

# Webly Supervised Learning Meets Zero-shot Learning: A Hybrid Approach for Fine-grained Classification

Li Niu, Ashok Veeraraghavan, and Ashu Sabharwal  
Department of Electrical and Computer Engineering, Rice University  
{ln7, vashok, ashu}@rice.edu

## Abstract

*Fine-grained image classification, which targets at distinguishing subtle distinctions among various subordinate categories, remains a very difficult task due to the high annotation cost of enormous fine-grained categories. To cope with the scarcity of well-labeled training images, existing works mainly follow two research directions: 1) utilize freely available web images without human annotation; 2) only annotate some fine-grained categories and transfer the knowledge to other fine-grained categories, which falls into the scope of zero-shot learning (ZSL). However, the above two directions have their own drawbacks. For the first direction, the labels of web images are very noisy and the data distribution between web images and test images are considerably different. For the second direction, the performance gap between ZSL and traditional supervised learning is still very large. The drawbacks of the above two directions motivate us to design a new framework which can jointly leverage both web data and auxiliary labeled categories to predict the test categories that are not associated with any well-labeled training images. Comprehensive experiments on three benchmark datasets demonstrate the effectiveness of our proposed framework.*

## 1. Introduction

In recent years, there is a surge of advance in the field of image classification due to the rapid progress in deep learning techniques and available large-scale image datasets like ImageNet [11]. However, fine-grained image classification, which aims to classify myriads of subcategories belonging to one category such as dog breeds or bird species, remains to be a very challenging problem. In order to distinguish the subtle differences among fine-grained categories, a large number of well-labeled training images are required. However, human annotation for large-scale fine-grained categories is a rather expertise and difficult task because of the following reasons: 1) fine-grained annotation is usually in

high demand of professional knowledge; 2) the number of subcategories belonging to one category is generally very huge (e.g., 14,000 species of birds in the world [19]) and thus it is almost impossible to exhaustively collect training images for all the species. Therefore, lack of well-labeled training images becomes a critical issue for fine-grained classification. In this paper, we consider an extreme case in which we do not have any well-labeled training images for a given set of test categories. In this case, there are currently two main research directions, i.e., Webly Supervised Learning (WSL) approaches and Zero-Shot Learning (ZSL) approaches, which will be detailed separately as follows.

The Webly Supervised Learning (WSL) approaches crawl freely available web images from public website (e.g., Flickr) by using category names as queries. However, when applying the classifier learnt based on web images to the test images, the performance will drop sharply due to the label noise issue and domain distribution mismatch. In particular, on one hand, the labels of web images are usually very inaccurate. On the other hand, images may be compressed or edited before being uploaded to public websites so that the data distribution between the web images and the test images are quite different. Although several works [19, 48] on utilizing web data for fine-grained classification have been done to address the above issues, they involve various forms of human intervention or strong supervision (e.g., part location and bounding box). Instead, we focus on learning from web data without utilizing strong supervision or human intervention when crawling web images.

The Zero-Shot Learning (ZSL) [20] approaches assume that there exist well-labeled training data for a set of fine-grained categories (i.e., seen categories) while we need to recognize the instances from another set of test categories (i.e., unseen categories) which have no overlap with seen categories, that being said, we have no training data for test categories. The gap between seen categories and unseen categories are generally bridged by intermediate semantic information of all categories such as manually designed attributes [20] (e.g., color, shape, and material) or word vectors [24, 34] corresponding to category names ob-

tained based on free online corpus (*e.g.*, Wikipedia) [2]. However, the performance gap between ZSL and traditional supervised learning is still very large [2]. One problem for ZSL approaches is that they ignore the large amount of freely available web images.

To this end, we tend to propose a new learning scenario which combines webly supervised learning and zero-shot learning. To be exact, given a set of fine-grained test categories which are not associated with any well-labeled training images, we can crawl web images for test categories as weak form of supervision and also leverage the well-labeled images from other fine-grained categories. From another perspective, when given the entire set of all fine-grained categories belonging to one category (*e.g.*, 14,000 bird species), we only need to ask experts to label a few (*e.g.*, 100) fine-grained categories, and then can predict all the remaining fine-grained categories by virtue of web data. This learning scenario can be treated as zero-shot learning with additional noisy web training data for unseen categories, or learning from web data with additional well-labeled data from auxiliary categories.

In this learning scenario, we develop our framework as illustrated in Figure 1, from which it can be seen that we have a set of test categories without well-labeled training images and a set of auxiliary categories with well-labeled training images. Firstly, we crawl noisy web images for test categories by using their names as queries. Secondly, we extract deep visual features for all images, *i.e.*, web training images and well-labeled training images as well as test images. Thirdly, we extract semantic representations (*e.g.*, word vector) for all the categories based on their category names using the linguistic model trained on free online corpus (*e.g.*, Wikipedia). Finally, we input extracted visual features and semantic representations into our learning model and obtain the prediction results for the test set. As the core part of our framework, our learning model can tackle the label noise and domain shift issue of web images, and simultaneously transfer the knowledge from auxiliary categories to test categories. According to our framework, we tend to alleviate the human annotation burden as much as possible by virtue of web data. In this sense, “webly supervised” in our framework actually has double meanings: 1) we crawl noisy web training images from public websites for test categories; 2) we use semantic information of all categories obtained based on free online corpus to bridge the gap between test categories and auxiliary categories. To avoid ambiguity, in the remainder of this paper, we refer to the test categories with noisy web training images as weakly-supervised categories and the auxiliary categories with well-labeled training images as fully-supervised categories.

The major contributions of this paper are threefold: 1) to the best of our knowledge, this is the first work to propose the learning scenario for fine-grained image classification

Figure 1: The flowchart of our framework which jointly utilizes both web data and auxiliary categories for fine-grained image classification.

by jointly utilizing web data and auxiliary categories; 2) in this learning scenario, we develop a novel learning model, which unifies WSL and ZSL in one formulation with an efficient and effective solution; 3) extensive experiments on three benchmark datasets demonstrate the effectiveness our proposed framework.

## 2. Related Work

In this section, we will first introduce some works on fine-grained image classification by using web data. Then, we will describe existing ZSL methods. Since we need to address the domain shift between web images and test images, domain adaptation will also be briefly discussed.

**Learning from Web Data:** Learning from web data [21, 29, 32, 31] becomes increasingly popular and many research works have tended to address the label noise and domain shift issue. To name a few, NEIL in [10] relies on Multi-Instance Learning (MIL) to mitigate the label noise of web images. Several domain adaptation approaches were explored in [3] while a weakly supervised domain generalization approach was proposed in [28]. With the rapid development of deep learning, there are also several CNN approaches on learning from web images for image classification [46, 40, 9, 12, 51, 55]. However, the above works did not focus on the fine-grained setting.

