Learning from Web Data: the Benefit of Unsupervised Object Localization

Xiaoxiao Sun, Liang Zheng, Yu-Kun Lai, Jufeng Yang

Abstract—Annotating a large number of training images is very time-consuming. In this background, this paper focuses on learning from easy-to-acquire web data and utilizes the learned model for fine-grained image classification in labeled datasets. Currently, the performance gain from training with web data is incremental, like a common saying “better than nothing, but not by much”. Conventionally, the community looks to correcting the noisy web labels to select informative samples. In this work, we first systematically study the built-in gap between the web and standard datasets, i.e., different data distributions between the two kinds of data. Then, in addition to using web labels, we present an unsupervised object localization method, which provides critical insights into the object density and scale in web images. Specifically, we design two constraints on web data to substantially reduce the difference of data distributions for the web and standard datasets. First, we present a method to control the scale, localization and number of objects in the detected region. Second, we propose to select the regions containing objects that are consistent with the web tag. Based on the two constraints, we are able to process web images to reduce the gap, and the processed web data is used to better assist the standard dataset to train CNNs. Experiments on several fine-grained image classification datasets confirm that our method performs favorably against the state-of-the-art methods.

Index Terms—Web Data, Standard Dataset, Built-in Gap, Unsupervised Object Localization

I. INTRODUCTION

D
eep Convolutional Neural Networks (CNNs) have achieved great success in solving recognition challenges including image classification [1], scene recognition [2], [3] and fine-grained image classification [4]–[6]. However, the commonly used CNN models such as AlexNet [1], VGGNet [7], GoogleNet-Inception [8] and ResNet [9] have a huge number of parameters, and their performance heavily depends on the availability of a large number of labeled training examples. In practice, getting reliably annotated images at a large scale is usually expensive and time-consuming, which prevents CNNs from being sufficiently trained for new image recognition tasks.

This paper considers the task of leveraging web data to improve recognition accuracy. The advantage of web data is primarily in its large-scale availability, such as millions of images with user-supplied tags from social websites or image search engines [4], [10]–[12]. Intuitively, more training data leads to higher classification ability of CNN models. However, there exists a saying when employing web data, which is “better than nothing, but not by much”. Existing works suggest that the labeling noise contained in webly-labeled images usually negatively influences the performance of CNN models. However, we believe the phenomenon that the content of web images is more complex than images from standard datasets is the crux of the problem. Fig. 1 shows images from Food-101 [13], Stanford Dogs [14], MIT Indoor67 [15] datasets and web data. The objects of Web Dog (WD-#) and Web Food (WF-#) images are selected by the boxes. Red boxes on images mean that the contained objects are consistent with the web labels, and yellow boxes denote that the objects are not consistent with the labels. For (c), the images are from the MIT Indoor 67 dataset (I-#) and the Web Indoor (WI-#) scene images.

![Fig. 1. Sample images of the standard and web datasets. The image pairs at each row of (a), (b) or (c) have the same label. The objects in Web Dog (WD-#) and Web Food (WF-#) images are selected by the boxes. Red boxes on images mean that the contained objects are consistent with the web labels, and yellow boxes denote that the objects are not consistent with the labels. For (c), the images are from the MIT Indoor 67 dataset (I-#) and the Web Indoor (WI-#) scene images.](image-url)

As discussed in [16], performance on the test set of a target task often decreases when the training data is augmented with data from other datasets. Torralba et al. [16], [17] argue that
the reason of this phenomenon is the input space of each image dataset is different, i.e., the datasets are biased. Regarding web data, we predict the label probabilities of sampled web images and test images from the Food-101 dataset using a benchmark model (fine-tuned on the Food-101 dataset with recognition accuracy of 84.31%). Fig. 2 shows the predicted label probabilities for two images from the web (red) and standard (green) datasets, which are unexpectedly significantly different. For example, the image baby-back-ribs from Food-101 has a salient rib closer to the center of the image. In contrast, the web baby-back-ribs image contains four dishes, but only the one in the top left contains ribs. Hence, the localization and scale of the labeled content, as well as the number of relevant objects result in the difference between the data from the two sources. To further illustrate that bias exists between web and standard datasets generally rather than occasionally, we conduct rigorous experiments to quantitatively analyze the difference between standard and web datasets based on 5 criteria: relative data bias, cross-dataset generalization [16], [17], the scale of related content, density of domain information, and label quality (see Section IV-C).

In this paper, we design abundant experiments to evaluate the web and standard datasets, and experimental results (in Section IV-C) confirm our conjecture that the complexity of web data results in the bias between the web and the standard datasets. Furthermore, we address the problem of learning from web images by reducing dataset bias in two aspects: (1) normality of content/information distribution on images (referred to as Form bias, i.e., whether the location, size and number of classification objects/content on web data are similar to these conditions on standard data). (2) label consistency (referred to as Label bias, i.e. whether the label is accurate to describe the web image). Specifically, we present an unsupervised object localization method to generate content regions related to the task on web images. Considering standard datasets without bounding box (bbox) annotations, we employ the “labeled-content-centric” prior to generate the bboxes, and use the bboxes to train a region proposal network (RPN). Then, we generate object proposals of web images and add two constraints for the regions: (1) a form constraint to control the scale and density of objects based on the bbox of region proposal, and (2) a label constraint to control the consistency of subject and the label based on the score from RPN and the probability from the benchmark recognition model. Finally, the eligible regions are collected as the de-biased web dataset. The evaluation results on the Food-101, Stanford Dogs (Dog-120) [14] and MIT Indoor 67 [15] datasets, as well as images retrieved from the web show that the designed framework can substantially reduce bias. Meanwhile, extensive experimental results show that reducing bias effectively improves the efficiency of using web data.

The contributions of this paper are three-fold.

1) We systematically study the built-in gap between web and standard datasets, and perform quantitative analysis on dataset bias from various aspects. As far as we know, there is no existing work that processes web data from this perspective.

