Fine-grained Classification using Heterogeneous Web Data and Auxiliary Categories
Li Niu, Ashok Veeraraghavan, and Ashu Sabharwal

Abstract—Fine-grained classification, which aims to distinguish the subtle difference among various fine-grained categories belonging to one coarse-grained category, remains a very challenging problem, because of the absence of well-labeled training data caused by the high cost of annotating a large number of fine-grained categories. In the extreme case, given a set of test categories without any well-labeled training data, the majority of existing works can be grouped into the following two research directions: 1) crawl noisy labeled web data for the test categories as training data, which is dubbed as webly supervised learning; 2) transfer the knowledge from auxiliary categories with well-labeled training data to the test categories, which corresponds to zero-shot learning setting. Nevertheless, the above two research directions have many limitations and critical issues to be addressed. The first direction, webly supervised learning, often suffers from domain shift. The second direction, zero-shot learning, is struggling to achieve compelling results compared with conventional supervised learning. The issues of the above two directions motivate us to develop a novel approach which can jointly exploit both noisy web training data from test categories and well-labeled training data from auxiliary categories. To this end, we propose a new learning scenario for fine-grained classification based on heterogeneous data, which contains auxiliary categories that are manually designed by human experts, and word vectors, which can be obtained via free online corpus (e.g., Wikipedia) [10]. However, the performance of ZSL is still far below that of conventional supervised learning [10], especially when using free semantic information such as word vector.

I. INTRODUCTION

Recently, the field of image classification is greatly fueled by the rapid development in deep learning techniques and large-scale image datasets such as ImageNet [1]. However, fine-grained image classification, which targets at classifying abundant fine-grained categories belonging to one coarse-grained category (e.g., bird species and dog breeds), is still a very tough task. To identify the minor distinction among various fine-grained categories, sufficient well-labeled training images are in high demand. However, accurate human annotation for fine-grained categories is not easy to acquire due to the following reasons: 1) fine-grained annotation generally requires expertise, which raises the bar for human annotators; 2) there are usually myriads of fine-grained categories belonging to one coarse-grained category (e.g., in total 14,000 known bird species [2]) and hence it is infeasible to collect well-labeled training images for all fine-grained categories exhaustively. Therefore, the absence of well-labeled training images is a vital issue for fine-grained classification. In this work, we take an extreme case into consideration, that is, there are not any well-labeled training images for a given set of test categories. In this circumstance, the existing research works mainly fall into two research realms, i.e., Webly Supervised Learning (WSL) and Zero-Shot Learning (ZSL) [3], which will be elaborated separately in the following.

For Webly Supervised Learning (WSL), freely available web images are crawled from public websites (e.g., Flickr and Google) using category names as queries. Nevertheless, web images are loosely labeled, which means that their labels are very noisy and often inaccurate [4], [5]. When the classifier is learnt based on noisy training images, its performance on the test set will be significantly degraded. Moreover, when images are uploaded to public websites, they are often edited or compressed, leading to the dramatic data distribution mismatch between web images and test images, which is also referred to as domain shift [6]. Up to date, some works [2], [7] have been proposed for fine-grained classification based on web data, which tend to address the above two issues: label noise and domain shift. However, they rely on strong supervision (e.g., bounding box and part location) on web images or human intervention when collecting web images, which is often inaccessible in the real-world applications.

For Zero-Shot Learning (ZSL) [3], training categories (i.e., seen categories) and test categories (i.e., unseen categories) have no overlap. In other words, with training instances from seen categories, we need to recognize the test instances from unseen categories. To achieve this goal, intermediate category-level semantic information of all categories is used to bridge the gap between seen categories and unseen categories. There are various forms of category-level semantic information including attributes [3] (e.g., shape, color, and material), which are manually designed by human experts, and word vectors [8], [9] corresponding to category names, which can be obtained via free online corpus (e.g., Wikipedia) [10]. However, the performance of ZSL is still far below that of conventional supervised learning [10], especially when using free semantic information such as word vector.

To this end, we propose a new learning scenario for fine-grained classification, which unifies webly supervised learning and zero-shot learning. In particular, given a set of fine-grained categories without any well-labeled training images, we crawl noisy web images for these categories as training data and also utilize the well-labeled images from auxiliary...
fine-grained categories. From another point of view, given the entire set of fine-grained categories belonging to one coarse-grained category, we only need to ask human experts to annotate partial fine-grained categories, and then can predict the rest of fine-grained categories with the aid of web data. Therefore, our proposed learning scenario could be viewed as webly supervised learning with well-labeled data from auxiliary categories, or zero-shot learning with noisy web training data for unseen categories.

We develop our method in this learning scenario. Given a set of auxiliary categories and a set of test categories, we first crawl noisy web training images for test categories using the category names as queries. Then, deep visual features are extracted for all images including web training images, well-labeled training images, and unlabeled test images. Furthermore, intermediate category-level semantic representations are extracted for all categories. In practice, we use word vector [8], [9] as category-level semantic representation. To be exact, we train a linguistic model based on free online corpus (e.g., Wikipedia) and obtain the word vector corresponding to each category name. Finally, visual features and word vectors are fed into our learning model, yielding the prediction results of test images.

Another benefit of learning from web data is that web images are generally associated with additional textual information such as titles and tags, while test images do not have such textual information. The information which is only available for training instances but not available for test instances is dubbed as privileged information [11]. The privileged information can be utilized in the training stage to help learn a more robust model, so we take advantage of additional textual information as privileged information in our method. Particularly, we extract textual features from the surrounding textual information of web images, which are fed into our learning model together with visual features and word vectors.

The flowchart of our method is illustrated in Fig. 1. With visual features, textual features, and word vectors as input, our learning model can cope with the label noise and domain shift of web images, transfer the knowledge from auxiliary categories to test categories, and simultaneously take advantage of the privileged information. It is obvious that our method involves different types of web data with different functions: 1) web images are crawled from public websites (e.g., Flickr and Google) for test categories as noisy training data; 2) the surrounding textual information of web images are used as privileged information; 3) free online corpus (e.g., Wikipedia) is used to obtain semantic representations of all categories for the sake of filling in the gap between test categories and auxiliary categories. The details of our learning model will be fully introduced in Section III-C.

Our major contributions are fourfold: 1) as far as we are concerned, we propose the first learning scenario for fine-grained classification using both web data and auxiliary categories; 2) in this learning scenario, we propose a novel method unifying zero-shot learning and webly supervised learning, which can transfer the knowledge from auxiliary categories to test categories and simultaneously handle the label noise and domain shift of web data; 3) we further extend our method by using the surrounding textual information of web images as privileged information; 4) the effectiveness of our methods is verified by comprehensive experiments on three benchmark datasets.

This paper extends the preliminary conference version [12] in the following ways. From the theoretical aspect, we extend our model by using additional textual information as privileged information in Section III-C followed by an effective solution in Appendix B. From the experimental aspect, we evaluate the methods using privileged information in Section IV-B and also provide more quantitative and qualitative analyses in Section IV-A.

II. RELATED WORK

In this section, we will discuss some recent works on webly supervised learning and zero-shot learning. Moreover, as domain shift needs to be addressed when learning from web data, domain adaptation will also be briefly introduced. Finally, we will describe previous works using privileged information, since additional textual information is used as privileged information in our method.

