV*, pages 786–802, 2020.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Song *et al.*, 2020] Kaitao Song, Xiu-Shen Wei, Xiangbo Shu, Ren-Jie Song, and Jianfeng Lu. Bi-modal progressive mask attention for fine-grained recognition. *IEEE TIP*, 29:7006–7018, 2020.

[Sun *et al.*, 2020a] Guolei Sun, Hisham Cholakkal, Salman Khan, Fahad Khan, and Ling Shao. Fine-grained recognition: Accounting for subtle differences between similar classes. In *AAAI*, pages 12047–12054, 2020.

[Sun *et al.*, 2020b] Zeren Sun, Xian-Sheng Hua, Yazhou Yao, Xiu-Shen Wei, Guosheng Hu, and Jian Zhang. CRSSC: Salvage reusable samples from noisy data for robust learning. In *ACM MM*, pages 92–101, 2020.

[Sun *et al.*, 2021a] Zeren Sun, Huafeng Liu, Qiong Wang, Tianfei Zhou, Qi Wu, and Zhenmin Tang. Co-LDL: A co-training-based label distribution learning method for tackling label noise. *IEEE TMM*, 2021. DOI: 10.1109/TMM.2021.3116430.

[Sun *et al.*, 2021b] Zeren Sun, Yazhou Yao, Xiu-Shen Wei, Yongshun Zhang, Fumin Shen, Jianxin Wu, Jian Zhang, and Heng Tao Shen. Webly supervised fine-grained recognition: Benchmark datasets and an approach. In *ICCV*, pages 10602–10611, 2021.

[Szegedy *et al.*, 2015] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.

[Wang *et al.*, 2019] Qian-Wei Wang, Yu-Feng Li, and Zhi-Hua Zhou. Partial label learning with unlabeled data. In *IJCAI*, pages 3755–3761, 2019.

[Wei *et al.*, 2017] Xiu-Shen Wei, Chen-Lin Zhang, Yao Li, Chen-Wei Xie, Jianxin Wu, Chunhua Shen, and Zhi-Hua Zhou. Deep descriptor transforming for image Co-localization. In *IJCAI*, pages 3048–3054, 2017.

[Wei *et al.*, 2020] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *CVPR*, pages 13726–13735, 2020.

[Wei *et al.*, 2021a] Xiu-Shen Wei, Yang Shen, Xuhao Sun, Han-Jia Ye, and Jian Yang. A^2-Net: Learning attribute-aware hash codes for large-scale fine-grained image retrieval. In *NeurIPS*, pages 5720–5730, 2021.

[Wei *et al.*, 2021b] Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. Fine-grained image analysis with deep learning: A survey. *IEEE TPAMI*, 2021. DOI: 10.1109/TPAMI.2021.3126648.

[Wold *et al.*, 1987] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37–52, 1987.

[Xu *et al.*, 2021] Ning Xu, Congyu Qiao, Xin Geng, and Min-Ling Zhang. Instance-dependent partial label learning. In *NeurIPS*, pages 27119–27130, 2021.

[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 *CVPR*, pages 5192–5201, 2021.

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

[Yu *et al.*, 2019] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *ICML*, pages 7164–7173, 2019.

[Zhang and Yu, 2015] Min-Ling Zhang and Fei Yu. Solving the partial label learning problem: An instance-based approach. In *IJCAI*, pages 4048–4054, 2015.

[Zhang *et al.*, 2020] Chuanyi Zhang, Yazhou Yao, Xiangbo Shu, Zechao Li, Zhenmin Tang, and Qi Wu. Data-driven meta-set based fine-grained visual recognition. In *ACM MM*, pages 2372–2381, 2020.

[Zhu *et al.*, 2020] Yaohui Zhu, Chenlong Liu, and Shuqiang Jiang. Multi-attention meta learning for few-shot fine-grained image recognition. In *IJCAI*, pages 1090–1096, 2020.