HCL: Hierarchical Consistency Learning for Webly Supervised Fine-Grained Recognition

Hongbo Sun, Xiangteng He and Yuxin Peng

***Abstract*—Webly supervised fine-grained recognition aims to distinguish subordinate categories (e.g., bird species) with freely available web data. It has significant research and application value for alleviating the costly professional manual annotations' dependence in the fine-grained recognition task. Nevertheless, there exists label noise in web data to decrease the model's recognition performance. Most existing methods attempt to select clean data via loss analyses, which favors easy samples to hinder mining subtle differences contained in hard samples. Inspired by the intrinsic trait of consistent semantic predictions among different hierarchies of clean samples in fine-grained recognition, we propose a hierarchical consistency learning (HCL) approach for detecting noisy samples and capturing multi-hierarchy discriminative clues simultaneously. Specifically, our HCL approach works in a coarse-to-fine order, which first explores the semantic consistency between the image level and object level through prediction distribution conformance analyses. The open-set noise (i.e., samples irrelevant to any fine-grained subcategory) is thus detected, and the visual object information is highlighted with image-object contrastive learning. Then, the semantic consistency between object-level and part-level prediction distributions is utilized for detecting closed-set noise (i.e., samples mislabeled as other fine-grained subcategories), and local discriminative information is enhanced with object-part contrastive learning. Extensive experiments and analyses on three widely-used webly supervised fine-grained benchmark datasets demonstrate that the proposed HCL approach can achieve new state-of-the-art. The code is available at https://github.com/PKU-ICST-MIPL/HCL_TMM2023.**

***Index Terms*—Webly supervised fine-grained recognition, hierarchical consistency learning, open-set noise, closed-set noise**

I. Introduction

FINE-GRAINED image recognition task is to identify the exact subcategory of a given basic category, such as classifying various bird species [1], car types [2], and aircraft models [3], which has significant research value in many real-life fields, such as biodiversity monitoring, intelligent agriculture, and intelligent transport. The deep neural network has brought remarkable progress in the fine-grained recognition task [4] for its strong image representation ability, which highly depends on a large amount of labeled training data. However, labeling fine-grained training data is extremely labor-intensive and time-consuming, which demands domain-specific expert knowledge to distinguish the subtle differences among subcategories for accurate annotation [5]. Thus, it becomes a severe obstacle to the generalization and practicability of fine-grained recognition models.

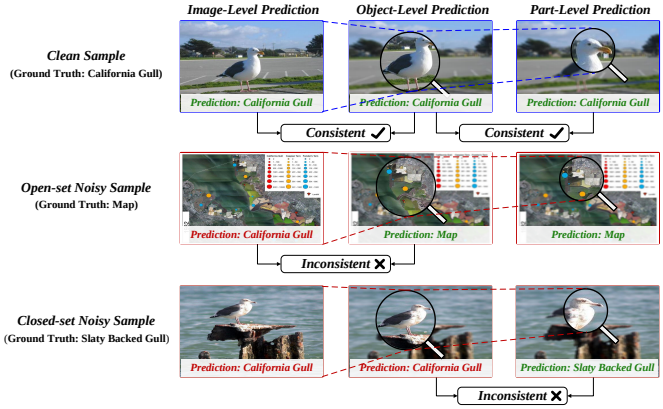

Fig. 1. The introduction of three kinds of web data in the webly supervised fine-grained recognition task. We propose a hierarchical consistency learning approach to imitate the human recognition process in a coarse-to-fine order for detecting noisy samples and mining discriminative clues simultaneously.

To alleviate the above problem, researchers begin to resort to the massive freely available web data to reduce reliance on manual annotations and obtain more practical fine-grained models [5]. Specifically, a series of webly supervised learning methods [5]–[13] are proposed for fine-grained recognition, which train the models with web images crawled from the public websites via querying class names. Though web images are cheap and easy to access, label noise exists due to errors of automatic tagging or non-expert labeling, which affects the model's recognition performance [14]. The label noise phenomenon is severe for fine-grained scenarios because of the high labeling demand for domain-specific expertise. Thus, webly supervised fine-grained recognition is very challenging and worthy of thorough studies, which simultaneously faces the problems of label noise and intrinsically subtle differences among fine-grained subcategories.

The label noise in the webly supervised fine-grained recognition task can be divided into two types, i.e., open-set label noise and closed-set label noise [15]. As shown in Fig. 1, the open-set label noise refers to the sample that does not belong to any fine-grained subcategory, which is out of distribution. For example, a "Map" image of the bird habitat is mislabeled as the "California Gull" because of the text's appearance in the image. The closed-set label noise refers to the samples

This work was supported by the National Natural Science Foundation of China (61925201, 62132001, 62272013) and Beijing Natural Science Foundation (4232005).

Hongbo Sun is with the Wangxuan Institute of Computer Technology, Peking University, Beijing 100871, China.

Xiangteng He and Yuxin Peng are with the Wangxuan Institute of Computer Technology, Peking University, Beijing 100871, China, and the Peng Cheng Laboratory, Shenzhen 518055, China.

Corresponding author: Yuxin Peng (e-mail: pengyuxin@pku.edu.cn).

mislabeled as other subcategories. For example, the "Slaty Backed Gull" image is mislabeled as "California Gull", which is an in-distribution noisy sample resulting from the extremely small inter-class variance. The open-set and closed-set label noise make the model confused or overfit during the training stage, which weakens the model's ability to recognize different subcategories to a great extent.

To deal with the above label noise problems, loss correction and sample selection are two leading solutions in webly supervised fine-grained recognition methods. Specifically, the loss correction methods attempt to modify the loss function for improving robustness [16]–[18] and estimate the noise label transition matrix [19], [20]. However, the loss correction methods generally cannot perform well in intricacy scenarios [21], and the transition matrix is hard to estimate for the open-set noise in the real world. Sample selection methods are intuitive to select clean samples for the model's training. Inspired by the observations that the deep learning models tend to fit simple samples with commonality before fitting hard samples [22], existing methods [5], [7], [8], [13], [23] generally select small-loss training samples as the clean samples for training. Nevertheless, the above sample selection methods based on loss analyses generally result in the domination of easy samples in the training process, affecting discriminative information mining in fine-grained recognition. Besides, the detected noisy samples are generally abandoned directly, which ignores their comparative value in enhancing the model's recognition ability.

Inspired by the human recognition process of progressively focusing on identifiable information for classification, we observe semantic consistency among different hierarchies when predicting the semantic label of the clean sample, as shown in Fig. 1. On the contrary, open-set noisy samples can be detected when analyzing the semantic consistency between the image-level and object-level prediction distributions because the visual object in the sample does not belong to the labeled subcategory. Closed-set noisy samples can be detected in a similar way when analyzing the semantic consistency between the object-level and part-level prediction distributions because the latter can provide subtle yet important local discriminative clues for predicting a specific subcategory, which differs from the given wrong label. Given all the above analyses, we propose a ***hierarchical consistency learning (HCL)*** approach for webly supervised fine-grained recognition. The main contributions can be summarized as follows:

- We propose to imitate the human progressive focusing recognition mechanism and utilize semantic consistency among different hierarchies to eliminate label noise as well as mine discriminative information from multiple hierarchies, which aims at simultaneously solving the two critical problems in the webly supervised fine-grained recognition task.
- Open-set noisy samples are detected by measuring the semantic consistency between the image-level and object-level prediction distributions. The model's recognition ability of the highlighted visual object is boosted via image-object contrastive learning.
- Closed-set noisy samples are detected by calculating the semantic consistency between the object-level and part-level prediction distributions. Local discriminative information is thoroughly mined for image classification via object-part contrastive learning.
- Extensive comparison experiments on three real-world webly supervised fine-grained benchmark datasets demonstrate that the proposed HCL approach achieves new state-of-the-art.

The rest of the paper is organized as follows: Section II briefly reviews the related work on fine-grained image recognition, webly supervised learning, and contrastive learning. Section III elaborates on the proposed HCL approach, and Section IV shows the experiments, analyses, and ablation studies. Finally, Section V concludes the paper.