Webly Supervised Learning: Learning from web data [13], [14], [15], [16], [4], [17], [18], which is also dubbed as webly supervised learning, has attracted more and more research...
interest from the computer vision community. Lots of works have been done to cope with label noise and domain shift when learning from web data. As a pioneering work, NEIL in [13] uses Multi-Instance Learning (MIL) to suppress the label noise of web images. Besides, a couple of domain adaptation methods were studied in [19] while a weakly supervised domain generalization method was developed in [20]. With the advance of deep learning techniques, several CNN approaches also emerged for learning from web images [16], [21], [4], [22], [23], [5]. Nevertheless, fine-grained classification is out of focus of the above works.

With regards to exploiting web data for fine-grained classification, label noise is mitigated in [2] via active learning, which in fact involves human intervention. In [7], bounding box annotations are needed for web data. The work in [24] leverages Flickr images to learn bird classifiers, whereas human annotators take part in the dataset collection. In a more recent work [25], label noise and domain shift are handled when learning from web images, but bounding boxes and part landmarks are required, which are unavailable in our application. Distinguished from all the above works, we focus on weaely supervised fine-grained classification without human intervention when collecting web data or strong supervision (e.g., bounding box and part location) on web data.

**Zero-Shot Learning:** Recently, many Zero-Shot Learning (ZSL) methods have been proposed [26], [10], [3], [27]. Moreover, several works [28], [29], [30] found that it is helpful to use unlabeled test instances from unseen categories in the training process, leading to semi-supervised/transductive ZSL. Inspired by semi-supervised/transductive ZSL, we also utilize unlabeled test instances in the training stage. For more details of existing ZSL approaches, please refer to two recent surveys [31], [32]. However, all the above ZSL methods ignore the large amount of freely available web images, which could be used for fine-grained classification. In contrast, we propose a new learning scenario, which treats seen categories as auxiliary categories and additionally exploit web images for test categories.

**Domain Adaptation:** Domain adaptation targets at addressing the domain shift between training set (i.e., source domain) and test set (i.e., target domain) by reducing the domain distribution mismatch between two domains. Analogous to semi-supervised/transductive ZSL, for domain adaptation, unlabeled test instances are also used in the training stage. Existing domain adaptation approaches can be roughly grouped into reweight based method [33], subspace based method [34], [55], [56], [47], [48], and generative model based method [39], [40]. Recently, many domain generalization methods [41], [42], [43], [44] have been proposed. Different from domain adaptation, domain generalization focuses on the setting in which the unlabeled test instances are unavailable in the training stage. The closest related work is the reweight based method [33], which assigns different weights on different training instances on the premise of Maximum Mean Discrepancy (MMD). However, the work in [33] only focuses on domain shift and does not favor our application.

**Learning using Privileged Information:** Moreover, our work is relevant to learning using privileged information (LUPI) in [45], in which training instances are associated with additional information (i.e., privileged information) that are not available for test instances. The key idea proposed in [45] is using privileged information to model the classification difficulty of training instances, which has been generalized to a wide range of applications such as ranking [46], clustering [47], metric learning [48], Gaussian process [49], hashing [50], and multi-instance learning [51]. However, all these works do not use privileged information in weaely supervised learning. The closest related works are [13], [20], which take advantage of textual information as privileged information when learning from web data. However, they do not borrow the idea of auxiliary categories from zero-shot learning. In contrast, we integrate privileged information into our method, which unifies weaely supervised learning and zero-shot learning.

### III. Our Learning Model

Recall that the flowchart of our method has been illustrated in Fig. 1 and described in Section 1 which will be recapped in the following. Given a set of auxiliary categories with well-labeled training images and a set of test categories, we crawl web training images with their surrounding textual information for test categories and obtain semantic representations (i.e., word vectors [8], [9]) for all categories. Then, we feed the visual features of all training images (i.e., well-labeled training images and web training images), the textual features of additional textual information, and word vectors of all categories into our learning model. Besides, we also feed unlabeled test images into our learning model in the training phase, which can help address domain shift and facilitate zero-shot learning [28], [29], [30], resulting in a semi-supervised learning model. The output of our learning model is the predicted semantic representations of test images, which are used for final categorization. Next, we will introduce how to transfer the knowledge from auxiliary categories to test categories, how to exploit web images from test categories, and how to take advantage of the additional textual information as privileged information.

In the remainder of this paper, for ease of representation, a matrix/vector is denoted by an uppercase/lowercase letter in boldface. \( A^{-1} \) is used to denote the inverse matrix of \( A \). Moreover, we adopt superscript ‘ to represent the transpose of a vector/matrix. We use \( O \) (resp., \( I \)) to denote the zero matrix (resp., identify matrix). Similarly, we use \( 1 \) to denote the vector with all ones. \( A \circ B \) (resp., \( (A,B) \)) is used to denote the element-wise product (resp., inner product) of two matrices.

#### A. Knowledge Transfer from Auxiliary Categories to Test Categories

We use \( X^a \in \mathbb{R}^{d \times n^a} \) to denote the visual features of well-labeled training images from \( C^a \) auxiliary categories, in which \( n^a \) is the number of training images and \( d \) is the dimension of visual feature. Similarly, we use \( X^t \in \mathbb{R}^{d \times n^t} \) to denote the visual features of test images from \( C^t \) test categories, in which \( n^t \) is the number of test images. Suppose that each category is associated with a \( m \)-dim semantic representation, the semantic
representation matrix of test (resp., auxiliary) categories is represented as \( A^t \in \mathbb{R}^{m \times C^t} \) (resp., \( A^a \in \mathbb{R}^{m \times C^a} \)). Then, the semantic representation matrix of well-labeled training data is represented as \( A^a \in \mathbb{R}^{m \times n^a} \), in which each column is the semantic representation of the category that each training instance belongs to. In a similar way, the semantic representation matrix of test data is given by \( A^t \in \mathbb{R}^{m \times n^t} \), which needs to be learnt. After obtaining \( A^t \), the category labels of test data can be inferred by comparing \( A^t \) with \( A^a \).

With the aim to transfer knowledge from auxiliary categories to test categories, inspired by \([29]\), we learn one visual-semantic dictionary \( D^a \in \mathbb{R}^{d \times m} \) (resp., \( D^t \in \mathbb{R}^{d \times m} \)) for auxiliary categories (resp., test categories), expecting \( D^a \) and \( D^t \) to be close to each other using a co-regularizer \( \| D^t - D^a \|_F^2 \).

The visual-semantic dictionary \( D^t \) (resp., \( D^a \)) maps from semantic representation space to visual feature space with the mapping error \( \| X^t - D^t A^t \|_2^2 \) (resp., \( \| X^a - D^a A^a \|_2^2 \)). It is worth mentioning that we adopt dictionary learning based method because it lays the foundation for unifying WSL and ZSL in a coherent manner.

The knowledge transfer has two steps. In the first step, the dictionary of auxiliary categories \( D^a \) is learnt as follows,

\[
\min_{D^a} \quad \frac{1}{2} \| X^a - D^a A^a \|_F^2 + \frac{1}{2} \| D^a \|_F^2, \tag{1}
\]

in which \( \| X^a - D^a A^a \|_F^2 \) is the mapping error of well-labeled training images from auxiliary categories and \( \| D^a \|_F^2 \) is a penalty term controlling the complexity of \( D^a \).

