Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

SPECIAL ISSUE

Robust Fine-Grained Visual Recognition with Images Based on Internet of Things

Zhenhuang Cai[1] | **Shuai Yan**[1] | **Dan Huang**[2]

[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, Jiangsu, China

[2]School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China

Correspondence
Zhenhuang Cai,
Email: czh@njust.edu.cn

Present address
School of Computer Science and Engineering, Nanjing University of Science and Technology, Jiangsu, China.

Abstract

Labeling fine-grained objects manually is extremely challenging, as it is not only label-intensive but also requires professional knowledge. Accordingly, robust learning methods for fine-grained recognition with web images collected from Internet of Things have drawn significant attention. However, training deep fine-grained models directly using untrusted web images is confronted by two primary obstacles: 1) label noise in web images and 2) domain variance between the online sources and test datasets. To this end, in this study, we mainly focus on addressing these two pivotal problems associated with untrusted web images. To be specific, we introduce an end-to-end network that collaboratively addresses these concerns in the process of separating trusted data from untrusted web images. To validate the efficacy of our proposed model, untrusted web images are first collected by utilizing the text category labels found within fine-grained datasets. Subsequently, we employ the designed deep model to eliminate label noise and ameliorate domain mismatch. And the chosen trusted web data are utilized for model training. Comprehensive experiments and ablation studies validate that our method consistently surpasses other state-of-the-art approaches for fine-grained recognition task in a real-world scenario. Simultaneously, this introduces a novel pipeline for fine-grained recognition with substantial efficacy in practical applications. The source code and models can be accessed at: https://github.com/NUST-Machine-Intelligence-Laboratory/DDN.

KEYWORDS
deep neural network, label noise, domain mismatch, fine-grained recognition, internet of things.

1 | INTRODUCTION

Discerning nuanced distinctions within fine-grained categories (*e.g.*, various bird species[1], or flowers[2]) typically demands a substantial volume of accurately annotated images. However, annotating objects at the subcategory level[3] often necessitates specialized expertise, thereby significantly constraining the viability of supervised fine-grained algorithms in practical real-world implementations.

To mitigate the need for extensive manual annotation and facilitate the development of pragmatic fine-grained models, the utilization of web images for training is gaining significant popularity. Nevertheless, as shown in Fig. 1 (a), due to the impact of inaccurate automated or non-expert annotations, as well as label corruption, web images typically come with noisy labels. As pointed out by Reference 4 and Reference 5, deep neural networks exhibit powerful capacity, allowing them to memorize incorrectly labeled training data. Therefore, it would be problematic to train deep models by directly leveraging web images with label noise[6,7]. To this end, Cui *et al.*[8] employed a universal iterative structure to bootstrap datasets for fine-grained categorization. For each round, the model selects several images according to confidence scores and forwards them to human annotators for precise manual labeling. Apart from this, Krause *et al.*[9] proposed to actively gather extensive amounts of fine-grained data for training networks. Generally, these strategies founded on either semi-supervised or active learning principles[8,9,10,11,12] have the capacity to effectively address common challenges inherent in webly supervised learning scenarios. Nevertheless, the requirement for varying degrees of human interaction restricts the scalability potential of these techniques.

 © 2023 Copyright Holder Name

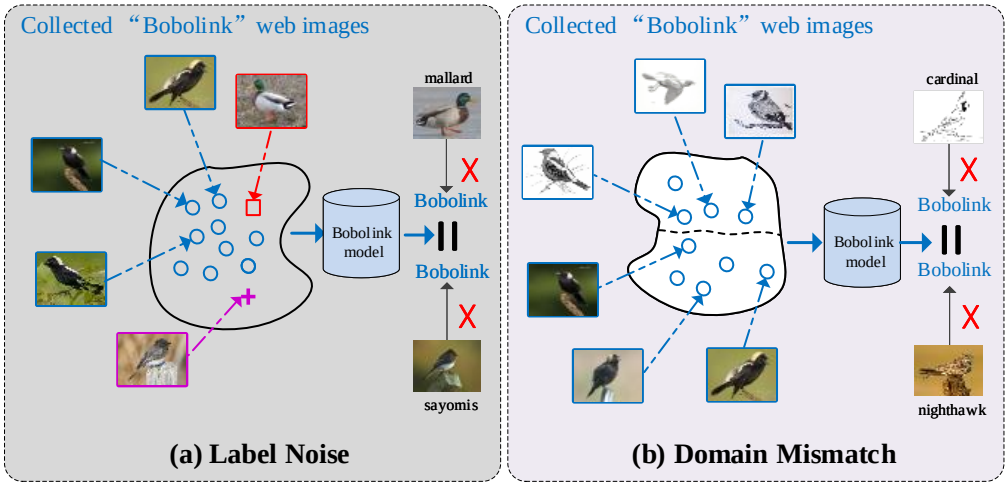

FIGURE 1 The issues of label noise and domain mismatch for web images. (a) The collected web images by text query *"Bobolink"* may include some incorrect instances, such as *"mallard"* image (red box) and *"sayomis"* image (purple box). Using these collected noisy web images for training is likely to result in identifying the test image *"mallard"* and *"sayomis"* as *"Bobolink"*. (b) Web images commonly come from various sources and contains multiple domains (*e.g.*, natural, sketch, and cartoon images), leading to domian mismatch between training data and test set. For instance, the test image *"cardinal"* may be classified as *"Bobolink"* despite being free of noise.

Recently, significant endeavors [13,14,15] have been directed towards automatically coping with label noise. Among these, several focus on calculating the noise transition probabilities across various category labels. For instance, Goldberger *et al.* [16] first devised an adaptation layer to simulate the noise transition matrix. Patrini *et al.* [17] presented loss correction to rectify noisy labels using the estimated noise transition matrix. However, accurate noise transition probability is too difficult to estimate due to the lack of prior knowledge, while inaccurate matrix estimation can even worsen the negative effects of noise. Another family of studies (*e.g.*, Bootstrapping [8], MentorNet [18], Active Learning [19], Decoupling [20] and Co-teaching [21]) focus on selecting correctly-labeled samples and removing noisy instances. These works usually operate on the assumption that the lower the noise rate of web data, the higher the performance of learned models. Despite promising results have been obtained in these works, as shown in Fig. 1 (b), none of them solve the domain mismatch problem which is also widely existed in the web data [22]. To minimize the domain discrepancy between web images used for training and test data, some methods [23,24] bolstered the diversity of the selected web images by leveraging query expansions or a Multiple Instance Learning (MIL) strategy.

Our approach is inspired by Reference 12, which utilizes deep MIL and noisy web images for fine-grained visual classification as well. However, we are different from Reference 12. Firstly, our method operates solely with web-supervised data, eliminating the need for manually annotated data. Reference 12 is a semi-supervised method and it still needs detailed annotations including part landmarks and bounding boxes for fine-grained classification. Secondly, while Reference 12 allocates fixed and equal weights to the instances within each bag, our approach diverges significantly. We propose an attention-based MIL pooling method, enabling the assignment of varying weights, particularly larger ones, to the key instances within the bags. This results in highly informative bag representations, facilitating the removal of noisy images and addressing domain mismatch effectively. Given the high similarity between subcategories in web-supervised fine-grained tasks, the identification and allocation of larger weights to key instances play a vital role in achieving superior performance. Thirdly, our proposed framework can mitigate label noise and alleviate domain discrepancy simultaneously in an end-to-end manner, while Reference 12 is not.

This decision that overcomes both problems simultaneously, rather than separately, stems from the intertwined nature of these challenges; tackling one while neglecting the other tends to fall short in significantly enhancing the performance of the model trained with web data. Our approach combines the bag-level MIL and the instance-level MIL into a single framework. And the later MIL employs attention mechanism to boost the recognition accuracy of the former one. Utilizing the combined output of two networks allows for more effective removal of noisy instances in bags. Furthermore, the way we construct bags in our deep MIL network contributes to bridging domain gap between web images and test set concurrently. Thorough experimentation and meticulous ablation studies showcase that our method surpasses the performance of cutting-edge approaches. Moreover,

we have made our source code publicly available, aiming to provide opportunities for researchers engaged in multimedia and affiliated domains to progress their research endeavors..