II. Related Work

This section briefly reviews related works about fine-grained image recognition, webly supervised learning, and contrastive learning.

A. Fine-grained Image Recognition

Fine-grained image recognition aims to recognize various subordinate categories belonging to the same basic category, such as bird species, which generally faces the challenge of large intra-class variance and small inter-class variance [4]. Existing fine-grained image recognition methods can be roughly summarized into two paradigms, i.e., (1) recognition by discriminative regions localization and (2) recognition by end-to-end feature learning. In the first paradigm [24]–[31], discriminative regions are utilized explicitly or implicitly for feature extraction and classification. Peng et al. [24] propose a two-level attention mechanism to promote discriminative regions localization, which does not need object bounding boxes and part annotations. Song et al. [28] propose the progressive mask attention model to discover discriminative parts gradually. Sun et al. [31] propose introducing the object structure information into the vision transformer to highlight significant regions and boost discriminative feature learning. Xu et al. [32] propose an ensemble learning transformer to select desired tokens to extract features for classification based on the attention map. In the second paradigm [33]–[36], high-order features are generally designed as robust image representation. Lin et al. [33] propose a bilinear convolution neural network framework to model local pairwise feature interactions to generate robust feature representation for classification. Tan et al. [36] devise a multi-scale selective hierarchical biquadratic pooling approach to model intra-layer and inter-layer feature interactions for extracting distinctive features. In addition to the above two main kinds of methods, Du et al. [37] propose to utilize the jigsaw puzzle generator for data augmentation, which facilitates the model to learn image information of different granularities. Chang et al. [38] propose disentangling feature learning for better classification performance with multi-granularity labels.

Though the above methods achieve promising fine-grained recognition performance, they generally lack sufficient fine-grained training data, which restricts their generalization ability and practicability to a large extent. Instead of the costly

and limited training data annotated by experts, utilizing the massive freely available web data for training fine-grained models becomes an alternative solution, spawning a series of webly supervised learning methods.

B. Webly Supervised Learning

Webly supervised learning methods utilize freely available web data to train deep learning models, which are obtained through retrieving search engines and social websites, such as Google Image Search Engine[1], Bing Image Search Engine[2], and Flickr[3], with class names as keywords. However, noisy samples in the web data generally exist due to errors of automatic tagging or non-expert labeling, which decreases the model's performance [5]. Existing webly supervised learning methods can be classified into two main types, i.e., the loss correction methods [16]–[20], [39] and sample selection methods [5], [7], [8], [12], [13], [23]. For the first loss correction paradigm, Ghosh et al. [16] propose mean absolute error loss, and Wang et al. [18] propose the symmetric cross-entropy loss. PNP [12] proposes to predict the noise type for input samples and adopts different training loss functions accordingly. Though the above loss design can promote the model's robustness to some extent, they generally cannot support intricate scenarios [21], such as indistinguishable noisy samples. To address this problem, some works attempt to model the label transition matrix for the noisy dataset. Goldberger et al. [39] propose a noise adaption layer to act as the transition matrix. Patrini et al. [19] propose to utilize the loss function to estimate the transition matrix. However, these methods are applicable to closed-set noisy scenarios, which cannot solve the open-set noise in the web data. In the second paradigm, sample selection methods generally select the clean data from the noisy dataset for training the model. Co-Teaching [7] trains two networks simultaneously to select small-loss training samples for the peer network. Yu et al. [23] propose the "update by disagreement" method for training the model. Peer-learning [5] simultaneously trains two networks to fetch clean data and direct each other in the training stage. JoCoR [8] proposes a joint loss to select clean samples for optimizing two networks. MS-DeJOR [13] exploits the symmetric Kullback Leibler divergence between the dual networks to improve the selection performance of clean data for better fine-grained classification accuracy.

Despite obtaining promising recognizing performance, the above sample selection methods generally tend to choose small-loss samples to result in the domination of easy samples, which affects mining the subtle differences for fine-grained classification. We propose to utilize the semantic consistency among different hierarchies to simultaneously take noisy sample detection and discriminative information mining of multiple hierarchies into account for webly supervised fine-grained recognition.

[1]https://images.google.com/
[2]https://www.bing.com/images/
[3]https://www.flickr.com/

C. Contrastive Learning

Contrastive learning aims to discover the pattern that is specific to one set relative to others [40]. Recently, contrastive learning has been widely applied to self-supervised representation learning [41]–[44]. SimCLR [43] proposes to conduct contrastive learning via constructing the positive pair by sampling two images from different transformations of the same image and the negative pair by sampling two different images. MoCo [42] trains a visual representation encoder by matching an encoded query to a dictionary of encoded keys with contrastive learning loss. Large-scale pre-trained vision-language models, such as the well-known CLIP [44], usually adopt contrastive learning to conduct the self-supervised pre-training task with massive paired data, which has achieved great success in general representation learning.

Inspired by the advantage of contrastive learning in enhancing feature representation, we propose multi-hierarchy contrastive learning among clean samples and noisy samples to boost the model' discriminability of noise and representation ability for image, object and part, which is beneficial for webly supervised fine-grained recognition.

III. Approach

This section introduces the overall pipeline of the proposed hierarchical consistency learning (HCL) approach and elaborates on each component.

A. Overview

The whole pipeline of the proposed HCL approach is shown in Fig. 2. Our HCL approach works in a coarse-to-fine order to imitate the human recognition process for detecting noisy samples and mining discriminative information simultaneously. It comprises the coarse stage of detecting open-set noisy samples and the fine stage of detecting closed-set noisy samples. Finally, the clean samples are naturally selected for training, and multi-hierarchy contrastive learning is utilized to enhance the discriminability of features for fine-grained recognition.

B. Open-set Noisy Samples Detection

In the fine-grained recognition task, different fine-grained subcategories generally only have small variances in local regions. Thus, the discriminative parts of the object play an essential role in the final classification result, which means that there exists inherent consistent semantic prediction among the image, object, and part in the fine-grained clean sample. Thus, clean samples, open-set noisy samples, and closed-set noisy samples can be detected based on the characteristic.

Open-set noisy samples refer to the samples that are mislabeled, where the visual objects do not belong to any subcategory. Thus, the object-level semantic prediction for the open-set noisy sample is generally different from the given label. For example, in Fig. 1, a "Map" image of the birds' habitat is mislabeled as the "California Gull". A significant variance exists between the object-level semantic prediction of "Map" and the annotated image-level semantic label of