In the second step, we aim to learn the dictionary of test categories \( D^t \) and semantic representations of test instances \( A^t \). Specifically, we minimize the mapping error \( \| X^t - D^t A^t \|_F^2 \) on the test images, similar to the first step. Besides, we enforce \( D^t \) to be close to \( D^a \) based on \( \| D^t - D^a \|_F^2 \), in which way the knowledge of auxiliary categories can be transferred to test categories. Moreover, considering that the semantic representations of the test instances belonging to the same category should be similar to each other, we expect \( A^t \) to be low-rank, which encourages the similar semantic representations of test instances to be grouped together. To achieve this goal, we bring in a nuclear norm \([52]\) regularizer \( \| A^t \|_* \), which enforces \( A^t \) to be low-rank. To this end, the objective function of the second step is given by

\[
\min_{D^t, A^t} \quad \frac{1}{2} \| X^t - D^t A^t \|_F^2 + \frac{\lambda_1}{2} \| D^t - D^a \|_F^2 + \lambda_2 \| A^t \|_*, \tag{2}
\]

in which \( \lambda_1 \) and \( \lambda_2 \) are trade-off parameters.

### B. Exploiting Web Images from Test Categories

Besides well-labeled training images from auxiliary categories, we additionally crawl web images by using \( C^t \) test category names as queries to form the web training set. The visual features of web images are represented as \( X^w \in \mathbb{R}^{d \times n^w} \), in which \( n^w \) is the number of web images. Analogous to \( A^a \), the semantic representation matrix of web images is given by \( A^w \in \mathbb{R}^{m \times n^w} \), in which each column is the semantic representation of the category that each web image is associated with (the category labels of web images may be inaccurate). Since web images and test images are from the same set of test categories, the same dictionary \( D^t \) is applied to the web images, leading to the mapping error of web images \( \| X^w - D^t A^w \|_F^2 \). Recall that we need to address two issues when learning from web images: label noise and domain shift, which will be detailed next.

To account for the label noise of web images, we replace the mapping error, i.e., Frobenius norm regularizer \( \| X^w - D^t A^w \|_F^2 \), with \( L_{2,1} \) norm regularizer \( \| X^w - D^t A^w \|_{2,1} \). The \( L_{2,1} \) norm of a matrix \( X \) is defined as \( \sum_{i} \| x_i \|_2 \), in which \( x_i \) is each column in \( X \). \( L_{2,1} \) norm encourages column-sparsity \([53]\), that being said, some columns have exceeding zero entries. After employing \( L_{2,1} \) norm, \( \| X^w - D^t A^w \|_{2,1} \) is expected to be column-sparse. The columns with exceeding non-zero (resp., zero) entries correspond to the outliers (resp., non-outliers), which is granted larger (resp., smaller) tolerance of error. In this way, we suppress the label noise of web images and learn a more robust dictionary \( D^t \) on test categories.

To account for the domain shift between web images (i.e., \( X^w \)) and test images (i.e., \( X^t \)), we employ an Maximum Mean Discrepancy (MMD) \([33]\) based regularizer \( \| \frac{1}{n^w} X^w \theta - \frac{1}{n^t} X^t \theta \|_2^2 \) with the weight vector \( \theta \) to be learnt. The idea of MMD-based regularizer is to reduce the distance between the center of weighted web images (i.e., \( \frac{1}{n^w} X^w \theta \)) and the center of test images (i.e., \( \frac{1}{n^t} X^t \)). This is accomplished by assigning higher weights on the web images which are closer to the center of test images. In other words, we identify the web images which are more likely to be sampled from the data distribution of test images, by learning the weight vector \( \theta \).

To take full advantage of the weight vector \( \theta \), we expect to identify the web images with not only closer distribution to test images but also relatively accurate labels. Specifically, besides the MMD-based regularizer \( \| \frac{1}{n^w} X^w \theta - \frac{1}{n^t} X^t \theta \|_2^2 \), we also employ the weights \( \theta \) in the \( L_{2,1} \) norm regularizer \( \| X^w - D^t A^w \|_{2,1} \), in which we use a diagonal matrix \( \Theta \) with the diagonal being \( \theta \) for ease of representation. By minimizing \( \| X^w - D^t A^w \|_{2,1} \), lower (resp., higher) weights are prone to be assigned to the columns of \( X^w - D^t A^w \) with exceeding non-zero (resp., zero) entries, which correspond to the outliers (resp., non-outliers). In this way, we collaboratively account for label noise and domain shift with the importance weight vector \( \theta \) shared by two regularizers. In contrast, most existing works address these two issues separately.

From another perspective, since the dictionary \( D^t \) used in \( \| (X^w - D^t A^w) \Theta \|_{2,1} \) is enforced to be close to the dictionary of auxiliary categories \( D^a \), auxiliary categories actually assist in dealing with the label noise of web images. To this end, we extend \([4]\) by using web images as follows,

\[
\min_{D^a, A^t, \theta} \quad \frac{1}{2} \| X^t - D^t A^t \|_F^2 + \frac{\lambda_1}{2} \| D^t - D^a \|_F^2 + \frac{\lambda_2}{2} \| A^t \|_* + \frac{\lambda_3}{2} \sum_{i} \| x_i \|_2 + \lambda_4 \| (X^w - D^t A^w) \Theta \|_{2,1}, \tag{3}
\]

s.t. \( 1^\theta = n^w, \quad 0 \leq \theta \leq b^1 \),

where \( \lambda_3, \lambda_4, \) and \( b \) are newly introduced trade-off parameters. Note that we impose a sum constraint and a box constraint on \( \theta \) in \([4]\), in which \( b \) is the upper bound of importance weights. The problem in \([3]\) is nontrivial to solve because of the \( L_{2,1} \) norm and low-rank regularizer. Hence, a novel solution is
developed based on inexact Augmented Lagrange Multiplier (ALM) [24]. For ease of optimization, an intermediate variable $Z^t \ (resp., \ E^w)$ is introduced to replace $A^t$ in $\|A^t\|$ (resp., $(X^w - D^t A^w)\Theta$) in (3). At the same time, we enforce $Z^t \ (resp., \ E^w)$ to be close to $A^t$ (resp., $(X^w - D^t A^w)\Theta$). Then, the problem in (3) can be rewritten as

$$\min_{D^t, A^t, \theta} \frac{1}{2}\|X^t - D^t A^t\|^2_F + \frac{\lambda_1}{2}\|D^t - D^a\|^2_F + \lambda_2\|Z^t\|_*$

$$+ \frac{\lambda_3}{n_w}\|X^w\theta - \frac{1}{n_t}\|X^t1\|^2 + \lambda_4\|E^w\|_{2,1}, \quad (5)$$

s.t. $1'\theta = n_w, \ 0 \leq \theta \leq b_1,$

$$E^w = (X^w - D^t A^w)\Theta,$$

$$Z^t = A^t.$$  

Then, after introducing the Lagrangian multiplier $R \ (resp., \ T)$ for the constraint in (6) (resp., (7)), we tend to minimize the augmented Lagrangian form of (5):

$$\min_{D^t, A^t, \theta, R, T} \frac{1}{2}\|X^t - D^t A^t\|^2_F + \frac{\lambda_1}{2}\|D^t - D^a\|^2_F + \lambda_2\|Z^t\|_*$$