In terms of utilizing web data for fine-grained image classification, label noise is removed in [19] using ac-

tive learning, which actually involves human intervention. In [48], web data together with additional bounding box annotations are required. Flickr images are leveraged in [42] to learn bird classifiers while crowd annotators participate in collecting the dataset. A more recent work [49] coped with the label noise and domain issue when utilizing web images. However, this work requires bounding boxes and part landmarks, which are not available in our setting. Distinctive from all the above works, we focus on learning from web data for fine-grained image classification without strong supervision (*e.g.*, part location and bounding box) or human intervention when crawling web images.

**Zero-Shot Learning:** In recent years, abundant Zero-Shot Learning (ZSL) approaches have been developed [37, 2, 20, 25]. More recently, some transductive or semi-supervised ZSL methods [47, 18, 22] demonstrated that it is effective to address the domain shift problem in ZSL by utilizing unlabeled test instances from unseen categories in the training phase. Hence, we also adopt a semi-supervised approach in our framework. Nevertheless, all the above ZSL approaches ignore the fact that there exist a large amount of freely available web images we can leverage for fine-grained image classification. In the contrast, we design a novel framework which can jointly leverage web images from test categories and well-labeled images from auxiliary categories.

**Domain Adaptation:** Domain adaptation methods [26] aim to address the domain shift issue, that is, to reduce the domain distribution mismatch between the source domain (*i.e.*, training set) and the target domain (*i.e.*, test set). There also exist domain generalization approaches [50, 27, 30] when the target domain is unseen in the training stage. The closest related is the work in [15], which reweights training instances based on MMD. However, the approach in [15] is not designed for webly supervised fine-grained image classification with auxiliary categories.

### 3. Our Formulation

In this paper, for ease of representation, a vector/matrix is denoted by a lowercase/uppercase letter in boldface. The transpose of a vector/matrix is denoted by the superscript  $\top$ . Moreover,  $\mathbf{A}^{-1}$  is used to denote the inverse matrix of  $\mathbf{A}$ . We use  $\mathbf{I}$  (*resp.*,  $\mathbf{O}$ ) to denote the identity matrix (*resp.*, zero matrix). Similarly, the vector with all ones is denoted as  $\mathbf{1}$ . We use  $\mathbf{A} \cdot \mathbf{B}$  (*resp.*,  $\mathbf{A} \odot \mathbf{B}$ ) to denote the inner product (*resp.*, element-wise product) of two matrices.

Recall that the flowchart of our framework has been introduced in Section 1, which is rephrased as follows. Given a set of test categories (*i.e.*, weakly-supervised categories) and auxiliary categories (*i.e.*, fully-supervised categories), we crawl noisy web images for test categories and obtain word vectors [24, 34] for all the categories. Then, we use the visual features of web training images and well-labeled training images as well as the semantic presentations of all

categories as input to our learning model. Moreover, we also utilize unlabeled test images in the training stage to address the domain distribution mismatch between web images and test images as well as the projection domain shift problem between fully-supervised categories and weakly-supervised categories in ZSL [47, 18, 22], leading to a semi-supervised learning model. With our learning model, we aim to obtain the predicted semantic representations of test images, which can be used for final category prediction. It is worth noting that the input and output of our learning model are specified in Figure 1.

Formally, we denote the visual features of training images from  $C^a$  fully-supervised categories as  $\mathbf{X}^a \in \mathbb{R}^{d \times n^a}$ , where  $d$  is the dimension of visual feature and  $n^a$  is the number of training images. Similarly, we denote the visual features of test images from  $C^t$  weakly-supervised categories as  $\mathbf{X}^t \in \mathbb{R}^{d \times n^t}$ , where  $n^t$  is the number of test images. We assume each category has a  $m$ -dim semantic representation and thus the semantic representation matrix of fully-supervised (*resp.*, weakly-supervised) categories is  $\mathbf{A}^a \in \mathbb{R}^{m \times C^a}$  (*resp.*,  $\mathbf{A}^t \in \mathbb{R}^{m \times C^t}$ ). Then, the semantic representation matrix of well-labeled training data is  $\mathbf{A}^a \in \mathbb{R}^{m \times n^a}$  with the semantic representation of each training instance equal to that of its category. Similarly, we use  $\mathbf{A}^t \in \mathbb{R}^{m \times n^t}$  to denote the semantic representation matrix of test data, which needs to be learnt. After learning  $\mathbf{A}^t$ , we can infer the category labels of test data by comparing  $\mathbf{A}^t$  with  $\mathbf{A}^a$ . Next, we describe how to transfer knowledge from fully-supervised categories to weakly-supervised categories as well as how to leverage noisy web training images.

#### 3.1. Knowledge Transfer

In order to transfer the knowledge from fully-supervised categories to weakly-supervised categories, inspired by [18], we learn two visual-semantic dictionaries  $\mathbf{D}^a$  and  $\mathbf{D}^t \in \mathbb{R}^{d \times m}$  separately for fully-supervised categories and weakly-supervised categories while enforcing  $\mathbf{D}^a$  and  $\mathbf{D}^t$  to be consistent to certain degree based on a co-regularizer  $\|\mathbf{D}^t - \mathbf{D}^a\|_F^2$ . We opt for dictionary learning approach because it lays the foundation for unifying WSL and ZSL elegantly. Specifically, the visual-semantic dictionary  $\mathbf{D}^a$  (*resp.*,  $\mathbf{D}^t$ ) is used to map from semantic representation space to visual feature space by minimizing the mapping error  $\|\mathbf{X}^a - \mathbf{D}^a \mathbf{A}^a\|_F^2$  (*resp.*,  $\|\mathbf{X}^t - \mathbf{D}^t \mathbf{A}^t\|_F^2$ ). The transfer process can be divided into two stages. In the first stage, we learn the dictionary of fully-supervised categories as

$$\min_{\mathbf{D}^a} \frac{1}{2} \|\mathbf{X}^a - \mathbf{D}^a \mathbf{A}^a\|_F^2 + \frac{1}{2} \|\mathbf{D}^a\|_F^2, \quad (1)$$

which is the same as in [18]. In the second stage, we minimize the mapping error  $\|\mathbf{X}^t - \mathbf{D}^t \mathbf{A}^t\|_F^2$  on the test images and enforce  $\mathbf{D}^t$  to be close to  $\mathbf{D}^a$  by using  $\|\mathbf{D}^t - \mathbf{D}^a\|_F^2$ . Additionally, considering that the semantic representations

of the test instances from the same category should be similar with each other, we expect  $\mathbf{A}^t$  to be low-rank. Thus, we introduce a novel nuclear norm [36] (convex approximation of rank function) regularizer  $\|\mathbf{A}^t\|_*$ , to enforce the semantic representation matrix of test data to be low-rank. Then, the formulation in the second stage can be written as

$$\min_{\mathbf{D}^t, \mathbf{A}^t} \frac{1}{2} \|\mathbf{X}^t - \mathbf{D}^t \mathbf{A}^t\|_F^2 + \frac{1}{2} \|\mathbf{D}^t - \mathbf{D}^a\|_F^2 + \frac{\lambda}{2} \|\mathbf{A}^t\|_*, \quad (2)$$

in which  $\lambda_1$  and  $\lambda_2$  are trade-off parameters. In the next section, we will extend (2) to leverage noisy web images.