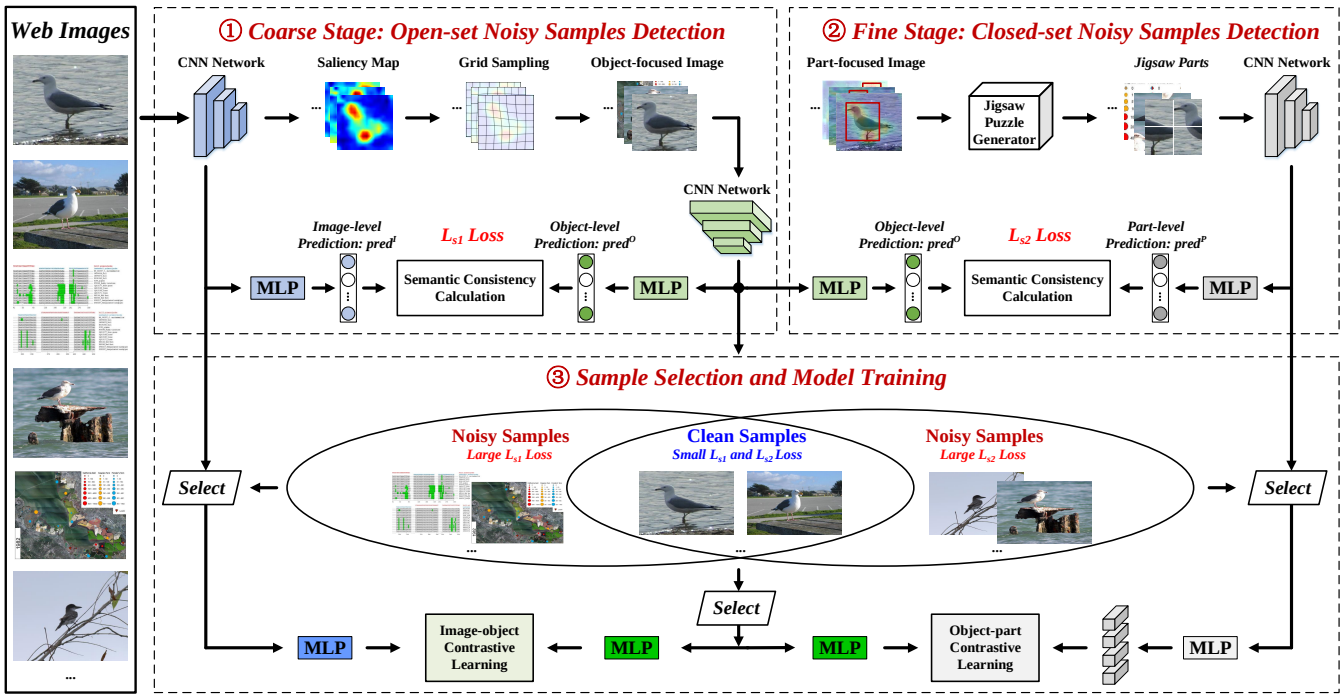

Fig. 2. The pipeline of the proposed HCL approach. In the coarse stage, the open-set noisy samples are detected via semantic consistency calculation between the image-level and object-level prediction distributions. In the fine stage, the closed-set noisy samples are detected via the semantic consistency calculation between the object-level and part-level prediction distributions. Finally, clean samples are naturally selected for training, and multi-hierarchy contrastive learning is adopted to enhance the features' discriminability for fine-grained recognition.

"California Gull". By contrast, clean samples have higher consistency between the image-level and object-level semantic predictions, which is thus utilized for detecting open-set noisy samples in the coarse stage.

For extracting and exploiting the object information without the bounding box annotation information, we first locate the visual object extent in the original image in an unsupervised way. Inspired by the observation that the high activation response area of feature maps extracted by the convolutional neural network (CNN) usually corresponds to significant object regions for final classification decision [45], we utilize the saliency map based on feature maps for locating the object extent. Specifically, the image is first fed into the CNN such as ResNet50 [46]. Feature maps are extracted from different convolution layers for average calculation processing to obtain the saliency map. As shown in Fig. 3, the saliency map obtained from the deep convolution layer focuses on the whole object in comparison with that obtained from the shallow convolution layer. Thus, we adopt the saliency map obtained from the last convolution layer for object localization. The above process is formulated as follows. For a given image $I(x, y)$, the saliency map is calculated as follows:

$$S(x,y) = \sum_{i=1}^{N} \omega_i FM_i(x,y), \tag{1}$$

$$w_i = \frac{1}{M}, \tag{2}$$

where $S(x, y)$ is the value of the saliency map in location (x, y), $FM_i(x, y)$ denotes the activation value of i_{th} feature map in the spatial location (x, y) extracted from $I(x, y)$, w_i denotes the weight and M denotes the number of feature maps. We find that simple average weighting can achieve promising visual object localization results.

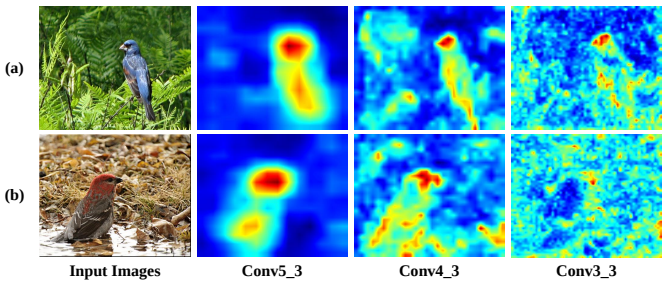

Fig. 3. Saliency maps obtained from different convolution layers, i.e., "conv5_3", "conv4_3", and "conv3_3" of ResNet 50. The saliency map obtained from the deep convolution layer focuses on the whole object.

We emphasize and amplify the visual object in the image to obtain the new object-focused image, as shown in Fig. 2, which imitates the human's gradual focus recognizing process. We utilize the grid sampling method to amplify the visual object while maintaining the surrounding information based on the above saliency map. Concretely, for the image $I(x, y)$, the object amplification with grid sampling aims to construct a mapping to scale up the high-saliency regions while suppressing the low-saliency regions with two functions $h(x, y)$ and $g(x, y)$. Then, the sampled image, i.e., the new object-focused image, can be obtained via $O(x, y) = I(h(x, y), g(x, y))$. The design of $h(x, y)$ and $g(x, y)$ is to map the image pixels

proportionally according to the normalized weight of the saliency map. An approximation solution is to obtain $h(x, y)$ and $g(x, y)$ that satisfies:

$$\int_0^{h(x,y)} \int_0^{g(x,y)} S(u, v)dudv = xy. \tag{3}$$

Following [47], the solution is as follows:

$$h(x, y) = \frac{\sum_{u,v} S(u, v)k((u, v), (x, y))u}{\sum_{u,v} S(u, v)k((u, v), (x, y))}, \tag{4}$$

$$g(x, y) = \frac{\sum_{u,v} S(u, v)k((u, v), (x, y))v}{\sum_{u,v} S(u, v)k((u, v), (x, y))}, \tag{5}$$

where the $k(\cdot, \cdot)$ denotes the Gaussian distance kernel function. It is utilized as regularization to avoid extreme cases, such as all the pixel values approaching the same value. In this way, the visual object in the image can be amplified based on the extracted saliency map to obtain the new object-focused image. More visualization results can be found in Fig. 4.

Feature maps extracted from different convolution layers of the CNN contain information of different scales and complement each other, as depicted in [48]. Thus, we first extract the feature maps FM_{N-2}, FM_{N-1}, and FM_N from the last three convolution blocks considering the semantic representation ability of high convolution layers, where N denotes the number of convolution blocks in CNN. Then, the new convolution blocks with two convolution layers are added to further extract features to obtain the F_{N-2}, F_{N-1}, and F_N. Maxpooling operation is utilized to transform the above feature maps into feature vectors f_{N-2}, f_{N-1}, and f_N. To get the comprehensive feature representation, we concatenate the above feature vectors as f_c:

$$f_c = concat(f_{N-2}, f_{N-1}, f_N). \tag{6}$$

By applying the classifiers consisting of two fully connected layers for the four feature vectors separately, we accordingly obtain the prediction vector p_{N-2}, p_{N-1}, p_N, and p_c, which contain the semantic prediction distribution information. By this means, we process the original image and object-focused image with two separate sub-networks, as shown in Fig. 2.

Given the phenomenon that deep learning models tend to fit simple samples owning commonality before learning hard samples [13], [22], the classification loss is an important indicator for detecting clean samples with consistent semantic prediction distribution:

$$L^I_{cls1} = \sum_{i=N-2}^{N} CE(p^I_i, y) + 2 \times CE(p^I_c, y), \tag{7}$$

$$L^O_{cls1} = \sum_{i=N-2}^{N} CE(p^O_i, y) + 2 \times CE(p^O_c, y), \tag{8}$$

where L^I_{cls1} and L^O_{cls1} are corresponding classification losses of the original image and object-focused image, $CE(\cdot)$ denotes the cross entropy calculation, p^I_i and p^I_c belong to set $pred^I$ which denote the prediction vectors for the original image, p^O_i and p^O_c belong to set $pred^O$ which denote the prediction vectors for the object-focused image, y is the given label. The small-loss samples typically correspond to clean samples, which have consistent semantics. Besides, we also utilize the Jensen–Shannon (JS) divergence to calculate the semantic consistency as follows:

$$L_{con1} = \sum_{i=N-2}^{N} JS(p^I_i||p^O_i) + 2 \times JS(p^I_c||p^O_c). \tag{9}$$

$JS(\cdot)$ denotes calculating the distribution variance of prediction vectors after normalization. A lower JS value represents a lower distribution difference, i.e., higher semantic consistency. Given all the above analyses, we propose the selection loss function L_{s1} for detecting open-set noisy samples in the coarse stage as follows:

$$L_{s1} = L^I_{cls1} + L^O_{cls1} + \lambda L_{con1}, \tag{10}$$

where λ is set to 10 to balance the loss items according to their loss values observed in the experiments. Based on the L_{s1}, we detect the large-loss samples as open-set noisy samples in the training process.