$$+ \frac{\lambda_3}{n_w}\|X^w\theta - \frac{1}{n_t}\|X^t1\|^2 + \lambda_4\|E^w\|_{2,1}$$

$$+ \frac{\mu}{2}\|E^w - (X^w - D^t A^w)\Theta + (R, E^w - (X^w - D^t A^w)\Theta)$$

$$+ \frac{\mu}{2}\|A^t - Z^t\|_F + \langle T, A^t - Z^t \rangle,$$  

(8)

in which $S = \{\theta | 1'\theta = n_w, 0 \leq \theta \leq b_1\}$ is the feasible set of $\theta$, and $\mu$ is a penalty parameter. We update the variables $\{E^w, Z^t, D^t, A^t, \theta\}$, the Lagrangian multipliers $\{R, T\}$, and the penalty parameter $\eta$ iteratively until the termination criterion is satisfied. The technical details of updating these variables are left to Appendix A. By minimizing (8), we can acquire the semantic representations of test instances $\tilde{A}^t$.

With the semantic representations of test images $A^t$ and test categories $\tilde{A}^t$, we use nearest neighbor (NN) classifier for final category prediction, following the strategy in [29]. In particular, we compare the semantic representation of each test instance (i.e., each column in $\tilde{A}^t$) with that of each test category (i.e., each column in $A^t$), and each test instance is assigned to the nearest test category.

C. Extension with Privileged Information

As mentioned in Section 3, web images are usually associated with additional textual information, which is unavailable for test images. The additional information which is only available for the training data but not available for the test data is referred to as privileged information. Learning using privileged information was initially proposed in [11]. In particular, they extend SVM to SVM+ by using privileged information, in which the slack variable $\xi_i$ in SVM is replaced by a slack function $\xi(\tilde{x}_i)$ based on the privileged information $\tilde{x}_i$. Formally, the objective function of SVM+ is

$$\min_{w, b, \tilde{w}, \tilde{b}} \frac{1}{2}\|w\|^2 + \frac{1}{2}\|\tilde{w}\|^2 + \sum_{i=1}^n \xi(\tilde{x}_i) \quad (9)$$

s.t. $y_i(wx_i + b) \geq 1 - \xi(\tilde{x}_i), \ \forall i,$

$$\xi(\tilde{x}_i) \geq 0, \ \forall i,$$

$$\xi(\tilde{x}_i) = \tilde{w}^T\tilde{x}_i + \tilde{b}, \ \forall i. \quad (10)$$

where $w \ (resp., \ \tilde{w})$ and $b \ (resp., \ \tilde{b})$ are the weight vector and bias of classification (resp., slack) function based on primal feature $x_i$ (resp., privileged information $\tilde{x}_i$), and $y_i$ is the binary label. Recall that in conventional SVM, the slack variable $\xi_i$ is used to model the difficulty of classifying each training instance. In analogy to $\xi_i$, SVM+ relies on the slack function $\xi(\tilde{x}_i)$ to model the classification difficulty, in which privileged information $\tilde{x}_i$ plays a role of teacher in the training process. To be exact, if one training instance is difficulty to classify, the value of its slack function $\xi(\tilde{x}_i)$ is allowed to be very large for tolerance of error. Otherwise, the value of its slack function $\xi(\tilde{x}_i)$ is enforced to be very small. Note that $\tilde{w}$ and $\tilde{b}$ are automatically learnt by solving (9).

Inspired by SVM+, we also use the slack function based on privileged information to model the mapping difficulty. Recall that in our problem, we learn a mapping $D^t$ on web images to map from semantic space $A^w$ to visual space $X^w$, leading to the mapping error $X^w - D^t A^w$, which stands for the mapping difficulty of each web image. Similar to SVM+, we tend to use privileged information (i.e., textual information) to approximate the mapping error. After denoting the aggregated textual feature of web images as $\tilde{X}^w \in \mathcal{R}^{d \times n_w}$ with $d$ being the dimensionality of textual feature, we learn a slack function based on $\tilde{X}^w$ to estimate the mapping error. Simply, we learn a matrix $W \in \mathcal{R}^{d \times d}$ and enforce $WX^w$ to be close to the mapping error $X^w - D^t A^w$. On one hand, the slack function $WX^w$ allows the web training images with great mapping difficulty to have large mapping error. On the other hand, when the number of training instances is not very large, the slack function can help avoid over-fitting by regulating the mapping error $\tilde{X}^w$.

To this end, we actually control the mapping error in two ways: 1) assign different weights $\Theta$ on the mapping error of different web images; 2) use additional textual information to approximate the mapping error of each web image. After adding a new regularizer $\|X^w - D^t A^w\|_F^2$ to (5), our new objective function can be written as

$$\min_{D^t, A^t, \theta, W} \frac{1}{2}\|X^t - D^t A^t\|^2_F + \frac{\lambda_1}{2}\|D^t - D^a\|^2_F + \lambda_2\|Z^t\|_*$$

$$+ \frac{\lambda_3}{n_w}\|X^w\theta - \frac{1}{n_t}\|X^t1\|^2 + \lambda_4\|E^w\|_{2,1}$$

$$+ \frac{\mu}{2}\|X^w - D^t A^w\|_F^2,$$  

(11)

s.t. $1'\theta = n_w, \ 0 \leq \theta \leq b_1,$

$$E^w = (X^w - D^t A^w)\Theta,$$

$$Z^t = A^t,$$

in which $\gamma$ is a trade-off parameter. Compared with (5), we have one more variable $W$ to learn. The problem in (11) can be solved similarly to (5) except updating $D^t$ and $W$. We leave the technical details of solving (11) to Appendix B. After solving (11), we obtain the semantic representations of test instances $\tilde{A}^t$ and the testing procedure is the same as in Section III-B. Note that we do not need textual information of test instances in the testing stage.
IV. EXPERIMENTS

In this section, we evaluate our methods for fine-grained image classification with or without privileged information on three benchmark datasets. Besides, we conduct extra experiments under the generalized setting in which the test instances may come from both auxiliary categories and test categories. Moreover, we also provide adequate ablation study and qualitative analysis.

A. Fine-grained Image Classification

Datasets: Experiments are conducted on the following three datasets which are popular in the zero-shot learning (ZSL) community, since our learning scenario can be treated as ZSL with additional web training images for unseen categories, as mentioned in Section I.

1) CUB [55]: Caltech-UCSD Bird (CUB) consists of 11,788 images from 200 bird species. Following [56], we adopt the standard train-test split with 150 auxiliary (resp., 50 test) categories.

2) SUN [57]: In Scene UNDERstanding (SUN) attribute dataset, each scene category has 20 images. Following [58], we adopt the standard train-test split with 707 auxiliary (resp., 10 test) categories.

3) Dogs [59]: Stanford Dogs dataset is composed of 19,501 images distributed in 113 dog breeds. We use the train-test split provided in [10], i.e., 85 auxiliary (resp., 28 test) 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 test categories as queries to collect the top ranked 100 images from Flickr website for each category after performing PCA-based near-duplicate removal [60].

Visual Features and Semantic Representations: We extract visual features for all images and semantic representations for all categories.