The primary contributions of this study can be summarized as follows: (1) We introduce a comprehensive end-to-end deep neural network model that tackles the challenges of mitigating label noise and alleviating domain discrepancy concurrently. By addressing these issues simultaneously, our model enhances the performance of web-supervised learning. Moreover, we have made our source code publicly available, aiming to foster further research and innovation among scholars in multimedia and related domains. (2) We perform extensive experiments regarding various baseline approaches to verify our proposed deep neural network model. Our experimental findings, observed on CUB200-2011 and Stanford Dogs datasets, provide empirical evidence of our method's superiority over existing state-of-the-art (SOTA). Additionally, through comprehensive ablation studies, we establish the strengths of our proposed model configurations. (3) Our work serves as a valuable pre-processing phase prior to direct web data learning. It aids in selecting appropriate instances, thereby enhancing the overall efficiency and effectiveness of web-based learning.

2 | RELATED WORK

2.1 | Fine-grained Visual Recognition

Fine-grained visual recognition focuses on distinguishing between subordinate categories, and previous works can be briefly categorized into three streams. The first cluster comprises strongly supervised learning approaches, hinging on manually annotated bounding boxes or detailed part labels for training [25,26,27,28,29,30,31,32,33,34,35]. The second category involves weakly supervised learning approaches, that solely rely on image-level labels in the training process, avoiding the need for detailed annotations [36,37,38,39,40,41,42,43,44]. The third class incorporates the use of web images for facilitating fine-grained recognition tasks [8,9,10,11,12,45]. Strictly speaking, these works [8,9,10,11,12,45] are not purely web-supervised learning, as all of them involve a certain level of human intervention. In contrast to these studies, our approach operates autonomously without the need for human intervention.

2.2 | Webly Supervised Learning

Learning directly from images found on the web, which circumvents the need for human intervention, is growing increasingly popular. However, the endeavor to train fine-grained recognition models using web-derived examples often encounters performance issues, primarily due to label noise [46]. Statistical learning has contributed significantly to solving the issue of noisy labels, particularly within the domain of theoretical analysis. This strategy can be classified into three distinct threads: surrogate loss, estimation of noise rates, and utilization of probabilistic models. Nonetheless, many of these theoretical approaches often come with specific priors or assumptions, yielding only moderate effectiveness when confronted with real-world complexities of label noise. Therefore, our primary emphasis lies in deep learning based approaches. To enhance the effectiveness of web-based learning, numerous deep learning approaches have been introduced to tackle the issue of label inaccuracies. To the best of our knowledge, prior research can be roughly divided into four sets. The initial category revolves around changing the loss function [47,48]. The second approach focuses on estimating the noise transition matrix [16,17]. The third approach put examples into buckets [49]. The fourth strategy tries to enhance data quality with a sample selection phase [8,12,20,18,21,50,51,52]. Nevertheless, these efforts do not specifically cater to fine-grained image recognition.

2.3 | Domain Adaptation

Our research is also pertinent to the field of domain adaptation. Chen *et al.* [53] proposed methods to tackle the issue of domain shift, specifically aiming to minimize the disparity between the source domain and the target domain. Since our methodology gathers training data from the internet, it shares resemblances with the approach described in Reference 54 that also capitalizes on supplementary data sources for its training regimen. Nevertheless, these techniques do not possess a specific focus on the context of fine-grained recognition under web-supervised conditions.

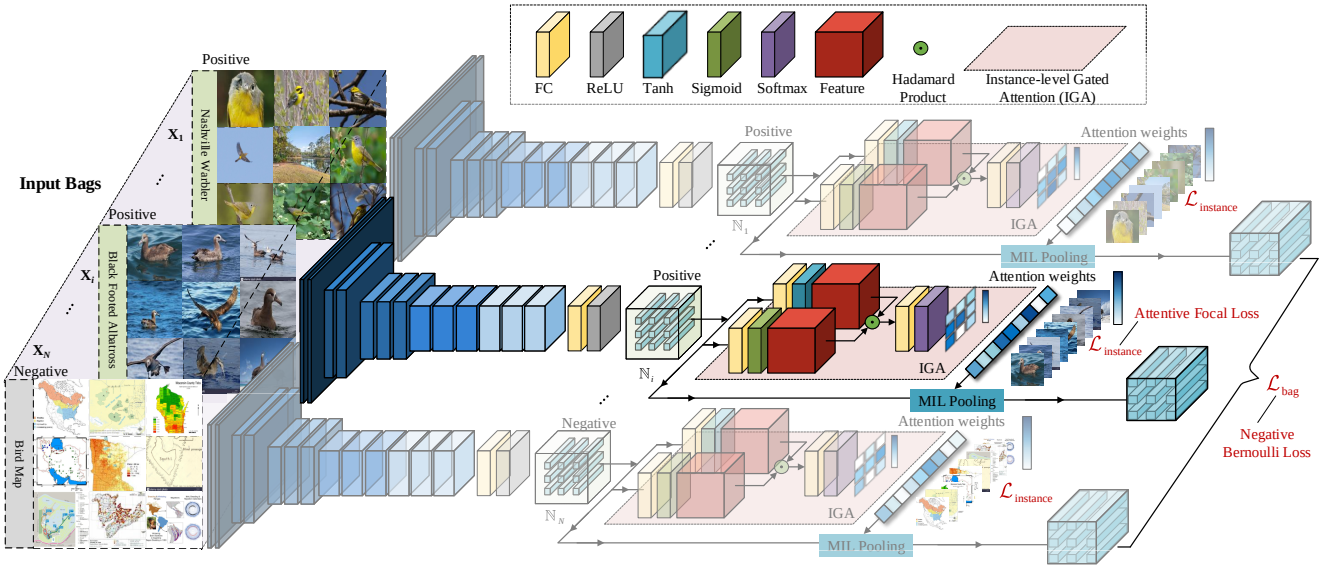

FIGURE 2 The framework of our deep neural network model. The input provided to our network consists of many "bags", each comprising multiple "instances" or "images". These bags first undergo initial processing in a backbone model (*i.e.*, VGG16) to extract features. Subsequently they enters a fully connected (FC) layer and a ReLU activation to generate a feature vector $\mathbb{N}$. Two branches are leverage to process the intermediate vectors. Our proposed attention block in the upper branch includes a FC layer, a *Tanh* layer and a sigmoid layer. Through attention block, the instance probability vector can be obtained. For the bottom branch, the proposed attention block generates weights to multiply $\mathbb{N}$ and estimate the positive probability of bags. Last, our deep neural network employs Attentive Focal Loss in the upper stream and the Negative Bernoulli Log Loss in the bottom stream. Our ultimate loss function results from combining the losses of both branches through a weighted summation.

3 | METHODOLOGY

Fig. 2 shows the framework of our proposed deep neural network model. Subsequent sections details the formulation of label noise, MIL with neural networks, novel attention mechanisms, and our approach's loss functions. Moreover, we also delve into domain mismatch, which is a key aspect of our proposed method.

3.1 | Formulation

Images collected from search engines such as Google and platforms like Flickr often exhibit loose and noisy labels. Consequently, directly using these images for training can lead to a sharp drop in classifier performance, especially with limited data[9]. Hence, prior to utilizing web images for model training, it's imperative to conduct noise removal to ensure more accurate learning outcomes.