C. Closed-set Noisy Samples Detection

The closed-set noisy samples refer to the samples that are mislabeled as other subcategories, such as the "Slaty Backed Gull" mislabeled as "California Gull" in Fig. 1, which are generally caused by the fine-grained essence, i.e., small inter-class variance. Many researches on fine-grained recognition adopt the localization and recognition paradigm, which detects and utilizes the information of salient regions for classification. This is because key parts in salient regions contain essential discriminative information for distinguishing various subcategories. The semantic prediction information of key parts in the closed-set noisy sample is generally inconsistent with that obtained from the visual object. For example, as shown in the third row of Fig. 1, the model's predictions for the original image and the visual object are "California Gull". However, when the model focuses on the local discriminative part, i.e., the beak shape and texture, it gets the prediction of "Slaty Backed Gull". Thus, there exists inconsistency between the object-level prediction and part-level prediction. Based on the above analyses, we propose to utilize the semantic consistency between object-level and part-level prediction distributions to detect the closed-set noisy samples in the fine stage, as shown in Fig. 2.

As described in Section III-B, the saliency map has presented the significance distribution. Thus, we adopt the OTSU algorithm [49] to binarize the saliency map and detect the largest connected area as the part-focused image, as shown in Fig. 2 and Fig. 4, which generally comprises key parts. Inspired by the claim in [48] that training data of different granularities can boost the model's recognition ability on local details, we adopt the jigsaw puzzle generator to shuffle and recompose the jigsaw parts of the part-focused image. By this means, the model is driven to focus on the key parts and extract local discriminative features for fine-grained recognition. Concretely, for preserving the completeness of parts, the jigsaw number is empirically set as 2×2. The image

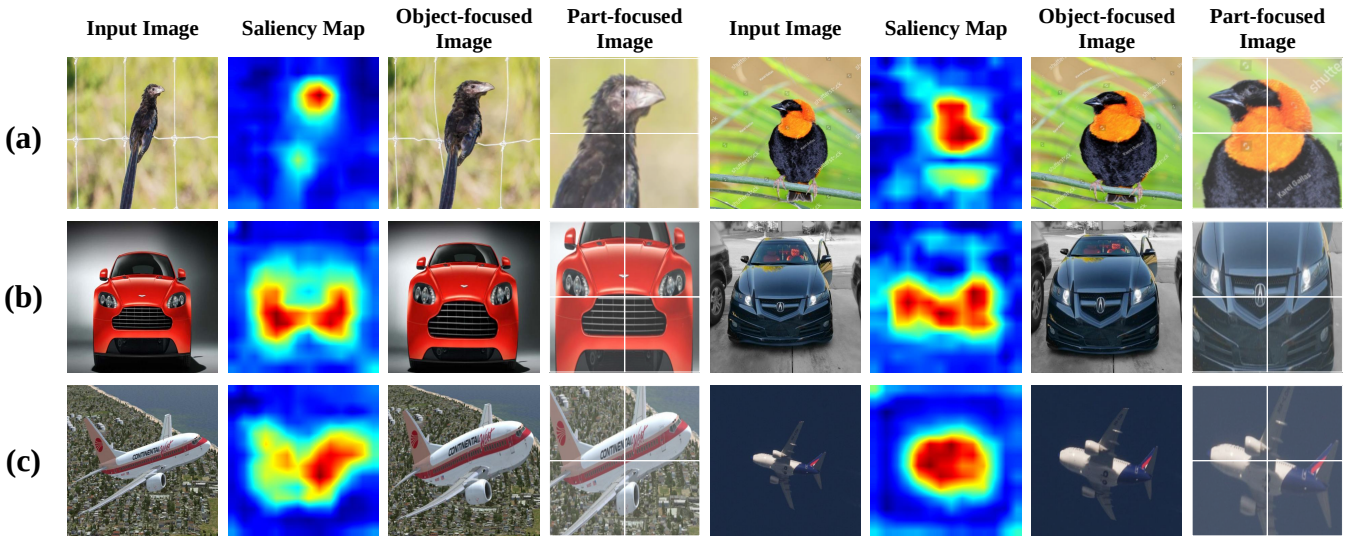

Fig. 4. The visualization examples of the proposed HCL approach's pipeline on three webly supervised fine-grained datasets. The visual object is amplified and the key parts are focused on with the saliency map, which is utilized to conduct semantic consistency analyses for detecting noisy samples.

of recomposed jigsaw parts is input into the third sub-network for feature extraction, and we adopt a similar way to Section III-B to get the selection loss as follows:

$$L_{cls2}^{O} = \sum_{i=N-2}^{N} CE(p_i^O, y) + 2 \times CE(p_c^O, y), \tag{11}$$

$$L_{cls2}^{P} = \sum_{i=N-2}^{N} CE(p_i^P, y) + 2 \times CE(p_c^P, y), \tag{12}$$

$$L_{con2} = \sum_{i=N-2}^{N} JS(p_i^O || p_i^P) + 2 \times JS(p_c^O || p_c^P), \tag{13}$$

$$L_{s2} = L_{cls2}^{O} + L_{cls2}^{P} + \lambda L_{con2}, \tag{14}$$

where L_{cls2}^{O} and L_{cls2}^{P} are classification losses for the object-focused image and part-focused image, respectively, $CE(\cdot)$ denotes the cross entropy calculation, p_i^O and p_c^O belong to set $pred^O$ which denote the prediction vectors for the object-focused image, p_i^P and p_c^P belong to set $pred^P$ which denote the prediction vectors for the part-focused image, y is the given label, L_{con2} denotes their prediction distributions' JS variance. Based on the semantic consistency analyses of the object-level and part-level prediction distributions, i.e., the L_{s2}, where λ is also set to 10 empirically to balance the loss items, the large-loss samples are treated as the closed-set noisy samples during the training process.

D. Sample Selection and Model Training

For obtaining clean samples with correct labels, we first select two small-loss sample sets according to Eq. 10 and Eq. 14 with the drop rate γ, i.e., selecting the first $1-\gamma$ small-loss samples. Then, samples from the intersection set of the above two selected sample sets are considered as clean samples, while the others are considered as noisy ones, as shown in Fig. 2. Inspired by the fact that the noisy samples also contain the information of other subcategories, we propose multi-hierarchy contrastive learning to improve the model's recognition ability.

Concretely, we conduct image-object and object-part contrastive learning, respectively. In image-object contrastive learning, the holistic feature representations of the original image and object-focused image, i.e., $f_c^{'I}$ and $f_c^{'O}$, are obtained in a similar way to f_c in Eq. 6. Then, the contrastive learning loss is constructed:

$$L_{CL1} = -E(\log \frac{\exp(sim(f_c^{'I}, f_c^{'O+})/\tau)}{\sum \exp(sim(f_c^{'I}, f_c^{'O})/\tau)}), \tag{15}$$

where $E(\cdot)$ denotes the expectation calculation, $sim(\cdot)$ denotes the cosine similarity calculation, τ is the temperature parameter which is set as 0.1 empirically, $(f_c^{'I}, f_c^{'O})$ denotes the image-object pair and $(f_c^{'I}, f_c^{'O+})$ denotes the positive pair which has the same label. It is noted that we abandon the positive pairs which contain noisy samples to avoid their disruption. Through the image-object contrastive learning, the model's ability to recognize visual objects of different subcategories is improved.