2) We propose an unsupervised object localization method and design two constraints based on the built-in gap between web and standard datasets to change the scale and density consistency of web images to approximate standard data.

3) We perform experimental evaluation on three publicly available datasets, i.e., Stanford Dogs, Food-101, MIT67. The experiments demonstrate that the proposed framework is able to improve the performance of learning from web images.

II. RELATED WORK

A. Deep Learning from Web Data

Leveraging web data [18]–[22] has been an alternative method to satisfy the data requirements of deep models. Previous work explores utilizing web data by reducing the label noise level or designing a noise-aware model [4], [10], [12], [23]. Different from these works, we propose to reduce dataset bias when employing web data.

The ever-growing web images have been an important data source for computer vision tasks. However, label noise is ubiquitous in real-world data, which influences the performance of the model. Efforts have been made to improve results when using web data [24]–[28]. Recent works [29] demonstrate that web data can be used to improve the performance of deep learning. To readily scale to ImageNet-sized problems, Izadinia et al. [30] present a method to directly learn from image tags in the wild and show that web data is useful for training CNN models. Chen and Gupta [31] propose a two-stage approach to training CNNs by exploiting noisy web data and the transferability of CNNs. This work assumes that label noise is conditionally independent of the image, and is directly affected by the noisy annotations produced by humans. Xiao et al. [10] use a probabilistic framework to handle noisy labels and train a classifier in an end-to-end learning procedure. Furthermore, Krause et al. [4] demonstrate that using massive web data for training CNN models can improve the performance of deep models on fine-grained datasets.

Joulin et al. [11] argue that ConvNets can learn from scratch in a weakly-supervised way, by utilizing 100M Flickr images annotated with noisy captions.

However, there exists a phenomenon when employing web data for training CNN models, which is “better than nothing,
but not by much”. In previous work, there is no significant improvement for the final classification accuracy, except for the cases which use over 10-100 times more images than the standard datasets [4]. Different from above work, a new perspective is considered in this work and systematic experiments show that dataset bias between web and standard datasets of the target task is the main reason for the limited benefit of web data. Therefore, we explore the potential of web data for training CNN models by reducing bias.

B. Built-in Gap between Datasets

With the recent development of deep learning, transfer learning has been successfully applied to object recognition problems. Based on the observation of transferability of deep neural networks, the works [1], [32] propose to conquer this problem by first initializing CNN parameters with a model pre-trained on a larger yet related dataset, and then fine-tuning it on the smaller dataset of the target task. However, there is a problem that the data distribution of the source task often fails to match the target task exactly (and vice versa). We define the mismatch as dataset bias, which has been researched in [16], [33]. These works prove that bias exists in visual datasets, and [34] further shows that the contents of images are often biased with human-centric labels. Moreover, [35] transfers human biases into machine systems to help with object recognition.

We believe the bottleneck of employing web data for training CNN models is the bias between web and standard datasets, so we present a method to reduce the bias for improving the performance of using web data. In [16], biases are summarized as selection bias, capture bias, category or label bias and negative set bias. [34] reveals the bias of human-centric annotations with images as keyword reporting bias. [36] considers the scale bias of object and scene datasets. Based on existing summarized biases, we systematically study dataset bias between webly-labeled data and standard data. Meanwhile, we present an unsupervised object localization method to reduce the bias, which can improve the effectiveness of using web data.

C. Object Localization Methods

1) Traditional Object Detection Methods: The sliding-window method has been used for a long time in which a classifier is applied on a dense image grid. For this kind of methods [37], [38], LeCun et al. apply convolutional neural networks to handwritten digit recognition and improve the final results. Viola and Jones [39] use improved object detectors to detect faces, leading to widespread adoption of these models. The introduction of Deformable Part Models (DPMs) [40] helps extend dense detectors to more general object categories and get best results on the PASCAL dataset [41]. However, with the development of deep learning [1], new detectors based on deep models quickly become the main approach for object detection.

2) Deep Object Detection Methods: The commonly used framework in modern object detection is a two-stage approach. Based on selective search work [42], the first stage will generate a set of candidate object proposals that contain all objects while filtering out the majority of negative locations. Then, the second stage classifies the proposals into foreground classes and background. On the other hand, some recent research focuses on developing one-stage methods, such as OverFeat [43], which is one of the first end-to-end object detectors based on deep networks, as well as SSD [44] and YOLO [45]. These methods can improve the processing speed but their accuracy is not good compared with two-stage methods, such as R-CNN [46]. Different from these works, we emphasize that our simple detector achieves top results not based on innovations in network design but due to our novel strategies.

3) Unsupervised Object Detection Methods: Unsupervised object discovery (also called image co-localization) [47], [48] is a fundamental problem in computer vision, where it needs to discover the common object emerging in only positive sets of example images (without any negative examples or further supervisions). Recently, there also appear several co-localization methods based on pre-trained deep convolutional models e.g., Li et al. [49]. These methods just treat pre-trained models as simple feature extractors to extract the fully connected representations, which do not sufficiently mine the information beneath the convolutional layers. Zhang et al. [18] propose a two-level attention framework for dealing with wepy-supervised classification, which achieves a new state-of-the-art. Our method handles web data based on RPN, which can (1) recognize noisy images and also (2) supply bounding boxes of objects.

III. Method

We aim to eliminate the dataset bias between the web and standard datasets, and then utilize such data to train a CNN model for classification. The causes of built-in bias are summarized as two aspects in this work: Form bias and Label bias (Section IV-C2). In our method, we mainly address the form and label biases under the following assumptions: (1) web data may contain different distributions, such as size, number and localization of classification objects/contents from the standard dataset (e.g. classification objects/contents for the dog recognition task refer to dogs appearing in the images); (2) labels of web images may be inconsistent with the contents of images. Specifically, We have a web dataset \( D_w = \{(x(1), y^{(1)}), \ldots, (x(N), y^{(N)})\} \), where the n-th image \( x^{(n)} \) has label \( y^{(n)} \) \( \in \{1, \ldots, N\} \), \( N \) is the number of images and \( L \) is the number of classes. Meanwhile, we have a standard dataset \( D_s \). Our target is to transform web dataset \( D_w \) to a new dataset \( D_{w^*} \), which is de-biased with respect to \( D_s \).