For supervised learning algorithms, training data are typically represented as pairs $\{(x_i, y_i)\}$, in which $x_i \in \mathbb{R}^d$ represents the feature vector while $y_i \in \{0, 1\}$ denotes the corresponding label. Due to the nature of collecting images from the web using text queries, it is not possible for webly supervised learning to obtain ground truth labels of these examples. For multi-instance learning (MIL) paradigm, the training data is typically structured into collections referred to as bags $\{\mathbf{X}_i\}$, where each bag encompasses multiple samples $\{x_{i,j}\}$. It's important to note that the true label information, denoted as $\{\mathbf{Y}_i\}$, pertains solely to the entire bag, without specific labels assigned to the instances within the bag, marked as $\{y_{i,j}\}$. To be specific, the assumptions for multi-instance learning problem can be written as:

$$\mathbf{Y}_i = \begin{cases} 0, & \sum_j y_{i,j} = 0, \\ 1, & \text{otherwise.} \end{cases} \tag{1}$$

Further, two statements in Eq. (1) can be reformulated as a compact form via a maximum operator.

$$\mathbf{Y}_i = \max_j y_{i,j}. \tag{2}$$

Then we can formulate noisy samples removal as a MIL problem by treating web data as individual "instance", while regarding the search keyword as the label for the "bag".

3.2 | MIL with Neural Networks

In the conventional MIL scenario, the norm is to express instances using features. Nevertheless, the method[55] demonstrates the utility of employing fully-connected deep neural networks for MIL problem. For tasks related to computer vision, the integration of MIL and deep neural networks has been verified to enhance overall performance[56]. For this work, we follow the idea in Reference 56 as it has the potential to facilitate a versatile form of transformation that can be defined using neural networks' parameters. Specifically, we employ specific transformations to convert the k-th images into a compact representation, $\mathbf{h}_{i,j} = f_\phi(x_{i,j})$, where $\mathbf{h}_{i,j} \in \mathrm{H}$ and $\mathrm{H} = [0, 1]$. The concept of utilizing deep neural networks to parameterize all transformations holds considerable appeal, as it affords a high degree of flexibility to the entire approach. This permits seamless end-to-end training of the network through back-propagation. The sole requirement is the differentiability of the MIL pooling process.

3.3 | Attention-based MIL Pooling

For multi-instance learning problem, the operating of pooling should exhibit permutation invariance. In a prior work[57], two distinct pooling techniques for MIL were introduced: the maximal pooling operator denoted as $\max_{j=1,\ldots,J}\{\mathbf{h}_{i,j}\}$, and the averaged pooling operator represented as $\frac{1}{J}\sum_{j=1}^{J}\mathbf{h}_{i,j}$. However, both the max and mean pooling mechanisms are fixed in advance and not subject to training. Hence, an adjustable and versatile MIL pooling technique becomes imperative in this context.

Attention mechanism We employ a deep neural network to compute the weighted average of instances. Furthermore, it's crucial to fulfill a specific requirement: the weights must collectively add up to 1 in order to maintain invariance with respect to the bag's size. A critical requirement here is that the aggregate of these weights should equate to 1, thus ensuring invariance to the size of a bag. In particular, let $\mathrm{H} = \{\mathbf{h}_{i,1}, ..., \mathbf{h}_{i,j}\}$ denote a collection consisting of J instances. The MIL pooling technique put forward can be described as follows:

$$\mathbf{z} = \sum_{j=1}^{J} a_{i,j}\mathbf{h}_{i,j}, \tag{3}$$

$$a_{i,j} = \frac{\exp\{\mathbf{w}^\top \tanh(\mathbf{V}\mathbf{h}_{i,j}^\top)\}}{\sum_{j=1}^{J}\exp\{\mathbf{w}^\top \tanh(\mathbf{V}\mathbf{h}_{i,j}^\top)\}}. \tag{4}$$

$\mathbf{w} \in \mathbb{R}^{L\times 1}$ and $\mathbf{V} \in \mathbb{R}^{L\times I}$ represent parameters. $\tanh(\cdot)$ is adopted to comprise of both positive and negative values, ensuring appropriate flow of gradients. Our introduced approach enables the identification of similarities and differences among instances.

Gated attention mechanism $\tanh(\cdot)$ exhibits near-linearity within the interval $x \in [-1, 1]$, potentially leading to limited efficacy in capturing intricate relationships. In this work, we propose to also utilize gating mechanism as well as $\tanh(\cdot)$ non-linearity, generating:

$$a_{i,j} = \frac{\exp\{\mathbf{w}^\top(\tanh(\mathbf{V}\mathbf{h}_{i,j}^\top) \odot \mathrm{sigm}(\mathbf{U}\mathbf{h}_{i,j}^\top))\}}{\sum_{j=1}^{J}\exp\{\mathbf{w}^\top(\tanh(\mathbf{V}\mathbf{h}_{i,j}^\top) \odot \mathrm{sigm}(\mathbf{U}\mathbf{h}_{i,j}^\top))\}}. \tag{5}$$

$\odot$ denotes an element-wise multiplication and $\mathbf{U} \in \mathbb{R}^{L\times I}$ are parameters. $\mathrm{sigm}(\cdot)$ denotes the sigmoid non-linearity. Given the attention-driven MIL pooling strategy suggested, the bag-level model could benefit from a potentially richer bag representation, as it enables the assignment of distinct weights to individual instances within the bag. Ideally, large attention weights should be directed towards crucial instances associated with the positive label ($\mathbf{Y}_i = 1$). Then key instances in the bag can be more easily discovered. Therefore, the proposed attention-based MIL pooling narrows the disparity and establishes a connection between the instance-level method and the bag-level method.

3.4 | Loss Function

Training the model with the aim of optimizing the objective according to the maximum value across image labels in Eq. (2) can present two potential issues. Firstly, gradient-based approaches may encounter problems with vanishing gradients, which can hinder effective learning. Secondly, this formulation is applicable only when the instance-level classifier is employed. For addressing these challenges and simplify the learning process, we propose an alternative approach. We optimize the log-likelihood function to train the MIL network, in which the labels of bags follow the Bernoulli distribution:

$$\mathcal{L}_{\text{bag}} = -(1 - \mathbf{Y}_i)\log(1 - \mathbf{Y}'_i) - \mathbf{Y}_i \log(\mathbf{Y}'_i). \tag{6}$$

Here, $\mathbf{Y}'_i$ signifies the likelihood of a specific bag being positive, while $\mathbf{Y}_i$ represents the label assigned to the input bag. The benefit of employing this bag-level loss function lies in its ability to establish a unified representation of a bag without introducing extra bias to the bag model. Nevertheless, given the absence of instance labels, the instance-level model may be learned inadequately, contributing extra error to the ultimate prediction. Therefore, we suggest incorporating an additional loss function at the instance level. The naive negative Bernoulli loss is:

$$\mathcal{L}_{\text{instance}} = \frac{1}{J}\sum_{j=1}^{J}(-(1 - y_{i,j})\log(1 - a_{i,j}) - y_{i,j}\log(a_{i,j})), \tag{7}$$

where $a_{i,j}$ represents the attention mechanism weights as delineated in Eq. (4) and Eq. (5). From our experimental findings, it's evident that the classical negative Bernoulli instance-level loss's effectiveness falls short of expectations. By analyzing the the gathered web instances' distributions, we find a significant disparity between the counts of positive and negative samples within each bag. Specifically, there are a lot more positive instances than negative ones. The significant imbalance problem may result in substantial performance degradation of the model in the process of training. Therefore, we introduce an innovative weighted focal loss function to serve as the final instance-level loss:

$$\begin{aligned}\mathcal{L}_{\text{instance}} = \frac{1}{J}\sum_{j=1}^{J}(&-y_{i,j}\cdot\alpha(1 - a_{i,j})^{\gamma}\log(a_{i,j}) \\ &-(1 - y_{i,j})\cdot(1 - \alpha)a_{i,j}^{\gamma}\log(1 - a_{i,j})),\end{aligned} \tag{8}$$

in which $\alpha \in [0, 1]$ is a weighting parameter for alleviating the imbalance problem. To be specific, we assign α to class 1 and its complementary $1 - \alpha$ to class 0. The attention mechanism weights in Eq. (4) are $a_{i,j}$ and the tunable focusing parameter is $\gamma \geq 0$. Subsequently, our ultimate loss function can be derived as follows:

$$\mathcal{L} = \lambda_1\mathcal{L}_{\text{bag}} + \lambda_2\mathcal{L}_{\text{instance}}, \tag{9}$$

in which λ_1 and λ_2 denote two parameters for regulating the impact of the bag-level loss and the instance-level loss, respectively.