During the training stage, we adopt the linear warm-up strategy with the parameter T_W as the number of warm-up epochs. The training loss of the two sub-networks designed for the original image and the object-focused image is calculated as follows:

$$L_1 = L_{s1}(x_c) + \alpha \times L_{CL1}(x), \tag{16}$$

where x denotes input samples and x_c denotes the selected clean samples in x, α is a weight parameter. The model learns both the fine-grained variances of different subcategories and the semantic consistency between the image level and object level with clean samples by optimizing the $L_{s1}(x_c)$. The feature representation is enhanced with all the samples by optimizing the contrastive learning loss $L_{CL1}(x)$.

Then, we adopt the object-part contrastive learning to mine the local subtle yet distinctive information for fine-grained classification. The feature of the part-focused image is obtained in a similar way to f_c in Eq. 6 and divided into four vectors, i.e., $f_c^{'P_n}$, by the maxpooling operation, which correspond to jigsaw parts, as shown in Fig. 2. The object-part contrastive learning loss is calculated as follows:

$$L_{CL2} = -E(\frac{1}{4}\sum_{n=1}^{4}\log\frac{\exp(sim(f_c^{'O}, f_c^{'P_n+})/\tau)}{\sum\exp(sim(f_c^{'O}, f_c^{'P_n})/\tau)}), \quad (17)$$

where $(f_c^{'O}, f_c^{'P_n})$ denotes the object-part pair and $(f_c^{'O}, f_c^{'P_n+})$ denotes the positive pair with the same label. The positive pairs that contain noisy samples are also abandoned. Through the object-part contrastive learning, the local distinguishable information of a specific subcategory is mined and enhanced from both the object and part levels. The training loss of two sub-networks designed for the object-focused image and part-focused image is calculated as follows:

$$L_2 = L_{s2}(x_c) + \alpha \times L_{CL2}(x), \quad (18)$$

where x and x_c denote input samples and selected clean samples, respectively, α is a weight parameter. Finally, the total training loss is obtained as follows:

$$L = L_1 + L_2. \quad (19)$$

Overall, noisy samples are filtered out and multi-hierarchy information is learned by three different sub-networks in the proposed HCL approach, which provide comprehensive feature representations to focus on the image, object, and part, respectively. They complement each other to achieve promising webly supervised fine-grained recognition performance.

E. Model Inference

In the test stage, the test image is input into the first sub-network to obtain the prediction vector p_c^I. Then, the visual object is emphasized and amplified to obtain the object-focused image based on the extracted saliency map. The part-focused image is obtained by cropping the salient image region from the original image. The object-focused image and the part-focused image are input into the second sub-network and the third sub-network to get the prediction vectors p_c^O and p_c^P, respectively.

Finally, we merge the prediction vectors from the three sub-networks by utilizing the following equation:

$$p = p_c^I + p_c^O + p_c^P, \quad (20)$$

where p is the final prediction vector of the test image. Webly supervised fine-grained recognition performance is boosted by filtering out noisy samples, considering multi-hierarchy information, and emphasizing distinguishable clues comprehensively in the proposed HCL approach.

IV. Experiments

In this section, we conduct comparison experiments with state-of-the-art methods on three webly supervised fine-grained benchmark datasets to validate the effectiveness of the proposed HCL approach. Meanwhile, ablation studies, parameter experiments, backbone network experiments, and visualization experiments are also conducted to verify the importance of each proposed component.

A. Datasets and Evaluation Metric

Three popular webly supervised fine-grained datasets, i.e., Web-Bird, Web-Car, and Web-Aircraft proposed in [5], are adopted in the experiments. Detailed descriptions of the three datasets are as follows:

- Web-Bird is a fine-grained web bird dataset. It contains 200 bird subcategories. 18,388 images crawled from the Internet are utilized for training. 5,794 clean, correctly labeled test images from the standard fine-grained dataset CUB-200-2011 [1] are utilized for test.
- Web-Car is a fine-grained web car dataset. It covers 196 car models. 21,448 images collected from the Internet are utilized for training and 8,041 clean, correctly labeled test images from the standard fine-grained dataset Stanford Cars [2] are utilized for test.
- Web-Aircraft is a fine-grained web aircraft dataset. It consists of 100 aircraft types. 13,503 images from the Internet are utilized for training and 3,333 clean, correctly labeled test images from the standard fine-grained dataset FGVC-Aircraft [3] are utilized for test.

We adopt the widely used classification accuracy to evaluate the performance of the proposed HCL approach and other comparison methods.

B. Implementation Details

In this work, we select ResNet50 [46] as the backbone network, and we only use the class labels of the images without other annotations during the training phase. The three CNNs in Fig. 2 are specifically designed to extract features and get the predictions from the original image level, visual object level, and local part level, which do not share parameters. They are trained together through the proposed hierarchical consistency learning. Input images are resized into the size of 550×550 and randomly cropped into the size of 448×448. Data augmentation and label smoothing with a smooth value of 0.1 are adopted in training. We adopt the stochastic gradient descent (SGD) optimizer with momentum set as 0.9 and weight decay set as 1e-5. We set the initial learning rate for the parameters from the pre-trained ResNet50 backbone network as 2e-4. As for the other newly added parameters, the initial learning rate is set as 2e-3. The drop rate γ is set as 0.35, 0.25, 0.15 for Web-Bird, Web-Car, and Web-Aircraft, respectively. The number of warm-up training epochs T_W is set as 10. The weight parameter α in Eq. 16 and Eq. 18 is set as 1. The total number of training epochs is set as 100, and the batch size is set as 30. The cosine annealing schedule is utilized to update the learning rate. In the test phase, the images are resized into

TABLE I
COMPARISON EXPERIMENTS WITH STATE-OF-THE-ART METHODS ON THREE WEBLY SUPERVISED FINE-GRAINED DATASETS, WEB-BIRD, WEB-CAR, AND WEB-AIRCRAFT DATASETS. **BOLD VALUE** INDICATES THE BEST PERFORMANCE AND UNDERLINED VALUE INDICATES THE SUBOPTIMAL PERFORMANCE.

Methods	Publications	Backbone	Accuracy(%)			
			Web-Bird	*Web-Car*	*Web-Aircraft*	*Average*
Decoupling [6]	NeurIPS 2017	ResNet50	71.6	79.4	75.9	75.6
Co-teaching [7]	NeurIPS 2018	ResNet50	76.7	85.0	79.5	80.4
Co-teaching+ [23]	ICML 2019	ResNet50	70.1	76.8	74.8	73.9
PENCIL [50]	CVPR 2019	ResNet50	75.1	81.7	78.8	78.5
JoCoR [8]	CVPR 2020	ResNet50	79.2	85.1	80.1	81.5
AFM [51]	ECCV 2020	ResNet50	76.4	83.5	81.0	80.3
Self-adaptive [9]	NeurIPS 2020	ResNet50	78.5	78.2	77.9	78.2
Peer-learning [5]	ICCV 2021	B-CNN (VGG-16)	76.5	78.5	74.4	76.5
PLC [10]	ICLR 2021	ResNet50	76.2	81.9	79.2	79.1
Jo-SRC [11]	CVPR 2021	ResNet50	81.2	88.1	82.7	84.0
Liu et al. [15]	TMM 2022	B-CNN (VGG-16)	78.5	82.2	75.4	78.7
Co-LDL [52]	TMM 2022	ResNet50	81.0	89.2	83.8	84.7
PNP [12]	CVPR 2022	ResNet50	81.9	90.1	85.5	85.8
CLAR-CRSSC [53]	TMM 2023	ResNet50	82.9	88.6	82.8	84.8
MS-DeJOR [13]	PR 2023	ResNet50	83.7	88.4	88.5	86.9
Our HCL method	**This paper**	**ResNet50**	**86.1**	**91.6**	**92.5**	**90.1**

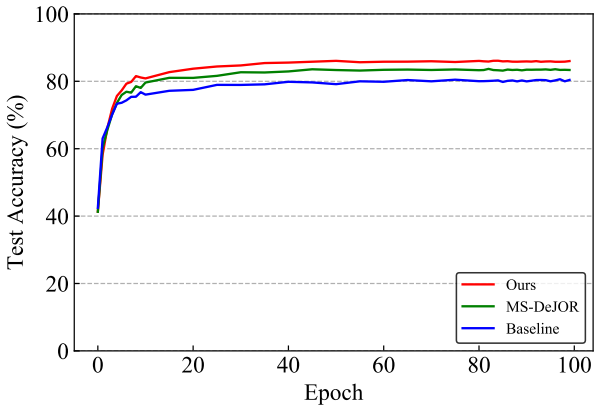

Fig. 5. The test accuracies of the baseline, MS-DeJOR, and ours on the Web-Bird dataset. The test accuracy data of MS-DeJOR in the figure is obtained by running its official code.

the size of 550×550 and cropped into the size of 448×448 from the center. All the experiments are implemented with Pytorch on two NVIDIA A40 GPUs.