A. Framework of the Presented Method

First, we capture objects related to the target task to normalize the scale, density and location of objects, which are important measures of form bias. Specifically, we train a detection model on the standard dataset \( D_s \) to detect target objects in web images, and then learn the level of relevance of detected regions with the tags of web images by modeling \( p(y|R^x_i) \), \( i = 1, 2, \ldots, K \) where \( x \) is a web image with label \( y \) and \( \{R^x_i\}_{i=1}^K \) refers to the set of detected regions of \( x \). \( R^x_i \)
is one of the regions, \( K \) is the number of regions. Second, we have the following two considerations for tags of web images: (1) whether target objects appear in the web image (e.g., a web image used for dog classification contains dogs) and (2) whether the label is consistent with the object present in the web image (e.g., the breed of dog in a web image is the same as the label of this image). To eliminate label bias, we couple the two considerations with the detected object proposals of web images, so that we can select object proposals of web images accurately.

Fig. 3 shows the framework of our method, in which the web dataset passes through the region proposal network and the constraints, and then becomes \( D_w \cdot \cdot \cdot \) (a) Training of region proposal networks has two-levels of parameters to control the size of bboxes corresponding to the red (region proposal) and blue (object detection) rectangles on the images. During generation of the proposals on the web dataset \( D_w \) (see (b) Prediction of RPNs), there is a two-stage selection based on the above two RPNs trained on bboxes of two sizes. Meanwhile, the basic model will predict the probability \( p(y | R^w_i) \) of proposals, which can be used to remove noisy regions. Finally, we take the output \( D_w \cdot \cdot \cdot \) and \( D_s \) as the training set to train the classification CNN model. The RPN and constraints are introduced as follows.

### B. Unsupervised Region Proposal Networks

As discussed above, candidate object proposals of web images are not only crucial for reducing form bias but also helpful to select regions consisting of relevant tags from web images. However, since there are no ground-truth annotations to locate object regions, we present a new unsupervised object localization method through joint region proposal and feature extraction via CNN layers without needing bounding box annotations. Given a standard dataset \( D_s = \{ (x^{(m)}, y^{(m)}) \}_{m=1}^{M} \), where \( M \) is the number of images, based on the assumption that standard images typically have one object in the center, we define the weak bounding box of an image \( x^{(m)} \) as follows:

\[
\left( t_m(x), t_m(y) \right) = \left( \frac{1 - \lambda}{2} w_m, \frac{1 - \lambda}{2} h_m \right),
\]

where \( w_m \) and \( h_m \) are the width and height of \( x^{(m)} \), and \( w_m^{(b)} \) and \( h_m^{(b)} \) are the width and height of its weak bounding box, which are defined as \( w_m^{(b)} = \lambda w_m \) and \( h_m^{(b)} = \lambda h_m \). \( t_m(x), t_m(y) \) are the coordinates of the top-left corner of the bounding box, and \( \lambda \) is the proportion used to set the size of bounding box.

We use the standard dataset and the weak bounding boxes defined above to train a region proposal network (RPN) where \( \lambda \) is set to \( \lambda_r \) for region proposal [50]. Fig. 3 shows the framework of the unsupervised region proposal network. Applying this RPN to web images generates region proposals. Given an image \( x \), the RPN will output a set of rectangular object proposals \( R_x = \{ R_1^x, R_2^x, \cdots, R_K^x \} \), each with an objectness score \( p(z | R_i^x) \) that estimates the probability of \( R_i^x \) containing a domain object. The regions produced by the RPN are not accurate enough because of the imprecise bounding
boxes used for training. This is however sufficient, as our goal is to generate candidate regions containing objects from web images rather than locating objects precisely.

To further select appropriate proposals, we use the size ratio \( \lambda \) for training the object detection RPN, which is smaller than \( \lambda_c \) used to train the RPN for generating region proposals (as shown in (a) Training of RPNs in Fig. 3). We train two RPNs by using two different sizes of bboxes for generation of content (red border) and object (blue border) region proposals respectively.

So far we have described how to get a set of candidate proposals which almost contain objects located close to the center. This makes the scale of related content and the density of domain information similar to the standard dataset. During the training of RPNs, one mini-batch is formed by 256 proposals and the loss is calculated based on classification loss and regression loss, which ensure the proposal contains the object, and its location and size are close to the ground truth.

C. Form and Label Constraints

The detection network maps the whole image to domain regions with associated probabilities, but the category of the object in the image has not yet been considered. Therefore, label bias still exists. We handle it with convolutional layers followed by a softmax layer. The convolutional features for detection and classification are shared, so that the classification network can be easily trained.

Given an image \( x \), let \( R_x^* = \{ R_1^x, R_2^x, \ldots, R_K^x \} \) be the locations of \( K \) proposal regions in the image. Note that we have adopted non-maximum suppression (NMS) [50] on all regions to reduce redundancy. In order to select the distinguishing regions \( R_{sub}^x \subset R^x \) from the proposals, we consider combining two score constraints by solving the following optimization problem:

\[
R_{sub}^x = \{ R_i^x \mid \Delta(R_i^x) = 1, i = 1, \ldots, K \},
\]

where \( \Delta(R_i^x) \) is defined as a scoring function over two constraints as follows:

\[
\Delta(R_i^x) = \Delta_{\text{form}}(R_i^x) \Delta_{\text{label}}(R_i^x),
\]

in which \( \Delta_{\text{form}}(R_i^x) \) denotes the form constraint and \( \Delta_{\text{label}}(R_i^x) \) denotes the label constraint.