3.5 | How to Formulate Noise Removal as a MIL Problem?

MIL represents a cluster of learning techniques applicable to scenarios with significant levels of labeling noise. Instances, referred to as elements within the MIL framework, are organized into collections known as bags, with a singular label associated with each bag. The bag encompassing all negative examples is deemed a negative bag ($Y_i = \sum_j y_{i,j} = 0$, $y_{i,j}$ denotes the label of j-th instance within the bag i and Y_i stands for the label of bag i), otherwise it is positive.

MIL has two levels including bag-level and instance-level. The primary objective of bag-level MIL involves training a model to forecast a label for a bag (*e.g.*, Y_i). In contrast, instance-level MIL aims to instruct a model in forecasting labels for individual instances (*e.g.*, $y_{i,j}$). In our work, we first predict the bag's label. Upon obtaining a negative label, we subsequently categorize all instances within the bag as noisy images. When the bag's label is positive, our focus shifts to predicting the instances' labels among the bag and treat those samples which have negative labels as noisy images.

TABLE 1 The ACA (%) performance on CUB200-2011. "Box/Part" stands for using bounding box or part annotation during training. The term "Data" indicates whether the training dataset has undergone manual annotation (referred to as "anno.") or if it has been sourced from the internet (referred to as "web").

	Method	CUB200-2011		
		Box/Part	Data	ACA(%)
†	Part-CNN [26]	✓	anno.	76.37
	Normalized CNNs [58]	✓	anno.	75.70
	Deep LAC [28]	✓	anno.	80.30
	Part-Stacked CNN [27]	✓	anno.	76.60
	Mask-CNN [25]	✓	anno.	85.70
‡	Two-level attention [38]		anno.	69.70
	Simon *et al.* [39]		anno.	81.01
	Zhang *et al.* [40]		anno.	80.26
	Multi-attention [42]		anno.	86.50
	Vision + Language [44]		anno.	85.55
	Bilinear [37]		anno.	84.10
	RA-CNN [36]		anno.	85.30
	Filter-bank [41]		anno.	86.70
	TASN [59]		anno.	89.10
	DCL [60]		anno.	87.80
	Com-Parts Model [61]		anno.	90.40
§	Xu *et al.* [12]	✓	anno.+web	84.6
	Cui *et al.* [8]	✓	anno.+web	80.7
	Niu *et al.* [11]		anno.+web	76.47
	Cui *et al.* [43]		anno.+iNat	89.29
$	Bergamo *et al.* [24]		web	70.13
	NEIL [62]		web	69.08
	WSDG [63]		web	70.61
	Sukhbaatar *et al.* [64]		web	70.47
	Xiao *et al.* [65]		web	70.92
	Decoupling [20]		web	70.56
	Co-teaching [21]		web	73.85
	Update-Drop [66]		web	77.22
	Ours		web	**79.92**

† : strongly ‡ : weakly § : semi $: webly

3.6 | Why takes advantage of two MIL networks?

Classifiers trained directly from web images often face a significant degradation in performance as a result of noisy labels and domain discrepancies. To solve these issues, numerous proposed solutions have emerged individually. However, addressing either issue in isolation typically proved ineffective in substantially enhancing the efficacy of learning from the web. Hence, our approach focuses on addressing both challenges simultaneously. The strength of our approach lies in the fusion of the bag-level MIL model and the instance-level MIL model that operates through attention mechanisms, enabling us to address the dual challenges of label noise and domain mismatch concurrently. Specifically, we solve the label noise by predicting the labels of the bag and instance. We alleviate the domain mismatch by assigning varied weights to specific instances within a bag. This is the primary reason why our approach outperforms other webly supervised methods.

3.7 | Datasets and Evaluation Metric

We assess our approach on two widely employed benchmark datasets known as CUB200-2011 [1] and Stanford Dogs [67]. Specifically, CUB200-2011 is a challenging fine-grained dataset, meticulously labeled with 200 distinct avian categories. The dataset was designed to facilitate research on subordinate classifications, a task unattainable using other well-known datasets that emphasize primary level categories. These images were sourced from the online platform Flickr and subsequently refined through evaluation by contributors on Amazon's Mechanical Turk platform. The Stanford Dogs dataset comprises photographs of 120 distinct canine breeds originating from various global regions. This dataset was constructed with pictures and annotations sourced from ImageNet, designed for fine-grained visual classification. The assessment metric employed is the Average

TABLE 2 The ACA (%) performance on Stanford Dogs. "Box/Part" stands for using bounding box or part annotation during training. The term "Data" indicates whether the training dataset has undergone manual annotation (referred to as "anno.") or if it has been sourced from the internet (referred to as "web").

	Method	Stanford Dogs		
		Box/Part	Data	ACA(%)
†	Yang *et al.*[29]	✓	anno.	38.01
	Chai *et al.*[30]	✓	anno.	45.60
	Gavves *et al.*[32]	✓	anno.	57.00
	Kanan[31]	✓	anno.	47.70
	HAR-CNN[34]	✓	anno.	49.40
	Chen *et al.*[33]	✓	anno.	52.00
	FOAF[35]	✓	anno.	53.50
‡	RA-CNN[36]		anno.	87.30
	FCAN[68]		anno.	84.20
	Simon *et al.*[39]		anno.	68.61
	PDFS[40]		anno.	71.96
	Cui *et al.*[43]		anno.	84.19
	DVAN[69]		anno.	81.50
§	Niu *et al.*[11]		anno.+web	85.16
$	Bergamo *et al.*[24]		web	78.64
	NEIL[62]		web	80.16
	WSDG[63]		web	80.20
	Sukhbaatar *et al.*[64]		web	81.15
	Xiao *et al.*[65]		web	81.67
	Ours		web	**85.47**

† : strongly ‡ : weakly § : semi $: webly

TABLE 3 Influence of different domains.

Test data	Training data source	ACA (%)
CUB200	Flickr	81.4
	Google Image Search Engine	79.9
Stanford Dogs	Flickr	86.5
	Google Image Search Engine	85.4

TABLE 4 Influence of different attention mechanisms.

Test data	Attention mechanisms	ACA (%)
CUB200	Attention	79.9
	Gated attention	79.2
Stanford Dogs	Attention	85.4
	Gated attention	86.7

Classification Accuracy (**ACA**), a broadly utilized measure for evaluating the efficacy of fine-grained image recognition techniques.

3.8 | Experimental Setting

Web Data Collection: To learn our proposed deep neural network model for mitigating issues related to inaccurate labels and domain discrepancies within web images, we employing imprecise search terms, such as "bird maps", to gather noisy web instances as the negative training data. To be specific, we obtain 556 noisy images for the CUB200-2011 dataset and 483 for the Stanford Dogs dataset. The retrieved web images are subsequently utilized as the positive samples for training.