1) Visual features: For each image, we use 4,096-dim output of the 6-th layer of VGG [61] model pretrained on ImageNet dataset as its visual feature.

2) Semantic representations: We employ two types of word vectors: GloVe [9] and Word2Vec [8], in which each word is associated with a real-valued vector. We train Word2Vec and GloVe linguistic models based on the latest Wikipedia corpus, with the dimension of word vector set as 400. Then, for each category, two word vectors corresponding to the category name from Word2Vec and GloVe models are concatenated as the category-level semantic representation, yielding an 800-dim vector. When one category has the name with more than one word, we simply use the average of the semantic representations corresponding to all words as its final semantic representation.

Baselines: We compare our approach with three sets of baselines: WSL baselines, ZSL baselines, and domain adaptation (DA) baselines. To the best of our knowledge, no existing method 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 as the combo baseline. Intuitively, the combo baseline should be the strongest baseline because it utilizes both web images and auxiliary categories.

For WSL baselines, we compare with NEIL [13], Bergamo and Torresani [19], WSDG [20], sukhaabtaar et al. [21], and Xiao et al. [16]. Note that Xiao et al. [16] leverages manually cleaned web data when training network and computing confusion matrix, which is not available in our application. Therefore, for fair comparison, we run [16] without using manually cleaned web data in the training process and estimate the confusion matrix based on semantic representations.

For ZSL baselines, we compare with the standard ZSL methods ESZSL [26], LatEm [62], SJE [10], DAP/IAP [3], Changpinyo et al. [63], Zhang and Koniusz [66] as well as transductive/semi-supervised ZSL methods Li et al. [30], Kodirov et al. [29], Zhang and Saligrama [64], Xu et al. [28], Shojaee and Baghshah [65], SE-GZSL [67] as baselines. The difference between transductive/semi-supervised ZSL approaches and standard ZSL approaches lies in whether unlabeled test data are available in the training phase.

For DA baselines, we compare with KMM [33], GFK [34], SA [55], TCA [35], and CORAL [37], in which web training images and test images are regarded as the source domain and the target domain respectively.

For the combo baseline, we select the most competitive WSL baseline [16] and ZSL baseline [64] based on their mean performance on three datasets, and average their test decision values, which is referred to as WSL+ZSL in Table I.

We also compare with one basic baseline LR, which simply learns a linear regressor based on web training images. Besides, to validate the WSL and ZSL components in our

| Dataset | CUB | SUN | Dogs | Avg |
|---------|-----|-----|------|-----|
| LR      | 68.39 | 62.50 | 77.67 | 69.52 |
| KMM     | 70.54 | 64.00 | 79.16 | 71.23 |
| GPR     | 70.37 | 62.50 | 79.51 | 70.79 |
| SA      | 68.67 | 63.00 | 80.18 | 70.62 |
| TCA     | 68.56 | 63.00 | 80.22 | 70.59 |
| CORAL   | 69.04 | 63.50 | 80.37 | 70.97 |
| NEIL    | 69.08 | 63.00 | 80.16 | 70.74 |
| Bergamo and Torresani | 70.13 | 64.00 | 78.64 | 70.93 |
| WSDG    | 70.61 | 66.00 | 80.20 | 72.27 |
| Sukhaabtaar et al. | 70.47 | 64.50 | 81.15 | 72.04 |
| Xiao et al. | 70.92 | 65.50 | 81.67 | 72.69 |
| ESZSL   | 38.08 | 65.00 | 37.21 | 46.77 |
| LatEm   | 35.15 | 66.50 | 35.99 | 45.88 |
| SJE     | 42.65 | 71.50 | 34.85 | 49.67 |
| DAPI/AAP | 28.91 | 57.50 | 33.15 | 39.85 |
| Changpinyo et al. | 41.83 | 72.00 | 39.91 | 51.25 |
| Li et al. | 32.36 | 72.50 | 43.15 | 49.34 |
| Kodirov et al. | 47.53 | 71.00 | 47.32 | 55.28 |
| Zhang and Saligrama | 44.08 | 76.50 | 48.09 | 56.23 |
| Xu et al. | 45.72 | 71.50 | 39.85 | 52.36 |
| Shojaee and Baghshah | 46.68 | 71.00 | 48.82 | 55.50 |
| Zhang and Koniusz | 47.66 | 73.00 | 43.98 | 54.88 |
| SE-GZSL | 46.84 | 75.00 | 43.91 | 55.25 |
| 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 | 80.04 | 80.42 |
| Ours_sim2 | 76.00 | 79.50 | 83.75 | 79.75 |
| Ours | 76.47 | 84.50 | 85.16 | 82.04 |
method (3), we report the results of our two special cases. Particularly, we remove the regularizer related to knowledge transfer (i.e., $\|D^t - D^w\|^2_F$) by setting $\lambda_1$ as 0, and refer to this special case as Ours_WSL. Similarly, we remove the regularizers using web data (i.e., $\frac{1}{n}X^w \theta - \frac{1}{1}X^1 \theta^2$ and $\|X^w - D^w A^w \|_2$) by setting $\lambda_3$ and $\lambda_4$ as 0, and this special case is referred to as Ours_ZSL. Moreover, to validate some regularizers individually in our method (3), we further compare with our two simplified versions. Specifically, we remove the regularizer $\|A^t\|$, (resp., $\|\frac{1}{n}X^w \theta - \frac{1}{1}X^1 \theta^2\|$) in (3) by setting $\lambda_2$ (resp., $\lambda_3$) as 0 and refer to this simplified version as Ours_sim1 (resp., Ours_sim2). For all methods, we use multi-class accuracy as the evaluation metric.

**Parameters:** Our method has trade-off parameters $b$, $\lambda_1$, $\lambda_2$, $\lambda_3$, and $\lambda_4$ in (3), which are determined by using the cross-validation strategy. In particular, following [65], we select the first $C^a$ categories according to default category indices from $C^a$ auxiliary categories as the validation categories, with $C^c$ satisfying $C^c = \frac{C^a}{C^a - C^c}$. It is worth mentioning that we need to additionally crawl web images for validation categories in order to use the cross-validation strategy. In the validation stage, we use $C^c$ categories as test categories and $C^a - C^c$ categories as auxiliary categories. Then, the optimal trade-off parameters are determined according to the validation performance through random search [68] within certain range. To be exact, we empirically traverse the parameters $\lambda_1$, $\lambda_2$, $\lambda_3$, and $\lambda_4$ within the range $[10^{-5}, 10^{-2}, \ldots, 10^3]$, and traverse the parameter $b$ within the range $[1.5, 2.0, \ldots, 5.0]$. The range of $b$ is explained as follows. The upper bound of importance weights should be larger than one yet not too large based on the mild assumption that no web training image is far more important than others.

In fact, our method is relatively robust when setting the trade-off parameters within certain range. By taking the Dogs dataset as an example, we explore the performance variation of our approach w.r.t. one parameter while the other parameters remain fixed as their optimal values. It can be seen from Fig. 2 that our method is relatively robust when varying one parameter within the range $[10^{-5}, 10^{-2}, \ldots, 10^3]$ while fixing the other parameters as their optimal values.

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

1. The WSL and DA baselines are better than LR, which shows the advantage of coping with label noise or domain shift. The ZSL baselines are worse than DA/WSL baselines on CUB and Dogs datasets, but generally better on the SUN dataset. There is no consistent winner between ZSL baselines and DA/WSL baselines because their performance highly depends on the purity of web images as well as the relation between auxiliary categories and test categories.