Form constraint. The form constraint controls the scale of the regions. We want to constrain the scale of objects such that they are not “too small” compared to the size of the regions. For identifying the object in a region, as described before, in addition to training an RPN with size ratio \( \lambda \), for regions containing relevant content, we also train another RPN with size ratio \( \lambda_s < \lambda \) for object detection. We then send each region proposal generated by the RPN model with \( \lambda_r \) into the object detection model (trained with size ratio \( \lambda_s \) ) and retain the proposal with maximum object score as the object proposal. Hence, we define the constraint as:

\[
\Delta_{\text{form}}(R_i^x) = \begin{cases} 1 & \text{IoU}(R_i^x) \geq \eta \\ 0 & \text{otherwise} \end{cases}
\]

where \( \text{IoU} \) means the Intersection-over-Union overlap between the object proposal and the region. \( \eta \) is the threshold value of IoU. The proposal objects are generalized by the detection model trained on the standard dataset.

Label constraint. The label of web images may contain errors, so we select proposals by adding a label constraint:

\[
\Delta_{\text{label}}(R_i^x) = \begin{cases} 1 & \tilde{y} = \hat{y}, p(x|R_i^x) \geq \epsilon \\ 0 & \text{otherwise} \end{cases}
\]

where \( p(y|R_i^x) \) is the probability of region \( R_i^x \) belonging to the category \( j \), \( j = 1, \ldots, L \), which is predicted by the benchmark model (fine-tuned on the standard dataset). \( \epsilon \) is the threshold value to control the score of the detected objects. \( \hat{y} \) is the predicted label of the image region. \( p(x|R_i^x) \) is the probability that \( R_i^x \) contains the classification content. The predicted label of the region is required to be the same as the web label \( \tilde{y} \).

Finally, for each web image, we obtain \( R_{sub}^x \), which is de-biased with the standard data. That is, by combining scale and label information, we map \( D_w \) to a new processed dataset \( D_{ws} = \{ R_{sub} \} \), which is used to assist the target task.

D. Training Process

We use \( D_{ws} = \{ R_{sub} \} \) as the auxiliary data and input it along with \( D_s \) into the CNN model. The parameters \( \theta \) of the model \( M \) are initialized on parameters \( \theta_{\text{pre}} \) of the pre-trained model \( M_{\text{pre}} \) (trained on ImageNet), and then the parameters are updated using stochastic gradient descent (SGD). The iteration of SGD updates current parameters \( \theta^* \) as:

\[
\theta^* = \theta + \gamma \cdot \frac{1}{|D^b|} \sum_{x,y \in D^b} \nabla_{\theta} [l(x,y)],
\]

where \( l(x,y) \) is a loss function, e.g., softmax. \( \nabla_{\theta} \) is computed by gradient back-propagation. \( D^b \) is a mini-batch randomly
drawn from the mixed training dataset \( D \), and \( \hat{\gamma} \) is the learning rate, which will reduce during training. The method is summarized in Algorithm 1.

IV. EXPERIMENTS

A. Experimental Setup

**Dataset.** We employ three standard datasets (where “standard” refers to public datasets of the target task), namely Food-101, Stanford Dogs and MIT Indoor67 to evaluate our method. We use these datasets for two reasons: 1) The diverse characteristics of both web-derived and well-labeled datasets, such as the number of instances in each image, object size and location, result in a significant gap which is under-researched in the literature. 2) The representative capacity to various classification tasks, i.e., foreground-oriented shape-based task (dog), foreground-oriented non-shape task (food) and hybrid content based task (indoor scene), can evaluate the robustness of the proposed method.

For each task, the web datasets are collected by keyword search from Google Images, Flickr and Twitter [51]. After downloading web images, the images that are near duplication with any image in the test set are removed following [52], which ensures the fairness of our tests. Finally, 214,743 images of food, 11,201 images of dog and 76,907 indoor scene web images are used for experiments based on the dataset of work [51]. For each source, we use a similar amount of data as the well-labeled images for the food and dog datasets. For indoor scenes, we found the downloaded web data are more complex (e.g., the contents of web images are richer and there are more similar images) than the other two tasks, so we use more web images (about 2 times compared to standard data) from each source of this task.

**Implementation.** The major model of the experiments is the ResNet [9], which has good performance for many classification tasks. For generating object proposals, we use our data to train the architecture used in [53]. For classification, the first step of our experiments is to fine-tune CNN models on the target datasets as the basic models. We use Caffe [54] in our experiments, and our models are trained on NVIDIA TITAN X GPUs. We use different experimental settings on the three datasets because the scales of datasets are different. The learning rate is initially set to 0.001, which is divided by 10 after 10 epochs. The total number of iterations is 30 epochs with a mini-batch including 20 224x224 images on Titan X (for ResNet110, the batch-size is set to 13 for Titan X). Since our goal is to obtain content regions related to the label rather than locate objects precisely, the settings of \( \lambda_r \) and \( \lambda_o \) for region detection and object detection do not need to be precise, as discussed in Section IV-B. The detailed discussion about \( \eta \) in Eq. 4 and \( \epsilon \) in Eq. 5 are also shown in Section IV-B. None of these parameters will seriously affect the results unless they are set to very small or very big values.