Overlap Removing Strategy: For web images crawled from Google or other search engines, there are many duplications intra the sub-classes due to the same search keywords. Apart from this, for fine-grained category recognition task, the repetitions also exists inter different sub-classes on account of the similarities of images belong to the same species. To combat the redundancies, we choose the image fingerprinting algorithm to address this problem. This method is often referred to as "image hashing". In practice, image hashing involves analyzing image content and generating a distinct value that serves to uniquely represent an image according to its visual attributes. Hashes for akin images should exhibit resemblance. The utilization of image hashing algorithms considerably simplifies the task of conducting near-duplicate image recognition. Finally, the "different hash" emerges within our scope of consideration, because it analyzes variances between neighboring pixel values to produce

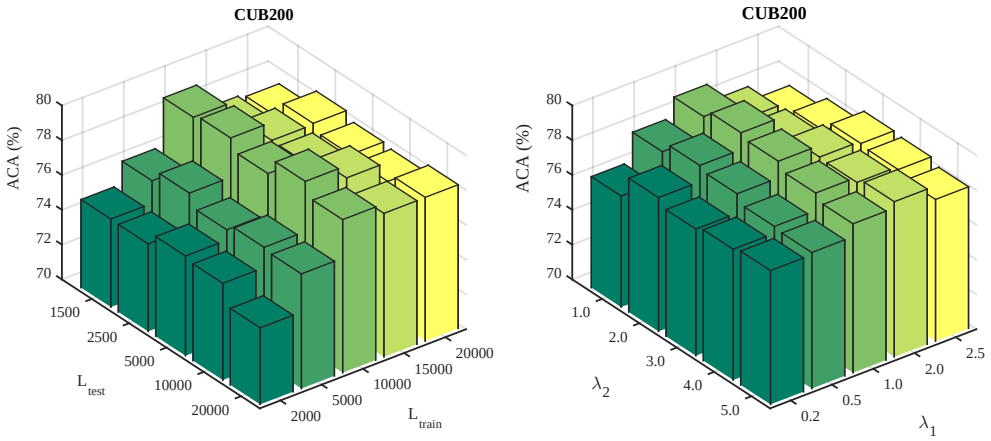

FIGURE 3 The ACA (%) performance of different parameter groups (L_{train}, L_{test}) and (λ_1, λ_2).

its unique identifier (*i.e.*, image fingerprint). Now that we have the hash values of all images in web dataset, the process then contains 2 steps. First is to detect replicate in same sub-class, in this case, we retain only one image from each near-duplicate group so as to reduce redundancy. As for different sub-class, the image appearing in two or more categories means one sample owns many labels at the same time, which will confuse the training model. So the second step is to remove all those images considered to be noise. When evaluate the similarity of hash values, we adopt Hamming distance to compute the quantity of differing bits within a hash. Two images with Hamming distance less than 2-bit are identified as duplicate, of course, the threshold can be adjusted.

Bags Generation: When training data is prepared, we subsequently generate L_{train} and L_{test} bags with a variance of θ and a mean of η. Instances are clustered within each bag, accompanied by their respective instance labels and bag label. During the experimentation, the values of L_{train} and L_{test} are chosen from {2000, 5000, 10000, 15000, 20000} and {1500, 2500, 5000, 10000, 20000}, respectively. Additionally, we search the parameter η from {15, 16, 18, 20, 25} and θ from {1, 2, 3, 4, 5}. We finally set $L_{\text{train}} = L_{\text{test}} = 10000$, $\eta = 16$, and $\theta = 2$ as the default value in our experiments.

Deep Model Learning: Following, we send the generated bags into our model for fine-grained visual recognition. The backbone of our proposed model relies upon a a pre-trained VGG-16. We eliminate the layers subsequent to the final pooling layer, yielding a feature map with a resolution of 7×7. Subsequently, a fully connected (FC) layer comprising 4096 neurons and a ReLU layer are incorporated. Additionally, we apply a sigmoid function to the probability vector of instance and normalize its value to [0, 1) for calculating the Attentive Focal Loss. Through experiments, we ultimately select $\alpha = 0.25$, $\gamma = 2$, $\lambda_1 = 1$, $\lambda_2 = 2$ as the default value. Furthermore, we conduct training for the deep model over 30 epochs. The starting learning rate is established at 10^{-5} and the weight decay parameter is set to 10^{-5}.

Fine-Grained Model Learning: Once we obtain the selected images using our model, we proceed to utilize these images as the training data to train a Bilinear CNN (BCNN)[37]. For the BCNN, we use the default parameter values as specified in Reference 37.

3.9 | Baselines

To demonstrate the effectiveness of our methodology, we incorporate four distinct groups of fine-grained SOTA approaches in baseline comparisons.

Strongly Supervised Methods: In this group, bounding boxes of objects or annotations indicating specific parts are required in the process of training. The baseline methods include Chen *et al.*[33], Mask-CNN[25], Part-Stacked CNN[27], Deep LAC[28], Pose Normalized CNNs[58], HAR-CNN[34], Chai *et al.*[30], Part-CNN[26], Yang *et al.*[29], Kanan[31], Gavves *et al.*[32], and FOAF[35].

Weakly Supervised Methods: This set needs image-level labels. Specifically, the baselines consist of RA-CNN[36], FCAN[68], Bilinear[37], DCL[60], Two-level attention[38], TASN[59], Simon *et al.*[39], Filter-bank[41], Multi-attention[42], Vision + Language[44], Zhang *et al.*[40], and Complementary Parts Model[61].

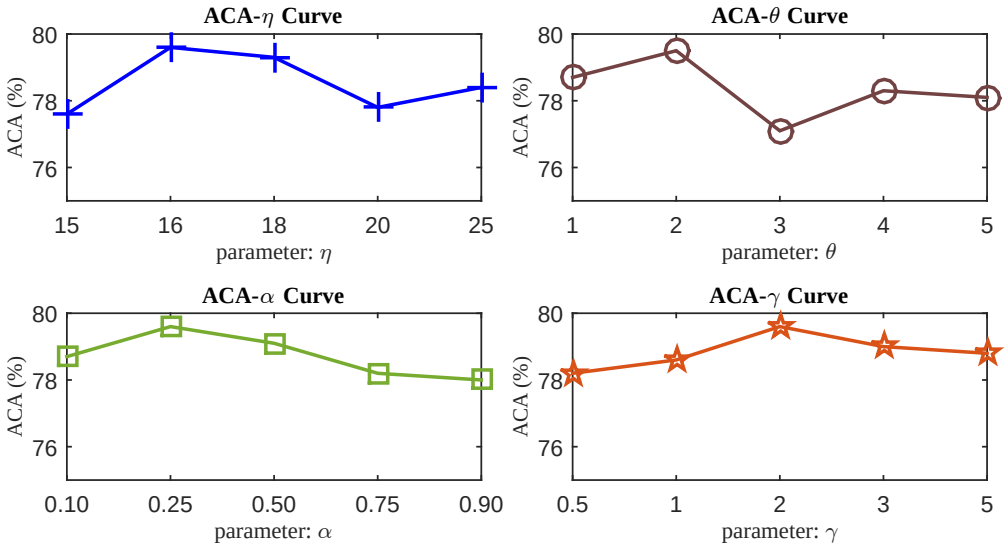

FIGURE 4 The influence of different parameters η, θ, α, and γ concerning ACA (%) results.

TABLE 5 Influence of denoising and augmenting.

Test data	Training data	ACA (%)
CUB200	Initial web instances	72.9
	cleansed web instances	79.9
	cleansed web instances + CUB200	85.1
Stanford Dogs	Initial web instances	76.3
	cleansed web instances	85.4
	cleansed web instances + Stanford Dogs	89.3

Semi-supervised Methods: This group requires a degree of human annotation involvement. The baselines contain Niu *et al.*[11], Cui *et al.*[43], Xu *et al.*[12], and Cui *et al.*[8].

Webly Supervised Methods: This set needs no human annotation. The method in[11] reproduced nearly all of the prominent techniques for fine-grained classification using webly supervised learning: NEIL[62], WSDG[63], Sukhbaatar *et al.*[64], Bergamo *et al.*[24], Xiao *et al.*[65]. In addition, we also directly utilize the performances of Decoupling[20] and Co-teaching[21] on CUB200-2011 dataset from Reference 66.

3.10 | Evaluations on Fine-grained Classification

Table 1 presents the outcomes of fine-grained approaches on the CUB200-2011 dataset, while Table 2 presents the outcomes for the Stanford Dogs dataset. A noticeable observation from these tables is that our approach has successfully surpassed other webly supervised methods. This improvement can be credited to the effectiveness of our method in addressing both the disparity between noisy web-derived data and labeled test data in terms of domains, and the diminishment of noise within the web data. It is worth mentioning that each webly supervised method is based on 100 web images per category, while strongly and weakly supervised methods have around 30 labeled images. Despite these efforts, webly supervised approaches still exhibit relatively lower performance compared to strongly, weakly, and semi-supervised approaches. The primary reason for this discrepancy may lie in the inherent label noise and domain inconsistencies found in web images. Despite employing different strategies within web-supervised approaches to mitigate the impact of these challenges, their performance remains unsatisfactory. However, this observation has inspired our conviction that substantial opportunities for enhancing web-supervised fine-grained techniques still remain untapped.