2. The transductive/semi-supervised ZSL methods [30], [29], [64], [28], [65], [67] generally perform more favorably than the standard ZSL methods [26], [62], [10], [3], [63], [66], which demonstrates that it is useful to include unlabeled test data in the training phase for ZSL.

3. Our method outperforms Ours_WSL and Ours_ZSL, which indicates the benefit of unifying WSL and ZSL. Our method also achieves better results than Ours_sim1 and Ours_sim2, which proves the effectiveness of our low-rank and MMD-based regularizers.

4) It is worth mentioning that the focus of this paper is a new learning scenario for fine-grained classification by using both web data and auxiliary categories, instead of a state-of-the-art WSL or ZSL approach. So there is no guarantee that Ours_WSL (resp., Ours_ZSL) can perform better than all WSL (resp., ZSL) baselines. However, when utilizing both web data and auxiliary categories, our method achieves significant improvement over the combo baseline WSL+ZSL, which shows that a naive combination can hardly take full advantage of both web data and auxiliary categories. In contrast, we unify ZSL and WSL coherently in our method, which significantly advances fine-grained image classification.

**Computational Efficiency:** By taking the CUB dataset as an example, we compare the running time of two special cases of our method (i.e., Ours_WSL and Ours_ZSL from Table 1) and their naive Combo (average $A^t$ for prediction). The running time of Combo is the sum of running time of two special cases. We run all methods on the same server with Intel Xeon 3.33-GHz CPUs and 32-GB RAM in a single thread. The running time and accuracies of various approaches are reported in Table II from which we can observe that our method is more effective and efficient than Combo.

| Method | Ours_WSL | Ours_ZSL | Combo | Ours |
|--------|----------|----------|-------|------|
| Time (s) | 1630.35 | 819.26 | 2449.61 | 1953.92 |
| Accuracies (%) | 69.42 | 70.58 | 76.47 |

Utilizing More Web Images: Since we only use 100 web training images for each test category, it is interesting to explore whether the performance will keep increasing by using more web training images. We study the performance variation w.r.t. different numbers of web training images. Specifically, we crawl various numbers of web images for each test category (i.e., [100, 200, \ldots, 1000]) to construct the web training set while keeping the other experimental settings unchanged. The accuracies with various numbers of web training images on three datasets are plotted in Fig. 3 from which it can be seen that for the CUB and Dogs datasets, the accuracy increases as the number of web training images grows within certain range. On the contrary, for the SUN dataset, the accuracy drops dramatically as the number of web training images grows. One possible explanation is that scene category names are more ambiguous than the dog/bird names. Furthermore, the scene category names in the SUN datasets are accompanied by additional “in_door” or “out_door” label, rendering it even more difficult to crawl semantically correct web images.

**Qualitative Analysis of Learnt Weights $\theta$:** In our method (3), higher weights are assumed to be assigned to the web training images with closer distribution to test images and relatively accurate labels. Therefore, the web images with higher weights are more likely to be non-outliers and visually resembling test images. We take the Dogs dataset as an example and rank the web images based on the importance weight vector $\theta$ learnt...
WSL in a qualitative fashion, we
(b) (i) (b) (g) WSL.
(c) (c) (d) (d) (e) WSL in
with our special case Ours at all.
ambiguous category name or not relevant to the category name
partially occluded dogs while some images are misled by the
test categories. For instance, given the entire set of all fine-
the test instances may come from both auxiliary categories and
(e) In detail, some images contain fake (e.g., printing) dogs or
while in the images in Fig. 5, some objects are very small
images in the top row in Fig. 4 have dominant centered objects
θ weights based on the weight vector
additionally show the web images associated with
5. From Fig. 4, it can be seen that
the top row of images with highest weights have accurate
labels. Moreover, the dogs occupy the substantial center of
the entire image and visually resemble the test images. In
contrast, the web images in the bottom row are quite noisy.
In detail, some images contain fake (e.g., printing) dogs or
partially occluded dogs while some images are misled by the
ambiguous category name or not relevant to the category name
at all.
To demonstrate the superiority of our method compared
with our special case Ours_WSL in a qualitative fashion, we
additionally show the web images associated with 5 highest
weights based on the weight vector θ learnt by Ours_WSL in
Fig. 5. By comparing Fig. 5 and Fig. 4, we observe that the
images in the top row in Fig. 4 have dominant centered objects
while in the images in Fig. 5 some objects are very small
with cluttered background (e.g., (a) and (d)) or even not dogs
(e.g., (e)), which indicates the advantage of using auxiliary
categories to deal with the label noise of web images. We
have similar observations on the other two datasets.
Generalized ZSL Setting: In some real-world applications,
the test instances may come from both auxiliary categories and
test categories. For instance, given the entire set of all fine-
grained categories belonging to one coarse-grained 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. Then, we aim to recognize the test
images from all fine-grained categories (e.g., 14,000). This
generalized setting is more enticing yet more challenging.
In fact, in the absence of web training images, this gener-
alized setting reduces to the generalized Zero-Shot Learning
(ZSL) [69], in which the test instances may come from both
seen and unseen categories.
To investigate the effectiveness of our method under the
generalized setting, we conduct extra experiments with the
test instances from a mixture of auxiliary categories and test
categories. Under the generalized setting, our method in [3]
can be readily applied with a little abusively used dictionary
D′ in ∥X′ − D′A′∥2 for all the categories instead of only
test categories. After the acquisition of A′, similarly as in
Section III-B we adopt nearest neighbor (NN) classifier to
predict test instances by comparing A′ with A = [A0, A′],
so that each test instance can be categorized into a test category
or an auxiliary category. Following the setting in [69], 20% of
the training instances in each auxiliary category are moved
from training set to test set, leading to a new test set composed
of the instances from both auxiliary and test categories. To be

Fig. 4: The web images in the top (resp., bottom) row are
associated with 5 highest (resp., lowest) weights based on the
weight vector θ learnt by our method.

Fig. 5: The web images associated with 5 highest weights
based on the weight vector θ learnt by our Ours_WSL.

Fig. 2: The performance variation of our method on the Dogs dataset by tuning one trade-off parameter and fixing other
trade-off parameters as their optimal values. Vertical dashed lines indicate the optimal value of each trade-off parameter.

Fig. 3: The performance variation of our method w.r.t. different numbers of web training images per category.

Fig. 4: The web images in the top (resp., bottom) row are
associated with 5 highest (resp., lowest) weights based on the
weight vector θ learnt by our method.
TABLE III: 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| 55.74| 47.01|
| WSL+LR     | 57.60| 35.11| 55.13| 49.28|
| Chao et al. [69]| 25.75| 20.77| 31.53| 26.02|
| Ours       | 59.60| 36.00| 65.89| 53.83|

more specific, the new test set of CUB (resp., SUN and Dogs) contains in total 200 (resp., 717 and 113) categories.