B. Parameters

In this section, we discuss the parameter settings of Eq. 1, Eq. 4 and Eq. 5. Here, we use “Name That Dataset” to evaluate how the parameters affect the gap between Food-101 and the processed data. First, there are two \( \lambda \) values for region detection and object detection, respectively. Fig. 4 shows the classification precision of “Name That Dataset” on Food-101 and Food-debiased when setting \( \lambda_r \) and \( \lambda_o \) to different values. As can be seen, \( \lambda_r = 0.9 \) is the best setting for region detection, because the minimum of binary classification results is at \( \lambda_r = 0.9 \), which means the gap is smaller than other settings. To make the experiment manageable, we fixed \( \lambda_r \) for region detection to 0.9 when discussing \( \lambda_o \) for object detection, and \( \lambda_o = 0.8 \) gets minimum results for classifying the two datasets. Second, Fig. 5 shows the influence of setting the IoU threshold \( \eta \) in Eq. 4 for generating object proposals and the object score threshold \( \epsilon \) in Eq. 5.

The classification precision of “Name That Dataset” on Food-101 and Food-debiased with different parameter settings are reported. It can be seen that none of these parameters affect the results significantly unless they are set to very small or very big values. The noise level affects the number of images fed into the CNN model selected by our label constraint module. For some images, the proposed method can retain more than one region proposal for one image. During the training of RPN, one mini-batch is formed by 256 proposals and the loss is calculated based on classification loss and regression loss, which ensure the proposal contains the object, and its location...
Fig. 6. CNN plays “Name That Dataset”, where the task is to separate images from standard datasets and web datasets as specified by each curve. The ‘-debiased’ means to classify standard data and data obtained using our method. “Ratio of used training set” is the proportion of training dataset actually used for training.

and size are close to the ground truth.

C. Evaluation of Built-in Gap Between Datasets

To analyze the change of the bias between web and standard datasets, we will measure dataset bias from five aspects: (1) Relative data bias: following [16], [17] we define the relative data bias as the uniformity between web and standard data, i.e., if they are misaligned. (2) Cross-dataset generalization: similar to the definition proposed in [16], [17], we use it to evaluate the generalization of different types of data, especially web data. (3) Scale of related content: it is designed based on [36], and originally means the scale of objects in an image, but in this paper, it represents the scale of the object relevant to the web label. (4) Density of domain information: it originally indicates the density of objects in a scene image, but here we use this to represent the number of classification objects occurring in a web image. (5) Label quality: it is determined by the noisy label of web data and represents the relationship between web label and web image content. Specifically, we measure the dataset bias as follows:

1) Measuring Dataset Bias: In this section, we will report the quantitative evaluation to demonstrate the existence of bias.

First, we employ the game “Name That Dataset” used in [16] on three datasets respectively as the measure of relative data bias. As shown in Fig. 6, for each dataset (Food-101, Dog-120 and Indoor-67), we create a mixed dataset (indicated by suffix -mix) by combining the same number of training images from the training dataset and web images, and we gradually increase the number of dataset images from 1/5 to the full set. This is used to train a CNN model for binary classification of whether the image is from the standard dataset or the web. With increasing training data, the web and standard images are easily classified by the model. For comparison, we also conduct the experiment on each standard dataset itself by labeling half of the images as 1 and the rest as 2, and the classification accuracy is stable at around 50%. The line chart illustrates that web and standard datasets have strong relative data bias and are not aligned, i.e., either of them is unique and identifiable from the other.

Second, the cross-dataset generalization experiments on web and standard datasets are designed following [16]. Table I shows the classification performance on food, dog and indoor scene datasets and the corresponding web data. Note that the same number of web images as the standard datasets are used in this experiment. Results in each row are trained using the specified dataset, and the three columns show accuracies when testing on the standard and web datasets as well as the mean of the two accuracies, e.g., in the first row, 84.31% accuracy is obtained by training and testing on Standard (Food-101), and 52.49% is obtained by training on Standard (Food-101) and testing on Web (Food). Furthermore, to prove that web dataset bias is universal rather than caused by the collection method used in our work, we conduct the same experiment using web data from L-Dog collected by Krause et al. [4]. In this experiment, we use 21,827 images (13,158 for training, 8,669 for testing) from the L-Dog dataset with the same classes as Stanford Dogs for a fair comparison.

As discussed in [16], the actual accuracies are not important, but the differences in performance are worth discussing. The performance when trained on one source and tested on the other source is substantially worse than training and testing on the same source (bold), so the two sources are not generalizable to each other. Meanwhile, the mean ± difference shows that web data is more generalizable than standard data, which conforms to the fact that standard datasets have more restrictions than web data.

Third, we further perform two experiments to measure the scale of related content and density of domain information. The scale of related content measures the size of the detected region containing the relevant content compared to the size of the entire image, and the density of domain information is the number of detected regions related to the subject in the image. The scale is defined as

$$scale = \frac{1}{d} \sum_{i=1}^{d} \frac{|I_i|}{|I|},$$  \hspace{1cm}(8)

where $d$ is the number of detected regions related to the
Dog Food Indoor

Fig. 7. Examples of the working mechanism of unsupervised object localization. We show examples of web images and the corresponding de-biased images. The web images contain the form bias and label bias, and there are many images with a density greater than one and our method handles them well.

subject of the domain task in the image, so the density is \(d \cdot |I_i|; i = 1 \ldots d\) is the size of the i-th region containing relevant objects, and \(|I|\) is the size of the image. In Table II, we report the average scales and densities of images on Food-101 and Stanford Dogs datasets. We can see that the densities of standard images are almost always 1, but they are getting bigger on web images. Meanwhile, the scale of related content is usually smaller in web images than in standard images. For example, Food-101 has \(\text{density} = 1.16\) and \(\text{scale} = 0.8536\) whereas Food-web has \(\text{density} = 1.94\) and \(\text{scale} = 0.6218\), which are distinguished from each other.

Different from object classification, indoor scene recognition requires knowledge about both scenes and objects. Calculating the scale and density of a scene image is not reasonable, but we can take a web scene image as a combination of several parts (some contain useful information and some are redundant information e.g., people). Therefore, we directly use the same process on scene data as for food and dog images without measuring the changes of these metrics.

Finally, for the label quality, we use the region proposal model (see Section III-B) to generate content regions for web images and find about 31.3% of webly-labeled food images have no detected regions. The result shows that about 1/3 of webly-labeled food images are outliers.