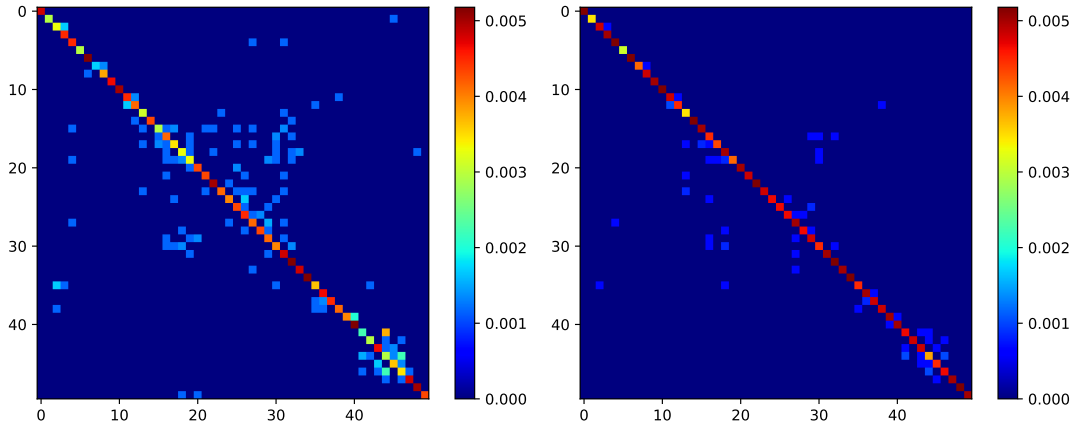

FIGURE 5 The misclassification confusion matrix (50 categories) sampled on the CUB200 dataset. The training data is (a) original web images and (b) purified web images, respectively.

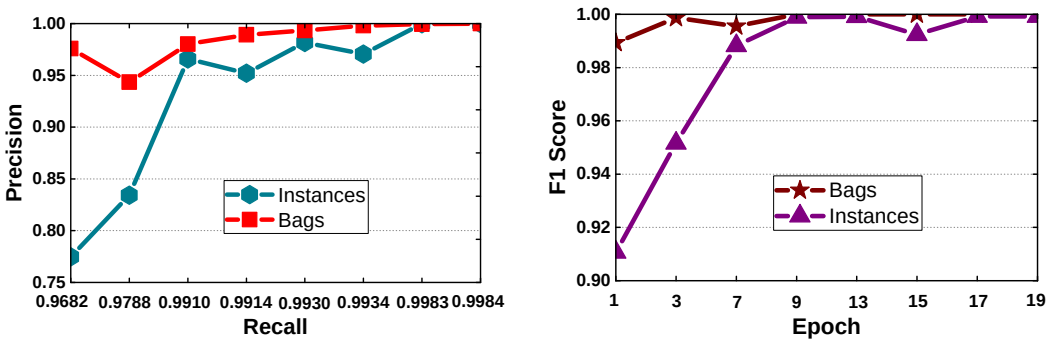

FIGURE 6 The Precision-Recall curve and F1 Score curve of our proposed deep neural network model.

4 | ABLATION STUDIES

We conduct comprehensive ablation studies to systematically examine our denoising model under web-supervised conditions. Subsequently, unless explicitly specified otherwise, we utilize CUB200-2011 as a representative case.

4.1 | Different Domains

To assess the impact of utilizing web data from various domains, we gather web instances for search keywords from two sources: Google and Flickr. Table 3 shows that the effectiveness of web pictures sourced from Flickr is slightly superior to those from Google. A potential rationale for this observation could be attributed to the origin of CUB200-2011 and Stanford Dogs datasets, both of which stem from Flickr. This convergence leads to a diminished domain disparity between the web-based images sourced from Flickr and the evaluation dataset.

4.2 | Different Attention Mechanisms

To analyze the effectiveness of different attention mechanisms stated in Section 3.3, we perform a comparative experiment about $a_{i,j}$ in Eq. (4) and Eq. (5). Table 4 provides a comparison of the efficacy of distinct attention mechanisms across two

datasets. The data in Table 4 indicates that, in the case of the CUB200 dataset, the attention mechanism outperforms gated attention mechanism; however, this scenario is reversed when considering the Stanford Dogs dataset.

4.3 | Hyper-parameters

We conduct a parameters analysis focusing on the parameters L_{train}, L_{test}, η, θ in bags generating and α, γ, λ_1, λ_2 in our deep neural network learning. In particular, our analysis delves into the interplay between parameter pairs, namely L_{train} and L_{test}, as they influence the process of bag generation. Additionally, we scrutinize the roles played by λ_1 and λ_2 in the context of Eq. (9). Furthermore, we visually examine the sensitivities of the remaining parameters. Fig. 3 illustrates the stable and consistent changing tendency of ACA concerning (L_{train}, L_{test}) and (λ_1, λ_2). Moreover, Fig. 4 showcases the sensitivities of the parameters η, θ, α, and γ concerning ACA.

4.4 | Denoising and Augmenting

Our work serves as a preliminary measure prior to direct web-based learning. To substantiate this assertion, we gather the top 100 web instances using the Google Image Search Engine to compose each distinct category. Subsequently, our model is employed to cleanse the label noise and alleviate domain disparity. The cleansed internet pictures are then utilized as the training dataset for the application of BCNN algorithm[37]. Furthermore, BCNN trained with the initial noisy web images is used to establish the baseline for comparison. Fig. 5 presents the misclassification confusion matrix by using the original web images (a) and purified web images (b). By comparing (a) and (b), we notice that directly leveraging web images for training tend to result in a relatively high misclassification probability. By performing our proposed deep neural network model and set the purified images as the training set for learning usually obtain a relatively small misclassification probability. In other words, the proposed deep neural network model has a realistic necessity before learning from the web.

Additionally, we combine the selected web data and the training set in the CUB200-2011 dataset to form the training data for learning the Bilinear model. Subsequently, we evaluate the model's performance on the CUB200-2011 test data. The experimental results are presented in Table 5. By observing Table 5, we can notice that the selected web images greatly improve the baseline performance, which demonstrates that the selected web data can enhance existing manually labeled datasets, leading to the development of a more robust model.

4.5 | Analysis for Bags and Instances

To assess the effectiveness of our deep denoising network model, we plot the Precision-Recall (PR) and F1 Score curves of for both bags and instances in Fig. 6. It can be observed that the area under PR curve for both bags and instances are relatively large. For the F1 Score curve, we notice that both bags and instances reach nearly 1 after 9 epochs.

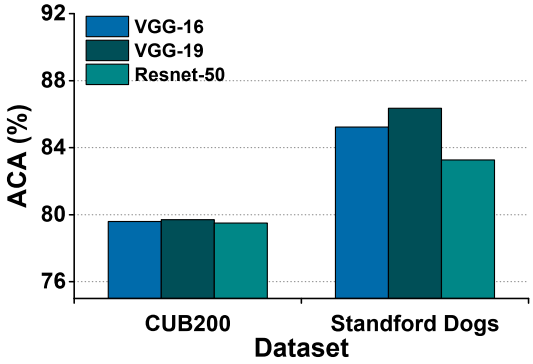

FIGURE 7 The influence of different deep neural network architectures.