In terms of baselines, we compare our method with basic linear regression which learns a linear regressor for each auxiliary category and test category, which is referred to as LR_mix in Table III. We also compare with WSL+LR which uses Xiao et al. [16] for test categories and linear regressor for auxiliary categories, considering that Xiao et al. [16] is the most competitive WSL baseline as reported in Table I. Moreover, we include generalized ZSL method in [69] as a baseline, which is specifically designed for the generalized ZSL setting [69]. The experimental results under the generalized setting are summarized in Table III. One observation is that the results drop sharply compared with those reported in Table I because the generalized setting is a more challenging task with the test instances come from both auxiliary categories and test categories. Nevertheless, our method still produces the best results on all three datasets, which indicates that our approach is still effective under the generalized setting.

B. Fine-grained Image Classification with Privileged Information

In this section, the experimental setting is basically the same as that in Section IV-A that is, we use the same datasets and the same splits of auxiliary/test categories.

Textual Features: For the web training images from test categories, we additionally extract textual features from the surrounding textual information of web images. Specifically, we crawl the tag, title, and comment of each web training image as raw textual information. Then, we build the vocabulary based on the top 2000 most frequent words from the aggregated raw textual information, during which we perform stop-word removal to eliminate the meaningless words. Finally, we encode the textual information of each web image into a 2000-dim term-frequency (TF) feature based on the vocabulary, leading to a 2000-dim textual feature for each web image.

Parameters: By using privileged information, we introduce another trade-off parameter $\gamma$ in [11], which is determined within the range $10^{-3}, 10^{-2}, \ldots, 10^3$ using the same cross-validation strategy as in Section IV-A and the details are omitted here.

Experimental Results: We evaluate our method using privileged information and SVM+ [11], which is trained based on web training images and their surrounding textual information. Similar to WSL+ZSL in Table I, we compare with a combo baseline WSDG_PI+ZSL, which averages the test decision values from WSDG_PI [20] and ZSL baseline [64]. Note that WSDG_PI [20] can handle the label noise of web data and utilize privileged information at the same time. Besides, we also include the results of LR, WSL+ZSL, and Ours from Table I for comparison. The results of various approaches are reported in Table IV.

From Table IV, we can see that SVM+ (resp., WSDG_PI+ZSL) outperforms LR (resp., WSL+ZSL), which shows that it is useful to utilize the additional textual information as privileged information. We also observe that after using privileged information, our approach achieves better performance on all three datasets, which shows the advantage of incorporating privileged information into our method for fine-grained image classification.

Generalized ZSL Setting: To further investigate the effectiveness of privileged information, we also evaluate our method under the same generalized setting as in Section IV-A in which the test instances come from both auxiliary categories and test categories. Recall that in Section IV-A we compare with LR_mix which learns a linear regressor for each auxiliary category and test category, and WSL+LR which uses Xiao et al. [16] for test categories and linear regressor for auxiliary categories. Here, to take advantage of privileged information, we learn SVM+ for each test category and SVM for each auxiliary category (auxiliary categories do not have privileged information), leading to the baseline SVM+_mix in Table V. We also use WSDG_PI [20] for test categories and linear regressor for auxiliary categories, leading to the baseline WSDG_PI+LR in Table V. Besides, we include the results of LR_mix, WSL+LR, and Ours from Table III for comparison.

Based on Table V we observe that SVM+_mix (resp., WSDG_PI+LR) achieves better results than LR_mix (resp., WSL+LR), which indicates that it is helpful to utilize the additional textual information as privileged information under the generalized setting. Another observation is that our method is further improved by using privileged information and produces the best results on all three datasets, which again shows the benefit of privileged information under the generalized setting.

TABLE IV: Accuracies (%) of different methods with or without privileged information (PI) on three datasets. The best results are highlighted in boldface.

| Dataset       | CUB  | SUN  | Dogs | Avg  |
|---------------|------|------|------|------|
| LR            | 68.39| 62.50| 77.67| 69.52|
| SVM+_mix      | 70.33| 64.00| 78.95| 71.09|
| WSL+ZSL       | 72.21| 78.50| 81.90| 77.55|
| WSDG_PI [20]+ZSL | 73.50| 79.00| 83.62| 78.71|
| Ours          | 77.94| 85.50| 86.82| 83.42|

TABLE V: Accuracies (%) of different methods with or without privileged information (PI) on three datasets under the generalized setting. The best results are highlighted in boldface.

| Dataset       | CUB  | SUN  | Dogs | Avg  |
|---------------|------|------|------|------|
| LR            | 68.39| 62.50| 77.67| 69.52|
| SVM+_mix      | 70.33| 64.00| 78.95| 71.09|
| WSL+ZSL       | 72.21| 78.50| 81.90| 77.55|
| WSDG_PI [20]+ZSL | 73.50| 79.00| 83.62| 78.71|
| Ours          | 77.94| 85.50| 86.82| 83.42|

TABLE III: 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| 55.74| 47.01|
| WSL+LR     | 57.60| 35.11| 55.13| 49.28|
| Chao et al. [69]| 25.75| 20.77| 31.53| 26.02|
| Ours       | 59.60| 36.00| 65.89| 53.83|
V. CONCLUSION

In this paper, a new learning scenario has been proposed for fine-grained image classification by using both web data and auxiliary categories. In this learning scenario, we have developed a method unifying webly supervised learning and zero-shot learning, which can transfer the knowledge from auxiliary categories to test categories and simultaneously handle the label noise and domain shift of web data. Moreover, our method has been further extended by taking advantage of the surrounding textual information of web images as privileged information. Comprehensive experiments have demonstrated the effectiveness of our proposed methods.

APPENDIX

A. Solution to (8)

For ease of representation, we rewrite the objective in (8) as follows,

\[ L_{D', A', \theta| E^w, \theta \in \mathcal{S}}^t = \frac{1}{2} \| \mathbf{X}^t - D^t A^t \|^2_F + \frac{\lambda_1}{2} \| D^t - D^a \|^2_F + \lambda_2 \| Z^t \|_S \]
\[ + \frac{\lambda_2}{2} \| \mathbf{E}^w - (\mathbf{X}^w - D^t A^w)\theta \|_2^2 + \lambda_3 \| \mathbf{E}^w \|_{2,1} \]
\[ + \frac{\mu}{2} \| |w - (\mathbf{X}^w - D^t A^w)\theta \|^2_F + \langle R, \mathbf{E}^w - (\mathbf{X}^w - D^t A^w)\theta \rangle \]
\[ + \frac{\mu}{2} \| A^t - Z^t \|^2_F + \langle T, A^t - Z^t \rangle. \]  (12)

To minimize (12), we update \( E^w, Z^t, D^t, A^t, \) and \( \theta \) one by one in an alternating fashion, which will be detailed in the following.

Update \( E^w \): The subproblem of (12) w.r.t. \( E^w \) is as follows,

\[ \text{min}_{E^w} \lambda_4 \| E^w \|_{2,1} + \frac{\mu}{2} \| E^w - \left( \mathbf{X}^w - D^t A^w \right)\Theta - \frac{R}{\mu} \|^2_F, \]

which has a closed-form solution [70]. Specifically, by denoting \( \mathbf{Q} = (\mathbf{X}^w - D^t A^w)\Theta - \frac{R}{\mu} \), if the optimal solution w.r.t. \( E^w \) is \( E^{w*} \), then the \( i \)-th column of \( E^{w*} \) is

\[ E^{w*}(:, i) = \begin{cases} \frac{q_i - 2 \lambda_4}{\| q_i \|_2} q_i, & \text{if } \frac{\lambda_4}{\mu} < \| q_i \|_2, \\ 0, & \text{otherwise}, \end{cases} \]  (13)

where \( q_i \) is the \( i \)-th column of \( Q \) and \( ||q_i||_2 \) is the L_2 norm of \( q_i \).