Overall, the above results demonstrate the existence of dataset bias between web and standard datasets.

2) Estimating the Culprits of Built-in Gap: We now investigate the causes of built-in gap between web and standard datasets.

Forming of datasets. It is a more generic concept of “capture bias” \cite{16} relating to focal length, finder frame and view etc. during shooting — that objects in images from the standard dataset are almost always in the center on the image, but objects in web images are not always single-object-in-the-center. Moreover, as shown in Fig. 7, the contents of web images are rich, not just containing one major object with a big scale. These factors result in the form bias.

Label noise level of datasets. It comes from the fact that images of a domain task (fine-grained classification) are often difficult to label for people because of the large intra-class and small inter-class variations among objects. This is particularly problematic for web data, which are usually labeled by non-experts without moderation. Image labels may be inaccurate or even entirely wrong, i.e., irrelevant content in the image. These “wrong” labels also contribute to the bias. The three kinds of biases are with respect to the information distribution, content and label of web images. For form and label bias, we show the results of reducing the influence of them as follows.

D. Performance of Reducing Bias

The effects of the proposed method on reducing bias between web and standard datasets are shown as follows.

In order to show the change of dataset bias quantitatively, we employ the game “Name That Dataset” again. We process the web data by our method, and select the same number of images from different sources to do the experiments. The results are shown in Fig. 6, after debiasing, the relative dataset bias becomes much lower (“debiased” datasets), comparable to the results on two halves of standard data (around 50%). Furthermore, with more processed training data (-debiased) added, the classification accuracies keep leveling at around the chance 50%. On the contrary, the classification accuracies of “data-web” tend to rise with more training data. Because the images after processing will be similar to the images from the standard dataset, the bias between web and standard datasets is reduced. Fig. 7 shows some examples of the images before and after processing, in which the scale and density of objects in the images are changed. Moreover, the noisy images are removed during processing. As shown in Fig. 8, the cross-dataset generalization results show that de-biased web
datasets become significantly more generalizable: training on standard datasets and testing on de-biased web data and vice versa both show significantly better performance than with original web data.

For the scale of related content and density of domain information, we report the measures before and after bias reduction using our method in Table II. Our method simultaneously removes different biases: the label constraint in our method can remove noisy web images and the form constraint controls the scale of the related object, and the density of domain information becomes more consistent with the target dataset. 95,672 (214,743), 68,355 (112,018) and 37,496 (76,970) are the numbers of images after (before) elimination for the three tasks. The numbers of proposals are 286,180, 167,925 and 153,772 for web food, dog and indoor scenes, respectively. After processing, the numbers of the regions are 121,327, 80,658 and 72,148.

To analyze the effect of using web data, we evaluate the performance of models trained on images from the following sources: the Food-101 dataset [13], web food images (Food-web), as well as web images after manual filtering (Human-Filter) and processed by Pseudo-label method [55]. The test set is the standard set of Food-101. As shown in Fig. 9, the improvements of performance are disparate when adding the same amount of standard data (red border) and webly-labeled data (green border). This may be due to the fact that web data contains noisy labels.

However, the filtered web data (blue and black) has a much lower noise level, yet the performance is still significantly worse than the standard data (red) of Food-101. We also use mixed data (Mix) of web and Food-101 as well as filtered web data (Mix-filter) to train CNNs. The number of mixed data is double of Food-101 and Food-web for the same x coordinate in the graph, but there is no significant improvement on the results after filtering the web data. The reason is that the web data and filtered data have poor cross-dataset generalization. On the contrary, the results of using the processed data are similar to the results of using the standard data (red). Furthermore, mixing the processed data with the standard data (Mix-Pro-Data) gets improved results than all the other methods. For the three tasks, the proposed method achieves consistently increasing performance with more processed images and the results are shown in Table III.

### Table II

| Dataset | Scale | Density |
|---------|-------|---------|
| Food-101 | 0.8536 | 1.16 |
| Food-web | 0.6218 | 1.94 |
| Food-debiased | 0.7775 | 1.23 |
| Dog-120 | 0.7381 | 1.00 |
| Dog-web | 0.5939 | 1.45 |
| Dog-debiased | 0.7465 | 1.18 |

Fig. 8. Cross-dataset generalization. Green, Red and Blue mean Standard, Web and Processed (debiased) web images respectively. The values are the recognition accuracy of the models that are trained on Standard (S), Web (W) and Processed (P) images respectively, and tested on Standard (S), Web (W) and Processed (P) data (columns). We also show some example images for each kind of data. For each column, the image with green border comes from standard dataset, and the two images in second and third rows are a pair, which are one web image (red) and the corresponding image after processing (blue).

Fig. 9. Performance comparison demonstrating the gain of adding web data and processed data. “Mix” means the web and standard data are added together, so the number of images is twice of only using Food-101 or Food-web.
E. Training CNN Models for Classification

We evaluate our method on three datasets. These datasets contain specific forms of objects (dogs), objects with irregular shapes (food), and even scenes which are of complex forms. The detailed results are as follows.

**Food-101.** It is a dataset with 1,000 images of each class. The training and testing are split into 3:1. The number of web images we used is 214,743. First, Table IV shows the comparisons with other works utilizing web images or weakly learning from web labels. To compare with these methods, we conduct experiments based on different models used in these works on the Food-101 dataset, e.g., AlexNet and VGGNet etc. “Clean-ft Accuracy” is the result of training the deep model on standard dataset and “Mix-ft Accuracy” is the performance of the model trained on web and standard datasets together.

Combining such methods with our bias removal significantly improves the performance of classification. “Method Accuracy” is the result of using the method in the first column for learning from web food images. “Our Accuracy” indicates the result of our proposed method. Compared with other methods using web data, the proposed method can further improve the performance. Moreover, no matter which deep model (AlexNet, CaffeNet, VGGNet, ResNet) is employed, the proposed algorithm shows its superiority consistently.