### 3.2. Utilizing Noisy Web Images

Besides well-labeled training set  $\mathbf{X}^a$  and test set  $\mathbf{X}^t$ , we crawl web images using  $\mathbf{C}^t$  weakly-supervised category names as queries to construct the web training set with visual features  $\mathbf{X}^w \in \mathbb{R}^{d \times n^w}$ , where  $n^w$  is the number of web training images. In analogy to  $\mathbf{A}^a$ , the semantic representation matrix of web data is denoted as  $\mathbf{A}^w \in \mathbb{R}^{m \times n^w}$  with the semantic representation of each web training instance equal to that of its pseudo category (the label may be inaccurate). Since web images and test images come from the same set of weakly-supervised categories, we apply the same dictionary  $\mathbf{D}^t$  to the web images and minimize the mapping error  $\|\mathbf{X}^w - \mathbf{D}^t \mathbf{A}^w\|_F^2$ . Recall that we need to account for two issues when learning from web images: the label noise of web images and the domain shift between web images and test images.

To suppress the label noise of web images, we replace the mapping error, i.e., Frobenius norm  $\|\mathbf{X}^w - \mathbf{D}^t \mathbf{A}^w\|_F^2$ , with  $L_{2,1}$  norm  $\|\mathbf{X}^w - \mathbf{D}^t \mathbf{A}^w\|_{2,1}$ , which is also referred to as group-lasso regularizer [52]. After employing group-lasso regularizer,  $\mathbf{X}^w - \mathbf{D}^t \mathbf{A}^w$  is expected to be column-sparse and the columns with non-zero entries correspond to the outliers, which is granted the larger error tolerance. In this way, we can learn a more robust dictionary  $\mathbf{D}^t$  on the weakly-supervised categories.

To address the domain distribution mismatch between web images (i.e.,  $\mathbf{X}^w$ ) and test images (i.e.,  $\mathbf{X}^t$ ), we employ an Maximum Mean Discrepancy (MMD) [15] based regularizer  $\frac{1}{n^w} \|\mathbf{X}^w\|^2 - \frac{1}{n^t} \|\mathbf{X}^t\|^2$  with weight vector  $\mathbf{1}$  to be learnt. The idea of MMD-based regularizer is to reduce the distance between the center of weighted web images (i.e.,  $\frac{1}{n^w} \|\mathbf{X}^w\|^2$ ) and the center of test images (i.e.,  $\frac{1}{n^t} \|\mathbf{X}^t\|^2$ ) by assigning higher weights on the web images which are closer to the center of test images.

To take full advantage of the weight vector  $\mathbf{1}$ , we expect to assign higher weights to the web training images with not only closer distribution to the center of test images, but also relatively accurate labels. Therefore, besides the MMD-based regularizer  $\frac{1}{n^w} \|\mathbf{X}^w\|^2 - \frac{1}{n^t} \|\mathbf{X}^t\|^2$ , we also add the weights  $\mathbf{1}$  in the group-lasso regularizer  $(\mathbf{X}^w - \mathbf{D}^t \mathbf{A}^w)_{2,1}$ , in which  $\mathbf{1}$  is a diagonal matrix with  $\mathbf{1}$  being its diagonal. In this case, lower weights are prone

to be assigned to the columns of  $\mathbf{X}^w - \mathbf{D}^t \mathbf{A}^w$  with non-zero entries, which correspond to the outliers. By using the importance weight vector  $\mathbf{1}$  in two regularizers, we collaboratively deal with the label noise issue and the domain shift issue, different from most previous works which usually address these two issues separately.

From another point of view, since the dictionary  $\mathbf{D}^t$  used in  $(\mathbf{X}^w - \mathbf{D}^t \mathbf{A}^w)_{2,1}$  is enforced to be close to the dictionary  $\mathbf{D}^a$  of auxiliary categories, we actually leverage auxiliary categories to help tackle the label noise of web images. To this end, we extend (2) to the following problem:

$$\begin{aligned} \min_{\mathbf{D}^t, \mathbf{A}^t} \quad & \frac{1}{2} \|\mathbf{X}^t - \mathbf{D}^t \mathbf{A}^t\|_F^2 + \frac{1}{2} \|\mathbf{D}^t - \mathbf{D}^a\|_F^2 + \frac{\lambda}{2} \|\mathbf{A}^t\|_* \\ & + \frac{3}{2} \frac{1}{n^w} \|\mathbf{X}^w\|^2 - \frac{1}{n^t} \|\mathbf{X}^t\|^2 + \frac{\lambda}{4} (\mathbf{X}^w - \mathbf{D}^t \mathbf{A}^w)_{2,1}, \quad (3) \\ \text{s.t.} \quad & \mathbf{1} = n^w, \quad 0 \leq \mathbf{1} \leq \mathbf{b}, \quad (4) \end{aligned}$$

where  $\lambda_3$ ,  $\lambda_4$ , and  $\mathbf{b}$  are trade-off parameters. Note that we regulate  $\mathbf{1}$  using a sum constraint and a box constraint in (4), in which  $\mathbf{b}$  is the upper bound of importance weights.

### 4. Optimization

In the first stage, the problem in (1) has a close-form solution  $\mathbf{D}^a = \mathbf{X}^a \mathbf{A}^a (\mathbf{A}^a \mathbf{A}^a + \mathbf{I})^{-1}$ . In the second stage, the problem in (3) is too hard to solve due to the group lasso regularizer and low-rank regularizer, so we develop a novel solution based on inexact Augmented Lagrange Multiplier (ALM) [5]. For ease of optimization, we introduce intermediate variable  $\mathbf{E}^w$  (resp.,  $\mathbf{Z}^t$ ) to replace  $(\mathbf{X}^w - \mathbf{D}^t \mathbf{A}^w)$  (resp.,  $\mathbf{A}^t$  in  $\mathbf{A}^t$ ) in (3). Then, the problem in (3) can be rewritten as

$$\begin{aligned} \min_{\mathbf{D}^t, \mathbf{A}^t} \quad & \frac{1}{2} \|\mathbf{X}^t - \mathbf{D}^t \mathbf{A}^t\|_F^2 + \frac{1}{2} \|\mathbf{D}^t - \mathbf{D}^a\|_F^2 + \frac{\lambda}{2} \|\mathbf{Z}^t\|_* \\ & + \frac{3}{2} \frac{1}{n^w} \|\mathbf{X}^w\|^2 - \frac{1}{n^t} \|\mathbf{X}^t\|^2 + \frac{\lambda}{4} \|\mathbf{E}^w\|_{2,1}, \quad (5) \\ \text{s.t.} \quad & \mathbf{1} = n^w, \quad 0 \leq \mathbf{1} \leq \mathbf{b}, \\ & \mathbf{E}^w = (\mathbf{X}^w - \mathbf{D}^t \mathbf{A}^w), \quad \mathbf{Z}^t = \mathbf{A}^t. \end{aligned}$$