C. Comparison With State-of-the-art Methods

We compare the proposed HCL approach with state-of-the-art (SOTA) methods on the above three webly supervised fine-grained (WSFG) datasets, and the results are shown in Table I and Fig. 5. We can observe that:

- The proposed HCL approach achieves better performance than all the comparison methods on the three WSFG datasets, achieving **86.1%**, **91.6%**, and **92.5%** test accuracy on Web-Bird, Web-Car, and Web-Aircraft, respectively. Compared with PNP [12], i.e., the representative SOTA method of the loss function correction paradigm introduced in Section II-B, our HCL approach achieves **4.2%**, **1.5%** and **7.0%** performance gains on the above three datasets. PNP proposes to predict noise probability for each sample with the noise predictor network and adopt distinct loss functions for different data types. Though achieving promising performance, it ignores discriminative information mining which is essential for fine-grained classification. We attribute the performance improvement brought by our HCL approach to utilizing multi-hierarchy information, which emphasizes the visual object and mines local discriminative clues. Besides, the features extracted from the original, object-focused, and part-focused images complement each other to achieve better fine-grained recognition performance.
- Compared with the representative SOTA methods MS-DeJOR [13] and CLAR-CRSSC [53] of the sample selection paradigm introduced in Section II-B, our HCL approach also achieves **2.4%**, **3.2%**, **4.0%** and **3.2%**, **3.0%**, **9.7%** performance gains on the three WSFG datasets, respectively. We attribute the improvements to the utilization of intrinsic semantic consistency among the image-level, object-level, and part-level prediction distributions in selecting clean samples. The noisy samples are thus eliminated and discriminative information from multiple hierarchies is learned and enhanced by contrastive learning, which is beneficial for improving webly supervised fine-grained recognition performance.
- On the average test accuracy metric, the proposed HCL approach can achieve **90.1%** fine-grained classification accuracy. It surpasses the suboptimal method by a margin of **3.2%** to verify the effectiveness and promising practicality of our proposed HCL approach in the real-world webly supervised fine-grained recognition scenario. Fig. 5 shows the test accuracy trends of the baseline, MS-DeJOR, and our proposed HCL approach with the increase of training epochs on the Web-Bird dataset, which presents the whole training process. The superiority of our proposed HCL approach in classification accuracy and performance stability can be observed in this figure.

TABLE II
ABLATION STUDIES OF EACH COMPONENT IN THE PROPOSED HCL APPROACH ON THE WEB-BIRD, WEB-CAR, AND WEB-AIRCRAFT DATASETS. W/O DENOTES WITHOUT, AND JS DENOTES JENSEN–SHANNON DISTRIBUTION VARIANCE CALCULATION, I.E., THE UTILIZATION OF EQ. 9 AND EQ. 13. **BOLD VALUE** INDICATES THE BEST PERFORMANCE.

Methods	Web-Bird(%)	Web-Car(%)	Web-Aircraft(%)
ResNet 50	78.8	86.1	87.9
ResNet 50 + multi-block features (Baseline)	81.2	87.7	89.3
Baseline + open-set noisy samples detection	85.2	91.1	91.4
Baseline + open-set & closed-set noisy samples detection (HCL)	**86.1**	**91.6**	**92.5**
HCL w/o JS	85.4	90.8	91.7

TABLE III
EXPERIMENTS ON DIFFERENT BACKBONE NETWORKS. **BOLD VALUE** INDICATES THE BEST PERFORMANCE.

Methods	Backbone	Web-Bird(%)	Web-Car(%)	Web-Aircraft(%)
ResNet50	-	78.8	86.1	87.9
ResNet101	-	79.4	86.9	88.9
ResNet152	-	80.4	87.8	89.3
Our HCL method	ResNet50	86.1	91.6	92.5
Our HCL method	ResNet101	86.7	92.2	93.0
Our HCL method	**ResNet152**	**87.2**	**92.4**	**93.6**

Our proposed HCL approach, denoted by the red curve, can perform better than MS-DeJOR, denoted by the green curve, and the baseline method, denoted by the blue curve, by distinct margins, which validates the effectiveness of the proposed approach and its components.

D. Ablation Studies

In this section, we conduct ablation studies on the three WSFG datasets to verify the effectiveness of each component in the proposed HCL approach. The experimental results are shown in Table II. We can observe that:

- Based on the standard ResNet 50, our baseline method concatenates the features from different convolution blocks to achieve the fine-grained classification accuracy of 81.2%, which outperforms the pure ResNet 50 by a margin of 2.4%. It verifies the effectiveness of the concatenation of multi-block features. Through detecting the open-set noisy samples in the coarse stage of our HCL approach, the classification accuracy improves from 81.2% to 85.2%. The reasons are analyzed as follows. In the coarse stage, the negative effect of the open-set noisy samples on the fine-grained classification training is first filtered out through the semantic consistency analysis between the image-level and object-level prediction distributions. Then, the noisy samples are utilized in image-object contrastive learning to enhance the model's recognition ability. When we further detect and utilize the closed-set noisy samples in the fine stage, the classification accuracy improves from 85.2% to 86.1%, which validates its effectiveness. The disruption caused by the closed-set noisy samples is removed, and the model's feature representation ability is enhanced through object-part contrastive learning, which mines and exploits local discriminative clues.
- We also conduct ablation studies about the distribution variance calculation in Eq. 9 and Eq. 13 on the three webly supervised fine-grained datasets. The fine-grained classification accuracy achieves consistent gains, such as improving from 85.4% to 86.1% on the Web-Bird dataset, which verifies its effectiveness. Through measuring the distribution variance and combing with the typical semantic label classification loss, the semantic consistency can be more comprehensively utilized for filtering out noisy samples accurately to obtain promising fine-grained recognition performance.

E. Parameter Experiments

We conduct parameter experiments about contrastive learning loss weight α, the number of warm-up epochs T_W, and the drop rate γ on the Web-Bird dataset. We first fix the γ as 0.25, T_W as 10 to change the contrastive learning weight α. Then, we fix the best α, set γ as 0.25 and change T_W. Finally, we fix the best α and T_W to change the drop rate γ. The experimental results are shown in Fig. 6, we can observe that:

- The proposed HCL approach can achieve the best recognition performance when α is set as 1, T_W is set as 10, and γ is set as 0.35. The fine-grained classification accuracy improves from 85.1% to 85.7% when raising α from 0 to 1, which verifies the effectiveness of contrastive learning in our HCL approach. We attribute the performance gain to the enhanced feature representation by contrastive learning.
- Slight performance fluctuation happens with higher or lower T_W values, which shows our proposed approach is relatively insensitive to the number of warm-up epochs. The drop rate γ affects the classification accuracy because it is an estimation of the ground-truth noise rate of the Web-Bird dataset, which is different for each webly supervised fine-grained dataset.