Furthermore, Table V presents the results on the Food-101 dataset of some detection methods, and the baseline as well as the ablation experiment results on ResNet50 and ResNet110 with web data. As can be seen, the result of ResNet-50+(web+filter)+ft is similar to ResNet-50+(web+RPN)+ft, because the regions generated by the RPN without the constraints still contain many noisy and redundant regions, which will influence the final results. Adding Δ_form helps remove redundant regions, so the result is improved by about 2%. Meanwhile, introducing Δ_label further filters the noisy regions to improve the performance. The ablation experiment results illustrate that each part of our method is efficient. Our result is 6.25% higher than the performance of directly using web data on ResNet-110. Meanwhile, we also compare with some existing unsupervised methods, such as [49] and [61]. Since the proposed method considers the characteristics of web images and adds constraints on the object proposals, the final generated regions are more similar to standard data than other localization methods.

**Stanford Dogs.** We also evaluate our method on Stanford Dogs to prove its robustness. Stanford Dogs contains 120 dog categories, with 12,000 images for training and 8,580 images for testing. The number of web images is 112,018 in this work. Table V shows the results of fine-tuning the ResNet on Stanford Dogs without web data achieves an accuracy of 81.26%. After adding web data without any pre-processing, the accuracy is only improved by 1.46%. However, by employing de-biased web data, the performance of the model improves substantially by 6.87%. For the dog dataset, the first round filtering removes almost 40% web images, which contain many useful images with abundant domain

| Method | Model | Clean-ft Accuracy (%) | Mix-ft Accuracy (%) | Method Accuracy (%) | Our Accuracy (%) |
|--------|-------|-----------------------|---------------------|--------------------|-----------------|
| Bottom-up [56] | AlexNet | 66.49 | 68.72 | 70.29 (+1.57) | 75.31 (+6.59) |
| Pseudo-label [55] | | | | 69.36 (+0.64) | |
| Weakly [11] | | | | 71.10 (+4.61) | |
| Boosting [23] | CaffeNet | 67.43 | 69.00 | 72.53 (+3.53) | 75.58 (+6.58) |
| PGM [10] | | | | 73.14 (+4.14) | |
| WSL [31] | | | | 73.21 (+4.21) | |
| Harnessing [24] | VGGNet | 74.32 | 76.98 | 79.02 (+2.04) | 81.93 (+4.95) |
| Goldfince [4] | ResNet50 | 84.31 | 84.89 | 86.75 (+2.86) | 90.41 (+5.42) |
| DPTL+subA [5] | | | | 88.78 (+3.89) | |

**TABLE IV**

**EXPERIMENTAL RESULTS ON THE Food-101 DATASET.** "CLEAN-Ft" MEANS FINE-TUNING ON THE MODEL PRE-TRAINED ON ImageNet WITH FOOD-101, AND "MIX-Ft" MEANS FINE-TUNING WITH BOTH STANDARD AND WEB DATA. METHOD ACCURACY AND OUR ACCURACY ARE ACCURACIES WITH THE METHOD ITSELF, AND THE PROPOSED METHOD. NUMBERS IN THE BRACKETS SHOW PERFORMANCE GAIN COMPARED WITH "MIX-Ft" AS THE BASELINE.

| Method | Model | Acc (%) | Food | Dog | Indoor |
|--------|-------|---------|------|-----|--------|
| Grad-CAM [57] | | 82.94 | 84.92 | 78.55 |
| Cascade R-CNN [58] | | 82.45 | 82.04 | 78.66 |
| RefineDet [59] | | 83.20 | 82.68 | 79.91 |
| DSOD [60] | | 84.75 | 81.93 | 78.51 |
| Cross-Domain WSD [61] | | 84.97 | 83.26 | 79.43 |
| ResNet-50+EdgeBox [62] | | 85.37 | 81.74 | 78.16 |
| ResNet-50+Selective Search [42] | | 84.82 | 81.43 | 77.69 |
| ResNet-50+DeepMask 1 [63] | | 86.59 | 83.31 | 78.94 |
| ResNet-50+DeepMask II [64] | | 86.34 | 83.86 | 79.25 |
| Mimic [49] | | 86.70 | 84.17 | 76.47 |
| Baseline | ResNet-50+ft | 84.31 | 81.26 | 79.63 |
| | ResNet-110+ft | 84.84 | 81.75 | 80.82 |
| | ResNet-50+(web)+ft | 84.89 | 82.72 | 82.31 |
| | ResNet-110+(web)+ft | 85.37 | 82.89 | 83.73 |
| Ablation | ResNet-50+(web+filter)+ft | 86.10 | 83.41 | 80.26 |
| | ResNet-50+(web+RPN)+ft | 86.64 | 82.65 | 79.52 |
| | ResNet-50+(web+RPN+Δ_form)+ft | 88.15 | 85.71 | 80.66 |
| | ResNet-50+(web+RPN+Δ_label)+ft | 89.32 | 85.24 | 82.09 |
| Ours | ResNet-50+Δ_form+Δ_label +ft | 90.41 | 87.95 | 84.93 |
| | ResNet-110+Δ_form+Δ_label +ft | 91.63 | 88.62 | 85.22 |

**TABLE V**

**EXPERIMENTAL RESULTS ON THE FOOD-101, STANFORD DOGS AND INDOOR-67 DATASETS.** "FT" MEANS FINE-TUNING ON THE MODEL PRE-TRAINED ON ImageNet. "+WEB" MEANS ADDING THE WEB DATA INTO THE STANDARD TRAINING SET. "+Δ_formΔ_label" REPRESENTS THAT DEBIASED WEB DATA IS USED. "ACC" IS CLASSIFICATION ACCURACY.
information.