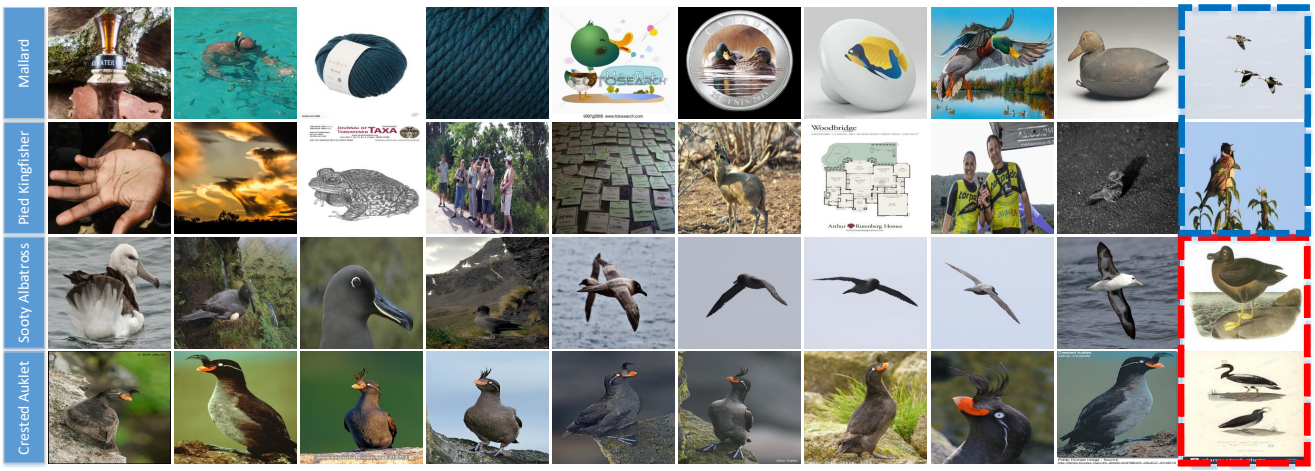

FIGURE 8 Accurate removal of noise and the unintentional removal of positive samples by our approach (**top**). Illustration of the successful retention of positive samples and erroneous preservation of noise by our method (**bottom**).

4.6 | Different Backbones

The impact of using various backbones on visual classification performance is widely recognized. To analyze this aspect, we replace the VGG-16 backbone in our denoising model with VGG-19 and ResNet-50[70]. The outcomes are presented in Fig. 7, revealing that the performance of the three backbone networks on CUB200-2011 is notably similar. However, when it comes to the Stanford Dogs dataset, the VGG-19 backbone network outperforms the other two choices, exhibiting the most impressive performance.

4.7 | Success and Failure Cases

We also analyze the success and failure cases of our deep denoising model. Specifically, we concentrate on two primary aspects: (1) the precise identification of noisy web images and the inadvertent exclusion of valuable images through our advanced deep learning framework, and (2) the accurate retention of positive images and the unintended omission of noisy instances by the proposed deep neural network model. The findings from the experiment are displayed in Fig. 8. Analyzing the upper part of Figure 8, it becomes evident that our advanced deep neural network architecture successfully eliminate incorrect labels. Meanwhile, our method erroneously removes a few positive instances (marked in blue boxes). This can be attributed to the distant positioning of the objects in the images, which makes it challenging to accurately capture the relevant features. Furthermore, when observing Fig. 8 (bottom), we notice that the selected web images exhibit relatively high accuracy, but a few noisy samples remain (*i.e.*, red boxes). Upon observing the remaining noise, it becomes apparent that there exists a comparable visual structure with the positive instances, contributing to the difficulty in distinguishing them accurately.

5 | CONCLUSIONS

In this study, we explored the issue of webly supervised fine-grained visual recognition tasks. Our main idea is to remove noisy labels and bridge domain gap between training set collected from web and test dataset, simultaneously. To be specific, we proposed an end-to-end deep neural network model to achieve this goal. Our method can also serves as a preprocessing step. Experimental results demonstrate that our method has attained the highest level of performance within the realm of fine-grained visual recognition under web-based supervision.

REFERENCES

1. Wah C, Branson S, Welinder P, Perona P, Belongie S. The caltech-ucsd birds-200-2011 dataset. In: California Institute of Technology. 2011.
2. Nilsback ME, Zisserman A. A visual vocabulary for flower classification. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2006:1447–1454.

3. Zhu F, Gao J, Yang J, Ye N. Neighborhood linear discriminant analysis. *Pattern Recognition.* 2022;123:108422.
4. Arpit D, Jastrzębski S, Ballas N, et al. A closer look at memorization in deep networks. In: International Conference on Machine Learning. 2017:233–242.
5. Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. Understanding deep learning requires rethinking generalization. In: International Conference on Learning Representations. 2017:1–15.
6. Chen X, Gupta A. Webly supervised learning of convolutional networks. In: IEEE International Conference on Computer Vision. 2015:1431–1439.
7. Zhu F, Yang J, Gao C, Xu S, Ye N, Yin T. A weighted one-class support vector machine. *Neurocomputing.* 2016;189:1–10.
8. Cui Y, Zhou F, Lin Y, Belongie S. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. In: IEEE Conference on Computer Vision and Pattern Recognition. 2016:1153–1162.
9. Krause J, Sapp B, Howard A, et al. The unreasonable effectiveness of noisy data for fine-grained recognition. In: European Conference on Computer Vision. 2016:301–320.
10. Xu Z, Huang S, Zhang Y, Tao D. Augmenting strong supervision using web data for fine-grained categorization. In: IEEE International Conference on Computer Vision. 2015:2524–2532.
11. Niu L, Veeraraghavan A, Sabharwal A. Webly supervised learning meets zero-shot learning: A hybrid approach for fine-grained classification. In: IEEE Conference on Computer Vision and Pattern Recognition. 2018:7171–7180.
12. Xu Z, Huang S, Zhang Y, Tao D. Webly-supervised fine-grained visual categorization via deep domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2016;40(5):1100–1113.
13. Yao Y, Hua X, Gao G, Sun Z, Li Z, Zhang J. Bridging the web data and fine-grained visual recognition via alleviating label noise and domain mismatch. In: ACM International Conference on Multimedia. 2020:1735–1744.
14. Zhang C, Wang Q, Xie G, Wu Q, Shen F, Tang Z. Robust Learning From Noisy Web Images Via Data Purification for Fine-Grained Recognition. *IEEE Transactions on Multimedia.* 2022;24:1198–1209.
15. Sun Z, Shen F, Huang D, et al. Pnp: Robust learning from noisy labels by probabilistic noise prediction. In: IEEE Conference on Computer Vision and Pattern Recognition. 2022:5311–5320.
16. Goldberger J, Ben-Reuven E. Training deep neural-networks using a noise adaptation layer. In: International Conference on Learning Representations. 2017:1–9.
17. Patrini G, Rozza A, Krishna Menon A, Nock R, Qu L. Making deep neural networks robust to label noise: A loss correction approach. In: IEEE Conference on Computer Vision and Pattern Recognition. 2017:1944–1952.
18. Jiang L, Zhou Z, Leung T, Li LJ, Fei-Fei L. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In: International Conference on Machine Learning. 2018:2304–2313.
19. Younesian T, Zhao Z, Ghiassi A, Birke R, Chen LY. Qactor: Active learning on noisy labels. In: Asian Conference on Machine Learning. 2021:548–563.
20. Malach E, Shalev-Shwartz S. Decoupling" when to update" from" how to update". In: Advances in Neural Information Processing Systems. 2017:960–970.
21. Han B, Yao Q, Yu X, et al. Co-teaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels. In: Advances in Neural Information Processing Systems. 2018:8527–8537.
22. Yao Y, Zhang J, Shen F, et al. Towards automatic construction of diverse, high-quality image datasets. *IEEE Transactions on Knowledge and Data Engineering.* 2019;32(6):1199–1211.
23. Yao Y, Zhang J, Shen F, Hua X, Xu J, Tang Z. Exploiting web images for dataset construction: A domain robust approach. *IEEE Transactions on Multimedia.* 2017;19(8):1771–1784.
24. Bergamo A, Torresani L. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In: Advances in Neural Information Processing Systems. 2010:181–189.
25. Wei XS, Xie CW, Wu J, Shen C. Mask-cnn: Localizing parts and selecting descriptors for fine-grained bird species categorization. *Pattern Recognition.* 2018;76:704–714.
26. Zhang N, Donahue J, Girshick R, Darrell T. Part-based R-CNNs for fine-grained category detection. In: European Conference on Computer Vision. 2014:834–849.
27. Huang S, Xu Z, Tao D, Zhang Y. Part-stacked cnn for fine-grained visual categorization. In: IEEE Conference on Computer Vision and Pattern Recognition. 2016:1173–1182.
28. Lin D, Shen X, Lu C, Jia J. Deep lac: Deep localization, alignment and classification for fine-grained recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. 2015:1666–1674.
29. Yang S, Bo L, Wang J, Shapiro LG. Unsupervised template learning for fine-grained object recognition. In: Advances in Neural Information Processing Systems. 2012:3122–3130.
30. Chai Y, Lempitsky V, Zisserman A. Symbiotic Segmentation and Part Localization for Fine-Grained Categorization. In: IEEE International Conference on Computer Vision. 2013:321–328.
31. Christopher K. Fine-grained object recognition with gnostic fields. In: IEEE Winter Conference on Applications of Computer Vision. 2014:23–30.
32. Gavves E, Fernando B, Snoek CG, Smeulders AW, Tuytelaars T. Local alignments for finegrained categorization. *International Journal of Computer Vision.* 2015;111(2):191–212.
33. Chen G, Yang J, Jin H, Shechtman E, Han TX. Selective Pooling Vector for Fine-Grained Recognition. In: IEEE Winter Conference on Applications of Computer Vision. 2015:860–867.
34. Xie S, Yang T, Wang X, Lin Y. Hyper-class augmented and regularized deep learning for fine-grained image classification. In: IEEE Conference on Computer Vision and Pattern Recognition. 2015:2645–2654.
35. Zhang X, Xiong H, Zhou W, Tian Q. Fused one-vs-all mid-level features for fine-grained visual categorization. In: ACM International Conference on Multimedia. 2014:287–296.
36. Fu J, Zheng H, Mei T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. 2017:4438–4446.
37. Lin TY, Roychowdhury A, Maji S. Bilinear CNN Models for Fine-grained Visual Recognition. In: IEEE Conference on Computer Vision. 2015:1449–1457.