Then, we aim to minimize the following augmented Lagrangian function:

$$\begin{aligned} L_{\mathbf{D}^t, \mathbf{A}^t, \mathbf{Z}^t} = \quad & \frac{1}{2} \|\mathbf{X}^t - \mathbf{D}^t \mathbf{A}^t\|_F^2 + \frac{1}{2} \|\mathbf{D}^t - \mathbf{D}^a\|_F^2 + \frac{\lambda}{2} \|\mathbf{Z}^t\|_* \\ & + \frac{3}{2} \frac{1}{n^w} \|\mathbf{X}^w\|^2 - \frac{1}{n^t} \|\mathbf{X}^t\|^2 + \frac{\lambda}{4} \|\mathbf{E}^w\|_{2,1} \\ & + \frac{\mu}{2} \|\mathbf{E}^w - (\mathbf{X}^w - \mathbf{D}^t \mathbf{A}^w)\|_F^2 + \mathbf{R}, \mathbf{E}^w - (\mathbf{X}^w - \mathbf{D}^t \mathbf{A}^w) \\ & + \frac{\mu}{2} \|\mathbf{A}^t - \mathbf{Z}^t\|_F^2 + \mathbf{T}, \mathbf{A}^t - \mathbf{Z}^t, \quad (6) \end{aligned}$$

in which  $\mathbf{S} = \{\mathbf{1} \mid \mathbf{1} = n^w, 0 \leq \mathbf{1} \leq \mathbf{b}\}$ ,  $\mu$  is a penalty parameter, and  $\{\mathbf{R}, \mathbf{T}\}$  are Lagrangian multipliers. We update the variables  $\{\mathbf{E}^w, \mathbf{Z}^t, \mathbf{D}^t, \mathbf{A}^t\}$ , the Lagrangian multipliers  $\{\mathbf{R}, \mathbf{T}\}$ , and the penalty parameter  $\mu$  iteratively



---

**Algorithm 1** Solving (5) with inexact ALM
 

---

```

1: Input:  $\mathbf{X}^a, \mathbf{A}^a, \mathbf{X}^w, \mathbf{A}^w, \mathbf{X}^t, \mathbf{D}^a$ .
2: Initialize  $\mathbf{R} = \mathbf{O}, \mathbf{T} = \mathbf{O}, \beta = 1, \mathbf{D}^t = \mathbf{D}^a, \mu = 0.1, \mu_{\max} = 10^6, \epsilon = 10^{-5}, N_{\text{iter}} = 10^6$ .
3: for  $t = 1 : N_{\text{iter}}$  do
4:   Update  $\mathbf{E}^w$  by using (7).
5:   Update  $\mathbf{Z}^t$  by using (9).
6:   Update  $\mathbf{D}^t$  by using (11).
7:   Update  $\mathbf{A}^t$  by using (12).
8:   Update  $\mu$  by solving (15).
9:   Update  $\mathbf{R}$  by  $\mathbf{R} = \mathbf{R} + \mu(\mathbf{E}^w - (\mathbf{X}^w - \mathbf{D}^t \mathbf{A}^w))$ .
10:  Update  $\mathbf{T}$  by  $\mathbf{T} = \mathbf{T} + \mu(\mathbf{A}^t - \mathbf{Z}^t)$ .
11:  Update the parameter  $\mu$  by  $\mu = \min(\mu_{\max}, (1 + \epsilon)\mu)$ .
12:  Break if  $\|\mathbf{E}^w - (\mathbf{X}^w - \mathbf{D}^t \mathbf{A}^w)\|_F < \epsilon$  and  $\|\mathbf{A}^t - \mathbf{Z}^t\|_F < \epsilon$ .
13: end for
14: Output:  $\mathbf{A}^t$ .
  
```

---

until the termination criterion is met. Updating the Lagrangian multipliers and  $\mu$  is trivial, which can be found in Algorithm 1. In the following, we will describe how to update  $\mathbf{E}^w, \mathbf{Z}^t, \mathbf{D}^t, \mathbf{A}^t$ , and  $\mu$  one by one.

**Update  $\mathbf{E}^w$ :** The subproblem of (6) w.r.t.  $\mathbf{E}^w$  is as follows,

$$\min_{\mathbf{E}^w} \frac{1}{2} \|\mathbf{E}^w\|_F^2 + \frac{\mu}{2} \|\mathbf{E}^w - (\mathbf{X}^w - \mathbf{D}^t \mathbf{A}^w)\|_F^2 - \frac{\mathbf{R}}{\mu} \|\mathbf{E}^w\|_F^2, \quad (7)$$

which has a close-form solution [23]. Specifically, by denoting  $\mathbf{Q} = (\mathbf{X}^w - \mathbf{D}^t \mathbf{A}^w) - \frac{\mathbf{R}}{\mu}$ , if the optimal solution w.r.t.  $\mathbf{E}^w$  is  $\mathbf{E}^w$ , then the  $i$ -th column of  $\mathbf{E}^w$  is

$$\mathbf{E}^w(:, i) = \begin{cases} \frac{q_i - \frac{4}{\mu} q_i}{2}, & \text{if } \frac{4}{\mu} < q_i, \\ 0, & \text{otherwise,} \end{cases} \quad (7)$$

where  $q_i$  is the  $i$ -th column of  $\mathbf{Q}$  and  $q_i$  is the  $L_2$  norm of  $q_i$ .

**Update  $\mathbf{Z}^t$ :** The subproblem of (6) w.r.t.  $\mathbf{Z}^t$  is as follows,

$$\min_{\mathbf{Z}^t} \frac{1}{2} \|\mathbf{Z}^t\|_F^2 + \frac{\mu}{2} \|\mathbf{Z}^t - (\mathbf{A}^t + \frac{\mathbf{T}}{\mu})\|_F^2, \quad (8)$$

which can be solved by using the Singular Value Threshold (SVT) method [6]. By denoting  $\mathbf{M} = \mathbf{A}^t + \frac{\mathbf{T}}{\mu}$  and the rank of  $\mathbf{M}$  as  $r$ , the singular value decomposition of  $\mathbf{M}$  can be represented as  $\mathbf{M} = \mathbf{U} \mathbf{V}$ , where  $\mathbf{U} \in \mathbb{R}^{m \times r}, \mathbf{V} \in \mathbb{R}^{r \times n^t}$ , and  $\mathbf{\Sigma} = \mathbf{R}^{r \times r}$  is a diagonal matrix with diagonal entries being the singular values of  $\mathbf{M}$ . Then, the solution w.r.t.  $\mathbf{Z}^t$  can be obtained as follows,

$$\mathbf{Z}^t = \mathbf{U} \mathbf{D}(\cdot)_+ \mathbf{V}, \quad (9)$$

where  $\mathbf{D}(\cdot)_+$  is a diagonal matrix with the diagonal being  $\{(\sigma_i - \frac{2}{\mu})_+ | \sigma_i = 1\}$ , in which  $\sigma_i$  is the  $i$ -th diagonal entry of  $\mathbf{\Sigma}$  and  $(\cdot)_+$  is a thresholding operator by assigning the negative entries to zeros.