We also conduct an experiment on another web dataset L-Dogs to illustrate the universality of the existence of bias. The L-Dogs dataset contains the same 120 classes and is collected from Google Images by Krause et al. [4]. The result of 85.90% for Goldfince [4] is achieved by using 342,632 images from 515-category dogs, and our method achieves higher accuracy by using only 1/3 of data in that work. The result tendency of experiments on L-Dogs is consistent with that on our collected web data, e.g. 82.59% for ResNet-110+(L-Dog) and 86.55% for ResNet-110+(L-Dog)+Δ form Δlabel+ft.

MIT indoor67. MIT Indoor67 contains 67 categories of indoor images, 5,360 for training and 1,340 for testing. Scene recognition requires knowledge about both scenes and various objects, making the task more challenging. As shown in Table V, the basic performance of ResNet50 is 79.63%, and the improvement for adding web data is 2.74%, and after removing the bias, the accuracy improves by 5.3%. Since the scale and density of domain information in scene images are more difficult to define than in object images, the improvement is not as significant as the results on object recognition datasets. Different from object classification, the intra-class variation is more obvious for scenes, so the web indoor scene images are more complex. Meanwhile, the results of other object detection methods can also prove that scene images are difficult to locate the labeled information, so more noisy regions (focusing on the object rather than the scene) will be introduced to influence the final results. Nevertheless, the classification accuracy of the proposed method is still higher than existing methods for scene recognition. Because the form constraint is a soft constraint for the content of the image, it will not break the wholeness of the scene.

In addition, we show the comparisons of related work on each task in Table VI. In these works, Goldfince [4] and DPTL [5] also use web data, and the amount of used data is larger than that used in our work. However, the proposed method gets better results by reducing the bias between the web and standard datasets. Compared to Progressive filter [51], our method can save more time for training, because Progressive filter needs to train the model for several rounds.

The proposed method effectively uses the web data to improve the performance of the classification models on all the three tasks.

### V. Conclusion

In contrast to previous works, in this paper, we reveal the phenomenon of web dataset biases and carry out rigorous quantitative analysis on them from various aspects. By conducting extensive experiments, we demonstrate that dataset bias causes the limited benefits of using web data. To address this problem, we present an unsupervised object localization method to provide critical insights into the object density and scale. Experiments show that the proposed method effectively reduces dataset bias. By employing de-biased web data, our method performs favorably against the state-of-the-art on multiple classification tasks. Web images are known to be easy and cheap to access. Although eliminating bias has not been fully solved, this work shows promising benefits towards this direction. How to use astronomically large webly-labeled data for a specific target learning task remains an open problem for future investigation.

### REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *NIPS*, 2012.

[2] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, “Learning deep features for scene recognition using places database,” in *NIPS*, 2014.

[3] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, “Learning hierarchical features for scene labeling,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[4] J. Krause, B. Sapp, A. Howard, H. Zhou, A. Toshev, T. Duerig, J. Philbin, and L. Fei-Fei, “The unreasonable effectiveness of noisy data for fine-grained recognition,” in *ECCV*, 2016.

[5] Y. Cui, Y. Song, C. Sun, A. Howard, and S. Belongie, “Large scale fine-grained categorization and domain-specific transfer learning,” in *CVPR*, 2018.

[6] 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.

[7] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *ICLR*, 2015.

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *CVPR*, 2015.

[9] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016.

[10] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang, “Learning from massive noisy labeled data for image classification,” in *ICCV*, 2015.
[68] Z. Xu, D. Tao, S. Huang, and Y. Zhang, “Friend or foe: Fine-grained categorization with weak supervision,” IEEE Transactions on Image Processing, 2017.

[69] A. Tatsuma and M. Aono, “Food image recognition using covariance of convolutional layer feature maps,” IEICE Transactions on Information and Systems, 2016.

[70] X. Zhang, H. Xiong, W. Zhou, W. Lin, and Q. Tian, “Picking deep filter responses for fine-grained image recognition,” in CVPR, 2016.

[71] D. Yoo, S. Park, J.-Y. Lee, and I. So Kweon, “Multi-scale pyramid pooling for deep convolutional representation,” in CVPR, 2015.

[72] R. Wu, B. Wang, W. Wang, and Y. Yu, “Harvesting discriminative meta objects with deep cnn features for scene classification,” in CVPR, 2015.

[73] K. Yanai and Y. Kawano, “Food image recognition using deep convolutional network with pre-training and fine-tuning,” in ICMEW, 2015.

[74] X. Zhang, H. Xiong, W. Zhou, and Q. Tian, “Fused one-vs-all features with semantic alignments for fine-grained visual categorization,” IEEE Transactions on Image Processing, 2016.

[75] M. Dixit, S. Chen, D. Gao, N. Rasiwasia, and N. Vasconcelos, “Scene classification with semantic fisher vectors,” in CVPR, 2015.

[76] A. Meyers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and K. P. Murphy, “Im2calories: towards an automated mobile vision food diary,” in CVPR, 2015.

[77] Y. Zhang, X. S. Wei, J. Wu, and J. Cai, “Weakly supervised fine-grained categorization with part-based image representation,” IEEE Transactions on Image Processing, 2016.

[78] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, “Places: A 10 million image database for scene recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 6, pp. 1452–1464, 2018.

[79] H. Hassannejad, G. Matrella, P. Ciampolini, I. De Munari, M. Mordonini, and S. Cagnoni, “Food image recognition using very deep convolutional networks,” in International Workshop on Multimedia Assisted Dietary Management, 2016.

[80] L. Niu, A. Veeraraghavan, and A. Sabharwal, “Webly supervised learning meets zero-shot learning: A hybrid approach for fine-grained classification,” in CVPR, 2018.

[81] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, “Deep layer aggregation,” in CVPR, 2018.

[82] P. Wang and N. Vasconcelos, “Towards realistic predictors,” in ECCV, 2018.