F. Experiments on Different Backbone Networks

To evaluate the performance of our proposed HCL method with different vision backbone networks, we conduct comparison experiments with ResNet50, ResNet101, and ResNet152 as the backbone network, respectively. Experimental results are shown in TABLE III. We can observe that:

- Compared with the pure ResNet50 network, our HCL approach with ResNet50 as the backbone network achieves

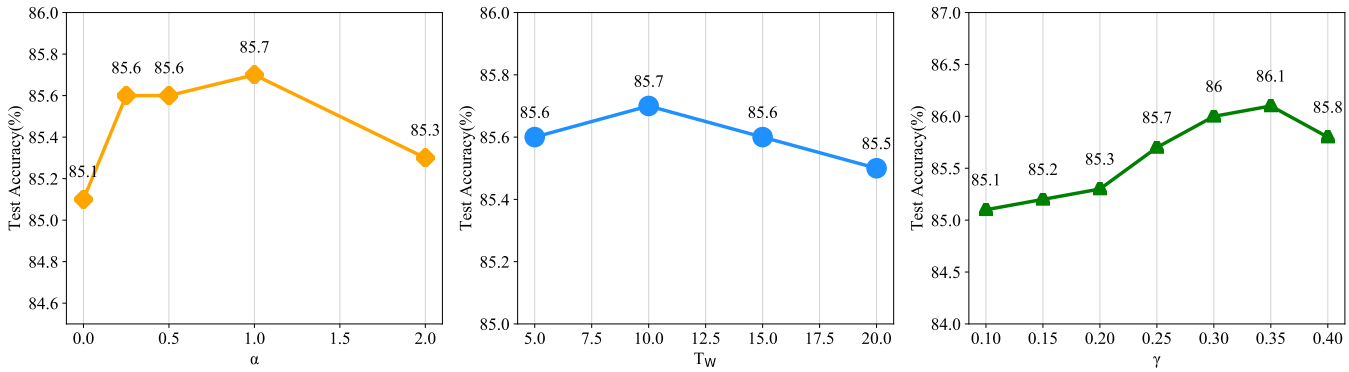

Fig. 6. Parameter experiments about contrastive learning loss weight α in Eq. 16 and Eq. 18, the number of warm-up epochs T_W and the drop rate γ in Section III-D on the Web-Bird dataset.

performance gains by 7.3%, 5.5%, and 4.6% on Web-Bird, Web-Car and Web-Aircraft, respectively. The same phenomenon can be observed when using other backbone networks. Thus, the effectiveness of our HCL approach on eliminating noisy samples and capturing multi-hierarchy discriminative clues for webly supervised fine-grained recognition is verified.

- By utilizing stronger vision backbone networks, our proposed HCL approach achieves better classification accuracy, which shows its extensibility and practicality to various CNN backbone networks. For example, our proposed HCL approach achieves 86.1%, 86.7%, and 87.2% classification accuracy on the Web-Bird dataset when we utilize ResNet50, ResNet101, ResNet152 as the backbone network, respectively. We attribute the performance gains to the more accurate salient regions extraction and more robust features brought by the stronger vision backbone, which contributes to highlighting visual object information and local discriminative information as well as eliminating noisy samples for fine-grained recognition. Overall, our proposed HCL approach can be effectively combined with various CNN backbone networks to capture multi-scale and multi-hierarchy information for webly supervised fine-grained recognition.

G. Visualization Experiments

We present the selected clean samples and detected noisy samples by the proposed HCL approach on three WSFG datasets in Fig. 7. The first four columns represent the selected clean samples, the fifth and sixth columns represent the detected open-set noisy samples, and the last column represents the detected closed-set noisy samples. Each row represents the sample selection on one specific subcategory. Samples in the red rectangles represent the wrongly selected clean samples. They comprise hard instances with subtle differences and animated model images. (1) For the subtle difference, the wrongly identified bird in the second row has a similar shape to the correct one and the blue background may confuse our proposed HCL model in color. The difference in the wing should be further mined by eliminating the background disruption to improve the recognition performance. (2) For the animated model, the wrongly identified car in the fourth row and the wrongly identified airplane in the sixth row are animated model images which are similar to the real instances. Our proposed HCL model is confused by the rare data format, which should be augmented for training. Overall, the proposed HCL approach can achieve promising performance in selecting clean samples and detecting noisy samples, which verifies its practicality in utilizing web data for fine-grained recognition.

H. Discussions

The limitations of our proposed HCL approach may rely on two aspects: (1) Time consuming: the coarse-to-fine information utilization manner of the proposed HCL approach takes a slightly longer training time. (2) Hard to recognize rare data format: the rare data, such as animated model images, confuses our HCL model to be misrecognized as clean samples. We will attempt the single-stage discriminative visual information utilization by part-attention design and data augmentation methods to improve the model's performance while alleviating the training time cost.

V. Conclusion

In this paper, we propose a hierarchical consistency learning (HCL) approach for webly supervised fine-grained recognition, which imitates the human recognition process to detect noisy samples and mine discriminative information simultaneously. In the fine-grained recognition scenario, there is inherent semantic consistency among the image, object, and part of the clean sample. The open-set noisy samples can be detected by analyzing the semantic consistency between the image-level and object-level prediction distributions while focusing on the visual object in the image. The closed-set noisy samples can be detected by calculating the semantic consistency between the object-level and part-level prediction distributions while mining local distinguishable clues. Multi-hierarchy discriminative information is thus utilized and enhanced for fine-grained recognition with multi-hierarchy contrastive learning. Extensive experiments on three public datasets present the superiority and practicality of the proposed HCL approach.

In the future, we will attempt to utilize the knowledge graph that describes visual object attribute information to help detect noisy samples, which is hopeful for improving the model's webly supervised fine-grained recognition performance.

Fig. 7. The sample selection results of the proposed HCL approach on the three webly supervised fine-grained datasets. The selected clean samples for training are on the left of the vertical dotted line. The detected noisy samples (including the first two columns of open-set noisy samples and the last column of closed-set noisy samples) are on the right of the vertical dotted line. Samples in the red rectangles are the wrongly selected clean samples.

REFERENCES

[1] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[2] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 554–561.

[3] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," *arXiv preprint arXiv:1306.5151*, 2013.

[4] X.-S. Wei, Y.-Z. Song, O. Mac Aodha, J. Wu, Y. Peng, J. Tang, J. Yang, and S. Belongie, "Fine-grained image analysis with deep learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 8927–8948, 2022.

[5] Z. Sun, Y. Yao, X.-S. Wei, Y. Zhang, F. Shen, J. Wu, J. Zhang, and H. T. Shen, "Webly supervised fine-grained recognition: Benchmark datasets and an approach," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 602–10 611.

[6] E. Malach and S. Shalev-Shwartz, "Decoupling" when to update" from" how to update"," *Advances in neural information processing systems*, vol. 30, 2017.

[7] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," *Advances in neural information processing systems*, vol. 31, 2018.

[8] H. Wei, L. Feng, X. Chen, and B. An, "Combating noisy labels by agreement: A joint training method with co-regularization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 726–13 735.

[9] L. Huang, C. Zhang, and H. Zhang, "Self-adaptive training: beyond empirical risk minimization," *Advances in neural information processing systems*, vol. 33, pp. 19 365–19 376, 2020.

[10] Y. Zhang, S. Zheng, P. Wu, M. Goswami, and C. Chen, "Learning with feature-dependent label noise: A progressive approach," *arXiv preprint arXiv:2103.07756*, 2021.

[11] Y. Yao, Z. Sun, C. Zhang, F. Shen, Q. Wu, J. Zhang, and Z. Tang, "Jo-src:

A contrastive approach for combating noisy labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5192–5201.

[12] Z. Sun, F. Shen, D. Huang, Q. Wang, X. Shu, Y. Yao, and J. Tang, "Pnp: Robust learning from noisy labels by probabilistic noise prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5311–5320.

[13] Z. Cai, G.-S. Xie, X. Huang, D. Huang, Y. Yao, and Z. Tang, "Robust learning from noisy web data for fine-grained recognition," *Pattern Recognition*, vol. 134, p. 109063, 2023.