**Update  $\mathbf{D}^t$ :** The subproblem of (6) w.r.t.  $\mathbf{D}^t$  is as follows,

$$\min_{\mathbf{D}^t} \frac{1}{2} \|\mathbf{X}^t - \mathbf{D}^t \mathbf{A}^t\|_F^2 + \frac{1}{2} \|\mathbf{D}^t - \mathbf{D}^a\|_F^2 + \frac{\mu}{2} \|\mathbf{E}^w - (\mathbf{X}^w - \mathbf{D}^t \mathbf{A}^w)\|_F^2 + \mathbf{R} \|\mathbf{E}^w - (\mathbf{X}^w - \mathbf{D}^t \mathbf{A}^w)\|_F. \quad (10)$$

By setting the derivative of (10) w.r.t.  $\mathbf{D}^t$  as zeros, we can obtain the close-form solution to  $\mathbf{D}^t$  as

$$\mathbf{D}^t = \frac{\mathbf{X}^t \mathbf{A}^t + \beta \mathbf{D}^a + (\mu \mathbf{X}^w - \mu \mathbf{E}^w - \mathbf{R}) \mathbf{A}^w}{\mathbf{A}^t \mathbf{A}^t + \mu \mathbf{A}^w \mathbf{A}^w + \beta \mathbf{I}}^{-1}. \quad (11)$$

**Update  $\mathbf{A}^t$ :** The subproblem of (6) w.r.t.  $\mathbf{A}^t$  is as follows,

$$\min_{\mathbf{A}^t} \frac{1}{2} \|\mathbf{X}^t - \mathbf{D}^t \mathbf{A}^t\|_F^2 + \frac{\mu}{2} \|\mathbf{A}^t - \mathbf{Z}^t\|_F^2 + \mathbf{T} \|\mathbf{A}^t - \mathbf{Z}^t\|_F,$$

which also has a close-form solution:

$$\mathbf{A}^t = \frac{\mathbf{D}^t \mathbf{X}^t + \mu \mathbf{Z}^t - \mathbf{T}}{\mathbf{D}^t \mathbf{D}^t + \mu \mathbf{I}}^{-1}. \quad (12)$$

**Update  $\mu$ :** The subproblem of (6) w.r.t.  $\mu$  is as follows,

$$\min_{\mu} \frac{3}{2} \frac{1}{n^w} \|\mathbf{X}^w\|_F^2 - \frac{1}{n^t} \|\mathbf{X}^t\|_F^2 + \frac{\mu}{2} \|\mathbf{E}^w - (\mathbf{X}^w - \mathbf{D}^t \mathbf{A}^w)\|_F^2 + \mathbf{R} \|\mathbf{E}^w - (\mathbf{X}^w - \mathbf{D}^t \mathbf{A}^w)\|_F. \quad (13)$$

After omitting the constant terms without  $\mu$ , the problem in (13) can be converted to

$$\min_{\mu} \frac{3}{2(n^w)^2} \|\mathbf{X}^w\|_F^2 - \frac{3}{n^w n^t} \|\mathbf{X}^w\|_F \|\mathbf{X}^t\|_F + \frac{\mu}{2} \|\bar{\mathbf{P}} - \mu \hat{\mathbf{p}} - \hat{\mathbf{r}}\|_F^2, \quad (14)$$

in which  $\bar{\mathbf{P}}$  is a diagonal matrix sharing the same diagonal with  $(\mathbf{X}^w - \mathbf{D}^t \mathbf{A}^w)(\mathbf{X}^w - \mathbf{D}^t \mathbf{A}^w)$ ,  $\hat{\mathbf{p}}$  is the diagonal of  $(\mathbf{X}^w - \mathbf{D}^t \mathbf{A}^w) \mathbf{E}^w$ , and  $\hat{\mathbf{r}} = (\mathbf{R} (\mathbf{X}^w - \mathbf{D}^t \mathbf{A}^w)) \mathbf{1}$ . The problem in (14) can be further simplified as

$$\min_{\mu} \frac{1}{2} \|\mathbf{H} - \mathbf{f}\|_F^2, \quad (15)$$

in which  $\mathbf{H} = \frac{3}{(n^w)^2} \mathbf{X}^w \mathbf{X}^w + \mu \bar{\mathbf{P}}$  and  $\mathbf{f} = \frac{3}{n^w n^t} \mathbf{X}^w \mathbf{X}^t \mathbf{1} + \mu \hat{\mathbf{p}} + \hat{\mathbf{r}}$ . The problem in (15) is a quadratic programming (QP) problem which can be solved by using existing QP solvers (e.g., Mosek). However, it is very time-consuming to use existing QP solvers, so we develop our own Sequential Minimal Optimization (SMO) [35] based algorithm to solve (15), which is much more efficient than those off-the-shelf QP solvers.

The whole algorithm using inexact ALM is summarized in Algorithm 1. Based on our experimental observation, the algorithm usually converges within 50 iterations. After obtaining the semantic representations of test images  $\mathbf{A}^t$ , given the semantic representation matrix of test categories  $\mathbf{A}^t$ , we adopt nearest neighbour (NN) classifier for final prediction, following the strategy in [18]. Specifically, we compare the semantic representation of each test instance (i.e., each column in  $\mathbf{A}^t$ ) with that of each test category (i.e., each column in  $\mathbf{A}^t$ ), and label each test instance with the nearest test category.

## 5. Experiments

In this section, we evaluate our method on three benchmark datasets with ablation study. Moreover, we conduct additional experiments under the generalized setting in which the test instances may come from both fully-supervised and weakly-supervised categories.

### 5.1. Fine-grained Image Classification

**Datasets:** We conduct experiments on the following three popular benchmark datasets which are commonly used for zero-shot learning (ZSL) tasks, since our learning scenario can be treated as ZSL with additional web training images for unseen categories, as mentioned in Section 1.

1) CUB [43]: Caltech-UCSD Bird (CUB) has in total 11,788 images from 200 bird species. Following [1], we use the standard train-test split with 150 fully-supervised (*resp.*, 50 weakly-supervised) categories.

2) SUN [45]: Scene UNDERstanding (SUN) attribute dataset has 20 images in each scene category. Following [16], we use the standard train-test split with 707 fully-supervised (*resp.*, 10 weakly-supervised) categories.