38. Xiao T, Xu Y, Yang K, Zhang J, Peng Y, Zhang Z. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In: IEEE Conference on Computer Vision and Pattern Recognition. 2015:842–850.
39. Simon M, Rodner E. Neural Activation Constellations: Unsupervised Part Model Discovery with Convolutional Networks. In: IEEE International Conference on Computer Vision. 2015:1143–1151.
40. Zhang X, Xiong H, Zhou W, Lin W, Tian Q. Picking Deep Filter Responses for Fine-Grained Image Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. 2016:1134–1142.
41. Wang Y, Morariu VI, Davis LS. Learning a discriminative filter bank within a cnn for fine-grained recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. 2018:4148–4157.
42. Zheng H, Fu J, Tao M, Luo J. Learning Multi-attention Convolutional Neural Network for Fine-Grained Image Recognition. In: IEEE International Conference on Computer Vision. 2017:5209–5217.
43. Cui Y, Song Y, Sun C, Howard A, Belongie S. Large Scale Fine-Grained Categorization and Domain-Specific Transfer Learning. In: IEEE Conference on Computer Vision and Pattern Recognition. 2018:4109–4118.
44. He X, Peng Y. Fine-grained image classification via combining vision and language. In: IEEE Conference on Computer Vision and Pattern Recognition. 2017:5994–6002.
45. Van Horn G, Branson S, Farrell R, et al. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In: IEEE Conference on Computer Vision and Pattern Recognition. 2015:595–604.
46. Luo H, Lin G, Shen F, Huang X, Yao Y, Shen H. Robust-EQA: robust learning for embodied question answering with noisy labels. *IEEE Transactions on Neural Networks and Learning Systems.* 2023:1-12.
47. Reed S, Lee H, Anguelov D, Szegedy C, Erhan D, Rabinovich A. Training deep neural networks on noisy labels with bootstrapping. In: International Conference on Learning Representations. 2015:1–12.
48. Flatow D, Penner D. On the robustness of convnets to training on noisy labels. *Technical Report, Stanford University.* 2017.
49. Zhuang B, Liu L, Li Y, Shen C, Reid I. Attend in groups: a weakly-supervised deep learning framework for learning from web data. In: IEEE Conference on Computer Vision and Pattern Recognition. 2017:1878–1887.
50. Li LJ, Fei-Fei L. Optimol: automatic online picture collection via incremental model learning. *International Journal of Computer Vision.* 2010;88(2):147–168.
51. Zheng W, Chen S, Fu Z, Zhu F, Yan H, Yang J. Feature selection boosted by unselected features. *IEEE Transactions on Neural Networks and Learning Systems.* 2021;33(9):4562–4574.
52. Zhu F, Gao J, Xu C, Yang J, Tao D. On Selecting Effective Patterns for Fast Support Vector Regression Training. *IEEE Transactions on Neural Networks and Learning Systems.* 2018;29(8):3610-3622.
53. Chen T, Zhang J, Xie GS, Yao Y, Huang X, Tang Z. Classification Constrained Discriminator For Domain Adaptive Semantic Segmentation. In: IEEE International Conference on Multimedia and Expo. 2020:1–6.
54. Hoffman J, Pathak D, Darrell T, Saenko K. Detector discovery in the wild: Joint multiple instance and representation learning. In: IEEE Conference on Computer Vision and Pattern Recognition. 2015:2883–2891.
55. Wang X, Yan Y, Tang P, Bai X, Liu W. Revisiting multiple instance neural networks. *Pattern Recognition.* 2018;74:15–24.
56. Oquab M, Bottou L, Laptev I, Sivic J, others . Weakly supervised object recognition with convolutional neural networks. In: Advances in Neural Information Processing Systems. 2014:1545–5963.
57. Ilse M, Tomczak JM, Welling M. Attention-based deep multiple instance learning. *arXiv:1802.04712.* 2018.
58. Branson S, Van Horn G, Belongie S, Perona P. Bird species categorization using pose normalized deep convolutional nets. *arXiv:1406.2952.* 2014.
59. Zheng H, Fu J, Zha ZJ, Luo J. Looking for the Devil in the Details: Learning Trilinear Attention Sampling Network for Fine-Grained Image Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. 2019:5012–5021.
60. Chen Y, Bai Y, Zhang W, Mei T. Destruction and Construction Learning for Fine-Grained Image Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. 2019:5157–5166.
61. Ge W, Lin X, Yu Y. Weakly supervised complementary parts models for fine-grained image classification from the bottom up. In: IEEE Conference on Computer Vision and Pattern Recognition. 2019:3034–3043.
62. Chen X, Shrivastava A, Gupta A. Neil: Extracting visual knowledge from web data. In: IEEE International Conference on Computer Vision. 2013:1409–1416.
63. Niu L, Li W, Xu D. Visual recognition by learning from web data: A weakly supervised domain generalization approach. In: IEEE Conference on Computer Vision and Pattern Recognition. 2015:2774–2783.
64. Sukhbaatar S, Bruna J, Paluri M, Bourdev L, Fergus R. Training convolutional networks with noisy labels. *arXiv:1406.2080.* 2014.
65. Xiao T, Xia T, Yang Y, Huang C, Wang X. Learning from massive noisy labeled data for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition. 2015:2691–2699.
66. Zhang C, Yao Y, Liu H, et al. Web-Supervised Network with Softly Update-Drop Training for Fine-Grained Visual Classification. In: AAAI Conference on Artificial Intelligence. 2020:12781–12788.
67. Khosla A, Jayadevaprakash N, Yao B, Li FF. Novel dataset for fine-grained image categorization: Stanford dogs. In: CVPR Workshop on Fine-Grained Visual Categorization. 2011.
68. Liu X, Xia T, Wang J, Yang Y, Zhou F, Lin Y. Fully convolutional attention networks for fine-grained recognition. *arXiv:1603.06765.* 2016.
69. Zhao B, Wu X, Feng J, Peng Q, Yan S. Diversified visual attention networks for fine-grained object classification. *IEEE Transactions on Multimedia.* 2017;19(6):1245–1256.
70. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. 2016:770–778.