[14] L. Niu, A. Veeraraghavan, and A. Sabharwal, "Webly supervised learning meets zero-shot learning: A hybrid approach for fine-grained classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7171–7180.

[15] H. Liu, H. Zhang, J. Lu, and Z. Tang, "Exploiting web images for fine-grained visual recognition via dynamic loss correction and global sample selection," *IEEE Transactions on Multimedia*, vol. 24, pp. 1105–1115, 2022.

[16] A. Ghosh, H. Kumar, and P. S. Sastry, "Robust loss functions under label noise for deep neural networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017.

[17] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *Advances in neural information processing systems*, vol. 31, 2018.

[18] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey, "Symmetric cross entropy for robust learning with noisy labels," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 322–330.

[19] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, and L. Qu, "Making deep neural networks robust to label noise: A loss correction approach," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1944–1952.

[20] X. Ma, H. Huang, Y. Wang, S. Romano, S. Erfani, and J. Bailey, "Normalized loss functions for deep learning with noisy labels," in *International conference on machine learning*. PMLR, 2020, pp. 6543–6553.

[21] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to reweight examples for robust deep learning," in *International conference on machine learning*. PMLR, 2018, pp. 4334–4343.

[22] D. Arpit, S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio *et al.*, "A closer look at memorization in deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 233–242.

[23] X. Yu, B. Han, J. Yao, G. Niu, I. Tsang, and M. Sugiyama, "How does disagreement help generalization against label corruption?" in *International Conference on Machine Learning*. PMLR, 2019, pp. 7164–7173.

[24] Y. Peng, X. He, and J. Zhao, "Object-part attention model for fine-grained image classification," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1487–1500, 2018.

[25] X. He, Y. Peng, and J. Zhao, "Fast fine-grained image classification via weakly supervised discriminative localization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 5, pp. 1394–1407, 2018.

[26] X. He, Y. Peng, and J. Zhao, "Which and how many regions to gaze: Focus discriminative regions for fine-grained visual categorization," *International Journal of Computer Vision*, vol. 127, no. 9, pp. 1235–1255, 2019.

[27] Z. Wang, S. Wang, H. Li, Z. Dou, and J. Li, "Graph-propagation based correlation learning for weakly supervised fine-grained image classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 289–12 296.

[28] K. Song, X.-S. Wei, X. Shu, R.-J. Song, and J. Lu, "Bi-modal progressive mask attention for fine-grained recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 7006–7018, 2020.

[29] J. He, J.-N. Chen, S. Liu, A. Kortylewski, C. Yang, Y. Bai, and C. Wang, "Transfg: A transformer architecture for fine-grained recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 852–860.

[30] Y. Hu, X. Jin, Y. Zhang, H. Hong, J. Zhang, Y. He, and H. Xue, "Rams-trans: Recurrent attention multi-scale transformer for fine-grained image recognition," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4239–4248.

[31] H. Sun, X. He, and Y. Peng, "Sim-trans: Structure information modeling transformer for fine-grained visual categorization," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 5853–5861.

[32] Q. Xu, J. Wang, B. Jiang, and B. Luo, "Fine-grained visual classification via internal ensemble learning transformer," *IEEE Transactions on Multimedia*, 2023, doi:10.1109/TMM.2023.3244340.

[33] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1449–1457.

[34] S. Kong and C. Fowlkes, "Low-rank bilinear pooling for fine-grained classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 365–374.

[35] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo, "Learning deep bilinear transformation for fine-grained image representation," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[36] M. Tan, F. Yuan, J. Yu, G. Wang, and X. Gu, "Fine-grained image classification via multi-scale selective hierarchical biquadratic pooling," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 18, no. 1s, pp. 1–23, 2022.

[37] R. Du, J. Xie, Z. Ma, D. Chang, Y.-Z. Song, and J. Guo, "Progressive learning of category-consistent multi-granularity features for fine-grained visual classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 9521–9535, 2022.

[38] D. Chang, K. Pang, Y. Zheng, Z. Ma, Y.-Z. Song, and J. Guo, "Your" flamingo" is my" bird": fine-grained, or not," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 476–11 485.

[39] J. Goldberger and E. Ben-Reuven, "Training deep neural-networks using a noise adaptation layer," in *International conference on learning representations*, 2017.

[40] J. Y. Zou, D. Hsu, D. C. Parkes, and R. P. Adams, "Contrastive learning using spectral methods," *Proceedings of Advances in Neural Information Processing Systems*, 2013.

[41] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," *arXiv preprint arXiv:1906.05849*, 2019.

[42] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[43] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[44] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[45] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[47] A. Recasens, P. Kellnhofer, S. Stent, W. Matusik, and A. Torralba, "Learning to zoom: a saliency-based sampling layer for neural networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 51–66.

[48] R. Du, D. Chang, A. K. Bhunia, J. Xie, Z. Ma, Y.-Z. Song, and J. Guo, "Fine-grained visual classification via progressive multi-granularity training of jigsaw patches," in *European Conference on Computer Vision*. Springer, 2020, pp. 153–168.

[49] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[50] K. Yi and J. Wu, "Probabilistic end-to-end noise correction for learning with noisy labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7017–7025.

[51] X. Peng, K. Wang, Z. Zeng, Q. Li, J. Yang, and Y. Qiao, "Suppressing mislabeled data via grouping and self-attention," in *European Conference on Computer Vision*. Springer, 2020, pp. 786–802.

[52] Z. Sun, H. Liu, Q. Wang, T. Zhou, Q. Wu, and Z. Tang, "Co-ldl: A co-training-based label distribution learning method for tackling label noise," *IEEE Transactions on Multimedia*, vol. 24, pp. 1093–1104, 2022.

[53] Z. Sun, Y. Yao, X.-S. Wei, F. Shen, J. Zhang, and X.-S. Hua, "Boosting robust learning via leveraging reusable samples in noisy web data," *IEEE Transactions on Multimedia*, vol. 25, pp. 3284–3295, 2023.

Hongbo Sun received the B.S. degree in electronic information engineering from Tianjin University, Tianjin, China, in 2016 and the M.S. degree in information and communication engineering from Tianjin University, Tianjin, China, in 2019. He is currently pursuing the Ph.D. degree in computer application technology with the Wangxuan Institute of Computer Technology, Peking University. His current research interests include fine-grained visual analysis and multi-modal content understanding.

Xiangteng He received the Ph.D. degree in computer application technology from Peking University, Beijing, China, in 2020. He is currently the Research Assistant Professor with the Wangxuan Institute of Computer Technology, Peking University. He has authored over 20 papers, including IJCV, IEEE TIP, IEEE TCSVT, CVPR, ICCV, ACM MM, ACM SIGIR, IJCAI and AAAI. His research interests include multi-modal content analysis, fine-grained visual analysis, image and video recognition and understanding, and computer vision. He was one of the recipients of 2020 CCF (China Computer Federation) Outstanding Doctoral Dissertation Award and 2018 Baidu Scholarship, and awarded Young Elite Scientists Sponsorship Program by CAST in 2022.

Yuxin Peng (Senior Member, IEEE) received the Ph.D. degree in computer applied technology from Peking University, Beijing, China, in 2003. He is currently the Boya Distinguished Professor with the Wangxuan Institute of Computer Technology, Peking University. He has authored over 200 papers, including more than 100 papers in the top-tier journals and conference proceedings. He has submitted 48 patent applications and been granted 39 of them. His current research interests mainly include cross-media analysis and reasoning, image and video recognition and understanding, and computer vision. He led his team to win the First Place in video semantic search evaluation of TRECVID ten times in the recent years. He won the First Prize of the Beijing Technological Invention Award in 2016 (ranking first) and the First Prize of the Scientific and Technological Progress Award of Chinese Institute of Electronics in 2020 (ranking first). He was a recipient of the National Science Fund for Distinguished Young Scholars of China in 2019, and the best paper award at MMM 2019 and NCIG 2018. He serves as the associate editor of IEEE TMM, TCSVT, etc.