3) Dogs [17]: Stanford Dogs dataset has 19,501 images from 113 dog breeds. We follow the provided train-test split in [2], *i.e.*, 85 fully-supervised (*resp.*, 28 weakly-supervised) categories.

4) Flickr image dataset: We construct the web training set by ourselves. Particularly, for each benchmark dataset (*i.e.*, CUB, SUN, and Dogs), we use the names of weakly-supervised categories as queries to collect the top ranked 100 images from Flickr website for each category after performing PCA based near-duplicate removal [54].

**Features:** We extract visual features for all the images and semantic representations for all the categories, which will be detailed next.

1) Visual features: For each image, we use 4,096-dim output of the 6-th layer of the pretrained VGG [39] model as its visual feature.

2) Semantic representations: We exploit two types of word vectors Word2Vec [24] and GloVe [34], in which each word can be represented as a real-valued vector. We train Word2Vec and GloVe language models based on the latest Wikipedia corpus, with the word vector dimension being 400. Then, we concatenate the word vectors corresponding to each category name from Word2Vec and GloVe as its semantic representation, leading to an 800-dim vector for each category. For those category names with more than one word, we average the semantic representations corresponding to all words within the category name as its final semantic representation.

**Baselines:** We compare with four sets of baselines: ZSL baselines, WSL baselines, domain adaptation (DA) baselines, and the combo baseline. Among these baselines, ZSL

Table 1: Accuracies (%) of different methods on three datasets. The best results are highlighted in boldface.

Dataset	CUB	SUN	Dogs	Avg
LR	<b>68.39</b>	<b>62.50</b>	<b>77.67</b>	<b>69.52</b>
KMM	70.54	<b>64.00</b>	79.16	71.23
GFK	70.37	<b>62.50</b>	79.51	70.79
SA	<b>68.67</b>	<b>63.00</b>	80.18	70.62
TCA	<b>68.56</b>	<b>63.00</b>	80.22	70.59
CORAL	<b>69.04</b>	<b>63.50</b>	<b>80.37</b>	70.97
NEIL	<b>69.08</b>	<b>63.00</b>	<b>80.16</b>	70.74
Bergamo and Torresani	70.13	<b>64.00</b>	78.64	70.93
WSDG	70.61	<b>66.00</b>	<b>80.20</b>	72.27
Sukhbaatar <i>et al.</i>	70.47	<b>64.50</b>	<b>81.15</b>	72.04
Xiao <i>et al.</i>	70.92	<b>65.50</b>	<b>81.67</b>	72.69
ESZSL	<b>38.08</b>	<b>65.00</b>	37.21	46.77
LatEm	35.15	<b>66.50</b>	35.99	45.88
SJE	42.65	71.50	34.85	49.67
DAP/IAP	28.91	57.50	33.15	39.85
Changpinyo <i>et al.</i>	41.83	72.00	39.91	51.25
Li <i>et al.</i>	32.36	72.50	43.15	49.34
Kodirov <i>et al.</i>	47.53	71.00	47.32	55.28
Zhang and Saligrama	44.08	76.50	48.09	56.23
Xu <i>et al.</i>	45.72	71.50	39.85	52.36
Shojaee and Baghshah	46.68	71.00	48.82	55.50
WSL+ZSL	72.21	78.50	81.90	77.53
Ours_WSL	69.42	65.50	80.43	71.78
Ours_ZSL	47.94	71.50	47.70	55.71
Ours_sim1	72.72	83.50	85.04	80.42
Ours_sim2	76.00	79.50	83.75	79.75
Ours	<b>76.47</b>	<b>84.50</b>	<b>85.16</b>	<b>82.04</b>

baselines cannot utilize web images while DA/WSL baselines cannot utilize auxiliary categories. Besides, as far as we are concerned, there is no existing method that can jointly utilize web data and auxiliary categories, so we combine the most competitive ZSL and DA/WSL baselines by simply averaging their test decision values. Intuitively, the strongest baseline should be the combo baseline, since it can utilize both web images and auxiliary categories.

For ZSL baselines, we include the standard ZSL methods ESZSL [37], LatEm [44], SJE [2], DAP/IAP [20], Changpinyo *et al.* [7], and transductive/semi-supervised ZSL methods Li *et al.* [22], Kodirov *et al.* [18], Zhang and Saligrama [53], Xu *et al.* [47], Shojaee and Baghshah [38] as baselines. The difference between standard ZSL and transductive/semi-supervised ZSL lies in whether to utilize unlabeled test data in the training stage.

For DA baselines, we compare with the following approaches: KMM [15], GFK [14], SA [13], TCA [33], and CORAL [41]. The web training images (*resp.*, test images) are treated as the source (*resp.*, target) domain.

For WSL baselines, we compare with NEIL [10], Bergamo and Torresani [3], WSDG [28], sukhaatar *et al.* [40],

and Xiao *et al.* [46]. Note that Xiao *et al.* [46] utilizes manually cleaned web data when computing confusion matrix and training network, which is not applicable in our scenario. Thus, for fair comparison, we evaluate [46] without using manually cleaned web data and estimate the confusion matrix based on semantic representations.

For combo baseline, we average the test decision values from the most competitive WSL baseline (*i.e.*, Xiao *et al.*) and ZSL baseline (*i.e.*, Zhang and Saligrama) in Table 1 for comparison, which is referred to as WSL+ZSL.

We also include one basic baseline LR, which simply learns a linear regressor based on web training images. In order to valid the WSL and ZSL components in our formulation (3), we report the results of our two special cases. To be exact, we remove the regularizer related to knowledge transfer (*i.e.*,  $\mathbf{D}^t - \mathbf{D}^a \frac{2}{F}$ ) by setting  $\gamma_1$  as 0, and refer to this special case as Ours\_WSL. Similarly, we remove the regularizers related to web data (*i.e.*,  $\frac{1}{n^w} \mathbf{X}^w - \frac{1}{n^t} \mathbf{X}^t \mathbf{1}^2$  and  $(\mathbf{X}^w - \mathbf{D}^t \mathbf{A}^w)_{2,1}$ ) by setting  $\gamma_3$  and  $\gamma_4$  as 0, and refer to this special case as Ours\_ZSL. In order to validate the regularizers used in our formulation (3), we further compare with our two simplified versions. In particular, we remove the regularizer  $\mathbf{A}^t$  (*resp.*,  $\frac{1}{n^w} \mathbf{X}^w - \frac{1}{n^t} \mathbf{X}^t \mathbf{1}^2$ ) in (3) by setting  $\gamma_2$  (*resp.*,  $\gamma_3$ ) as 0 and refer to this simplified version as Ours\_sim1 (*resp.*, Ours\_sim2). For all the methods, we use multi-class accuracy for performance evaluation.

**Parameters:** Our method has trade-off parameters  $b$ ,  $\gamma_1$ ,  $\gamma_2$ ,  $\gamma_3$ , and  $\gamma_4$  in (3), which are determined by cross-validation. Specifically, following the cross-validation strategy used in [38], we choose the first  $C^c$  categories based on the default category indices from  $C^a$  fully-supervised categories as validation categories, in which  $C^c$  satisfies  $\frac{C^c}{C^a} = \frac{C^t}{C^a + C^t}$ . Note that we need to additionally crawl web images and extract their visual features for validation categories. In the validation stage, we use  $C^a - C^c$  categories as fully-supervised categories and  $C^c$  categories as weakly-supervised categories. Then, we determine the optimal trade-off parameters based on the validation performance through random search [4] within certain range. Particularly, the parameters  $\gamma_1$ ,  $\gamma_2$ ,  $\gamma_3$ , and  $\gamma_4$  are empirically searched within the range  $[10^{-3}, 10^{-2}, \dots, 10^3]$ , and the parameter  $b$  is empirically searched within the range  $[1.5, 2.0, \dots, 5.0]$ .

**Experimental Results:** The experimental results of all methods are reported in Table 1, based on which we have the following observations:

1) The DA and WSL baselines outperform LR, which indicates the benefits of addressing domain shift or label noise. The ZSL baselines are worse than DA/WSL baselines on CUB and Dogs datasets, but generally better on the SUN dataset. The inconsistent superiority of ZSL or DA/WSL baselines highly depends on the purity of web images as well as the relation between auxiliary and test categories.

2) Our method performs much better than Ours\_WSL and Ours\_ZSL, which shows the advantage of unifying WSL and ZSL. Our method also outperforms Ours\_sim1 and Ours\_sim2, which validates the effectiveness of our low-rank and MMD-based regularizer.

3) Note that the focus of this paper is introducing a new learning scenario for fine-grained image classification with both web data and auxiliary categories, instead of proposing a state-of-the-art WSL or ZSL approach. Therefore, there is no guarantee that Ours\_WSL (*resp.*, Ours\_ZSL) can outperform all WSL (*resp.*, ZSL) baselines. However, when using both web data and auxiliary categories, our method yields significant improvement over the strongest combo baseline WSL+ZSL, which demonstrates that a naive combination cannot take full advantage of both web data and auxiliary categories. In the contrast, we unify ZSL and WSL coherently in one formulation, which greatly facilitates fine-grained image classification.

**Utilizing More Web Images:** Since we only utilize 100 web training images for each weakly-supervised category, it is interesting to explore whether the performance will increase with more web training images. We study the variation of performance w.r.t. different numbers of web training images. In particular, we crawl various numbers of web images for each weakly-supervised category (*i.e.*,  $[100, 200, \dots, 1000]$ ) to construct the web training set and keep the remaining experimental settings unchanged. The accuracies with various numbers of web training images on three datasets are reported in Figure 2, from which we observe that for the CUB and Dogs dataset, the accuracy increases as the number of web training images increases within certain range. However, for the SUN dataset, the accuracy drops dramatically when the number of web training images increases. This is possibly because that scene name is more ambiguous than the dog/bird name. Furthermore, the fine-grained scene categories in the SUN datasets are associated with additional “in\_door” or “out\_door” label, making it even harder to crawl the semantically correct web images.

**Qualitative Analysis of Learnt Weights :** Based on our formulation (3), higher weights are expected to be assigned to the web training images with closer distribution to the center of test images and relatively accurate labels. So the web images with higher (*resp.*, lower) weights are prone to be non-outliers (*resp.*, outliers). By taking the Dogs dataset as an example, we rank the web training images based on the learnt  $\mathbf{w}$ , and show the web images with 5 highest weights and 5 lowest weights in Figure 3, in which the numbers below images are their corresponding weights within the range  $[0, 1.5]$  since the cross validated  $b$  on the Dogs dataset is 1.5. From Figure 3, we observe that the top row of images with highest weights have accurate labels. Moreover, the dog occupies the large center of the entire

Figure 2: The performance variation of our method w.r.t. different numbers of web training images per category.

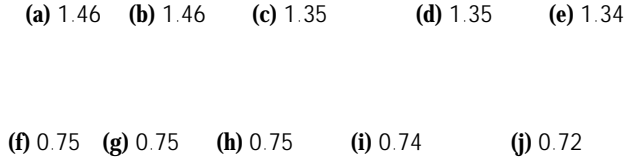


Figure 3: The web images in the top (*resp.*, bottom) row are associated with 5 highest (*resp.*, lowest) weights based on the learnt weight vector .

Table 2: Accuracies (%) of different methods on three datasets under the generalized setting. The best results are highlighted in boldface.

Dataset	CUB	SUN	Dogs	Avg
LR_mix	55.27	32.03	53.74	47.01
WSL+LR	57.60	35.11	55.13	49.28
Chao <i>et al.</i>	25.75	20.77	31.53	26.02
Ours	<b>59.60</b>	<b>36.00</b>	<b>65.89</b>	<b>53.83</b>

image, visually resembling the test images, which implies close data distribution between these web images and test images. In the contrast, the web images in the bottom row are quite noisy. We have similar observations on the other two datasets.

## 5.2. Extension to Generalized Setting

In some real-world applications, the test instances are likely to come from both fully-supervised categories and weakly-supervised categories. For example, given the entire set of all fine-grained categories belonging to one category (*e.g.*, 14,000 bird species), we annotate a few (*e.g.*, 100) fine-grained categories and crawl web images for

the remaining fine-grained categories. With our learning model, we hope to predict the test images which may come from any fine-grained category. Excluding web training images, this setting reduces to the generalized Zero-Shot Learning (ZSL) setting [8], in which the test samples may come from both seen and unseen categories. In order to validate the effectiveness of our method in the generalized setting, we additionally conduct experiments with the test set as the mixture of fully-supervised categories and weakly-supervised categories. Note that for the generalized setting, our method in (3) can be readily applied with a little abusively used dictionary  $\mathbf{D}^t$  in  $\mathbf{X}^t - \mathbf{D}^t \mathbf{A}^t \frac{2}{F}$  for all categories instead of only weakly-supervised categories. After obtaining  $\mathbf{A}^t$ , we use nearest neighbour (NN) classifier to label test instances based on  $\mathbf{A}^t$  and  $\bar{\mathbf{A}} = [\bar{\mathbf{A}}^a, \bar{\mathbf{A}}^t]$ , similarly as in Section 4. Following the setting in [8], we move 20% of the training instances from each fully-supervised category to the test set so that the new test set consists of the instances from both fully-supervised and weakly-supervised categories. Therefore, the new test set of CUB (*resp.*, SUN and Dogs) contains in total 200 (*resp.*, 717 and 113) categories.

For baselines, we compare our method with basic linear regression which learns a linear regressor for each fully-supervised or weakly-supervised category, which is referred to as LR\_mix in Table 2. We also compare with WSL+LR which uses Xiao *et al.* [46] for weakly-supervised categories and linear regressor for fully-supervised categories, considering that Xiao *et al.* [46] is the most competitive WSL baseline as reported in Table 1. Moreover, we include generalized ZSL method in [8] as a baseline, which claimed to achieve the state-of-the-art performance under the generalized ZSL setting [8].

The experimental results under the generalized setting are summarized in Table 2, from which we observe that the results suffer from significant drop when compared with those reported in Table 1. However, our method still achieves the best results on three datasets, which demonstrates the effectiveness of our method under the generalized setting.

## 6. Conclusion

In this paper, we have proposed a new learning scenario, *i.e.*, fine-grained image classification by jointly utilizing web data and auxiliary labeled categories. The superiority of our proposed framework has been demonstrated by comprehensive experiments.

## Acknowledgement

This work is supported by NIH 7000000356 and NSF IIS-1652633. This work is also supported by the NGA NHARP program HM0476-15-1-0007.



## References

- [1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label embedding for attribute-based classification. In *CVPR*, 2013. **6**
- [2] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, 2015. **2, 3, 6**
- [3] A. Bergamo and L. Torresani. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In *NIPS*, 2010. **2, 6**
- [4] J. Bergstra and Y. Bengio. Random search for hyperparameter optimization. *JMLR*, 13(Feb):281–305, 2012. **7**
- [5] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends<sup>R</sup> in Machine Learning*, 2011. **4**
- [6] J.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 2010. **5**
- [7] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, 2016. **6**
- [8] W.-L. Chao, S. Changpinyo, B. Gong, and F. Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *ECCV*, 2016. **8**
- [9] X. Chen and A. Gupta. Webly supervised learning of convolutional networks. In *ICCV*, 2015. **2**
- [10] X. Chen, A. Shrivastava, and A. Gupta. NEIL: Extracting visual knowledge from web data. In *ICCV*, pages 1409–1416, 2013. **2, 6**
- [11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. **1**
- [12] S. K. Divvala, A. Farhadi, and C. Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *CVPR*, 2014. **2**
- [13] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *ICCV*, 2013. **6**
- [14] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012. **6**
- [15] J. Huang, A. Smola, A. Gretton, K. Borgwardt, and B. Scholkopf. Correcting sample selection bias by unlabeled data. In *NIPS*, 2006. **3, 4, 6**
- [16] D. Jayaraman and K. Grauman. Zero-shot recognition with unreliable attributes. In *NIPS*, 2014. **6**
- [17] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei. Novel dataset for fine-grained image categorization. In *CVPR*, 2011. **6**
- [18] E. Kodirov, T. Xiang, Z. Fu, and S. Gong. Unsupervised domain adaptation for zero-shot learning. In *ICCV*, 2015. **3, 5, 6**
- [19] 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. **1, 2**
- [20] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *T-PAMI*, 2014. **1, 3, 6**
- [21] W. Li, L. Niu, and D. Xu. Exploiting privileged information from web data for image categorization. In *ECCV*, 2014. **2**
- [22] X. Li, Y. Guo, and D. Schuurmans. Semi-supervised zero-shot classification with label representation learning. In *ICCV*, 2015. **3, 6**
- [23] G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. In *ICML*, 2010. **5**
- [24] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013. **1, 3, 6**
- [25] L. Niu, J. Cai, and A. Veeraraghavan. Zero-shot learning via category-specific visual-semantic mapping. *arXiv preprint arXiv:1711.06167*, 2017. **3**
- [26] L. Niu, J. Cai, and D. Xu. Domain adaptive fisher vector for visual recognition. In *ECCV*, 2016. **3**
- [27] L. Niu, W. Li, and D. Xu. Multi-view domain generalization for visual recognition. In *ICCV*, 2015. **3**
- [28] L. Niu, W. Li, and D. Xu. Visual recognition by learning from web data: A weakly supervised domain generalization approach. In *CVPR*, 2015. **2, 6**
- [29] L. Niu, W. Li, and D. Xu. Exploiting privileged information from web dt for action nd event recognition. *IJCV*, 118(2):130–150, 2016. **2**
- [30] L. Niu, W. Li, D. Xu, and J. Cai. An exemplar-based multi-view domain generalization framework for visual recognition. *T-NNLS*, 2016. **3**
- [31] L. Niu, W. Li, D. Xu, and J. Cai. Visual recognition by learning from web data via weakly supervised domain generalization. *T-NNLS*, 28(9):1985–1999, 2017. **2**
- [32] L. Niu, X. Xu, L. Chen, L. Duan, and D. Xu. Action and event recognition in videos by learning from heterogeneous web sources. *T-NNLS*, 28(6):1290–1304, 2017. **2**
- [33] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *T-NV*, 2011. **6**
- [34] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. **1, 3, 6**
- [35] J. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. 1998. **5**
- [36] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010. **4**
- [37] B. Romera-Paredes and P. H. Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, 2015. **3, 6**
- [38] S. M. Shojaee and M. S. Baghshah. Semi-supervised zero-shot learning by a clustering-based approach. *arXiv preprint arXiv:1605.09016*, 2016. **6, 7**
- [39] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. **6**
- [40] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus. Training convolutional networks with noisy labels. *ICLR*, 2015. **2, 6**
- [41] B. Sun, J. Feng, and K. Saenko. Return of frustratingly easy domain adaptation. In *AAAI*, 2016. **6**

- [42] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, and S. Belongie. Building a bird recognition app and large scale dataset with citizen scientists. In *CVPR*, 2015. 3
- [43] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 6
- [44] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele. Latent embeddings for zero-shot classification. In *CVPR*, 2016. 6
- [45] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 6
- [46] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang. Learning from massive noisy labeled data for image classification. In *CVPR*, 2015. 2, 7, 8
- [47] X. Xu, T. Hospedales, and S. Gong. Transductive zero-shot action recognition by word-vector embedding. *IJCV*, 2017. 3, 6
- [48] Z. Xu, S. Huang, Y. Zhang, and D. Tao. Augmenting strong supervision using web data for fine-grained categorization. In *ICCV*, 2015. 1, 3
- [49] Z. Xu, S. Huang, Y. Zhang, and D. Tao. Webly-supervised fine-grained visual categorization via deep domain adaptation. *T-PAMI*, 2016. 3
- [50] Z. Xu, W. Li, L. Niu, and D. Xu. Exploiting low-rank structure from latent domains for domain generalization. In *ECCV*, 2014. 3
- [51] Z. Xu, L. Zhu, and Y. Yang. Few-shot object recognition from machine-labeled web images. *CVPR*, 2017. 2
- [52] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc.*, 2006. 4
- [53] Z. Zhang and V. Saligrama. Zero-shot recognition via structured prediction. In *ECCV*, 2016. 6
- [54] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva. Places: An image database for deep scene understanding. *arXiv preprint arXiv:1610.02055*, 2016. 6
- [55] B. Zhuang, L. Liu, Y. Li, C. Shen, and I. Reid. Attend in groups: a weakly-supervised deep learning framework for learning from web data. *CVPR*, 2017. 2