Update \( Z^t \): The subproblem of (12) w.r.t. \( Z^t \) is as follows,

\[ \min \lambda_2 \| Z^t \|_S + \frac{\mu}{2} \| Z^t - (A^t + \frac{T}{\mu}) \|^2_F, \]  (14)

which can be solved based on Singular Value Threshold (SVT) method [71]. By denoting \( M = A^t + \frac{T}{\mu} \) and the rank of \( M \) as \( r \), the singular value decomposition of \( M \) can be represented as \( M = U \Sigma V^t \), \( U \in \mathbb{R}^{m \times r}, V \in \mathbb{R}^{r \times n} \), and \( \Sigma = \mathbb{R}^{r \times r} \) is a diagonal matrix with diagonal entries being the singular values of \( M \). Then, the solution w.r.t. \( Z^t \) can be obtained as follows,

\[ Z^t = U D(\Sigma) V^t, \]  (15)

where \( D(\Sigma) \) is a diagonal matrix with \( \{ (\sigma_i - \frac{\lambda_3}{\mu} )_+ \}_{i=1}^r \) being the diagonal elements, in which \( \sigma_i \) is the \( i \)-th diagonal entry of \( \Sigma \) and \( (\cdot)_+ \) is an operator setting the negative entries to zero.

Update \( D^t \): The subproblem of (12) w.r.t. \( D^t \) is as follows,

\[ \min_{D^t} \frac{1}{2} \| X^t - D^t A^t \|^2_F + \frac{\lambda_1}{2} \| D^t - D^a \|^2_F + \frac{\mu}{2} \| E^w - (X^w - D^t A^w)\Theta \|^2_F + \langle R, E^w - (X^w - D^t A^w)\Theta \rangle. \]  (16)

We set the derivative of (16) w.r.t. \( D^t \) as zero, and obtain the closed-form solution to \( D^t \) as

\[ D^t = \left( X^t A^t + \lambda_1 D^a + (\mu X^w \Theta - \mu E^w - R) \Theta A^t \right)^{-1}. \]  (17)

Update \( A^t \): The subproblem of (12) w.r.t. \( A^t \) is as follows,

\[ \min_{A^t} \frac{1}{2} \| X^t - D^t A^t \|^2_F + \frac{\mu}{2} \| A^t - Z^t \|^2_F + \langle T, A^t - Z^t \rangle, \]  (18)

which also has a closed-form solution:

\[ A^t = \left( D^t D^t + \mu I \right)^{-1} (D^t X^t + \mu Z^t - T). \]

Update \( \theta \): The subproblem of (12) w.r.t. \( \theta \) is as follows,

\[ \min_{\theta \in S} \frac{\lambda_3}{2 n \| w \|^2} \| \mathbf{X}^w \mathbf{w}^t \mathbf{X}^w \theta - \frac{\lambda_3}{n \| w \|^2} \| \mathbf{X}^w \mathbf{w}^t \mathbf{X}^t \mathbf{1} \]
\[ + \frac{\mu}{2} \| \theta \|_1 + \mu \| \theta - \hat{\theta} \|_1. \]  (20)

After omitting the constant terms without \( \theta \), the problem in (19) can be converted to

\[ \min_{\theta \in S} \frac{\lambda_3}{2 n \| w \|^2} \| \theta \theta^t \mathbf{X}^w \mathbf{w}^t \mathbf{X}^w \theta - \frac{\lambda_3}{n \| w \|^2} \| \mathbf{X}^w \mathbf{w}^t \mathbf{X}^t \mathbf{1} \]
\[ + \frac{\mu}{2} \| \theta \|_1 + \mu \| \theta - \hat{\theta} \|_1, \]  (21)

in which \( \bar{\mathbf{P}} \) is a diagonal matrix sharing the same diagonal with \( \left( \mathbf{X}^w - D^t A^w \right)^t \left( \mathbf{X}^w - D^t A^w \right) \), \( \bar{\mathbf{P}} \) is the diagonal of \( \left( \mathbf{X}^w - D^t A^w \right)^t \mathbf{X}^w \), and \( \bar{\mathbf{r}} = (\mathbf{R} \circ \left( \mathbf{X}^w - D^t A^w \right))^t \mathbf{1} \). The problem in (20) can be further simplified as

\[ \min_{\theta \in S} \frac{1}{2} \| \theta \|_1 + \mu \| \theta - \hat{\theta} \|_1, \]  (21)
programming (QP) problem and could be solved using off-the-shelf QP solvers (e.g., Mosek). However, based on our experimental observation, existing QP solvers are not very efficient. So we develop our own Sequential Minimal Optimization (SMO) [22] based algorithm to solve [21], which is much more efficient than those off-the-shelf QP solvers. Simply speaking, we select the most violating pair of variables in \( \theta \) for update in each iteration sequentially until the objective of (21) converges.

The whole algorithm using inexact ALM is summarized in Algorithm 1. Based on our experimental observation, the algorithm usually converges within 50 iterations.

### B. Solution to (11)

The problem in (11) can be solved similarly to (5). We write the augmented Lagrangian function of (11) as follows,

\[
\mathcal{L}_{\mathbf{D}, \mathbf{A}, \mathbf{Z}, \mathbf{W}} = \frac{1}{2} \| \mathbf{X}^t - \mathbf{D}^t \mathbf{A}^t \|_F^2 + \frac{\lambda_1}{2} \| \mathbf{D}^t - \mathbf{D}^q \|_F^2 + \lambda_2 \| \mathbf{Z}^t \|_* + \frac{\gamma_1}{2} \| \mathbf{X}^w \|_2^2 + \lambda_4 \| \mathbf{E}^w \|_2^2 + \frac{\gamma_2}{2} \| \mathbf{E}^w - (\mathbf{X}^w - \mathbf{D}^t \mathbf{A}^t w) \|_F^2 + \langle \mathbf{R} \mathbf{E}^w - (\mathbf{X}^w - \mathbf{D}^t \mathbf{A}^t w) \mathbf{E}^w \rangle + \frac{\mu}{2} \| \mathbf{A}^t - \mathbf{Z}^t \|_F^2 + \langle \mathbf{T} \mathbf{A}^t - \mathbf{Z}^t \rangle, \tag{22}
\]

which can be minimized by updating \( \mathbf{E}^w, \mathbf{Z}^t, \mathbf{D}^t, \mathbf{A}^t, \mathbf{W} \) and one by one iteratively until the termination criterion is met. Compared with the solution to (5), the only difference lies in the steps of updating \( \mathbf{D}^t \) and \( \mathbf{W} \), which will be elaborated below.

#### Update \( \mathbf{D}^t \): The subproblem of (22) w.r.t. \( \mathbf{D}^t \) is as follows,

\[
\min_{\mathbf{D}^t} \frac{1}{2} \| \mathbf{X}^t - \mathbf{D}^t \mathbf{A}^t \|_F^2 + \frac{\lambda_1}{2} \| \mathbf{D}^t - \mathbf{D}^q \|_F^2 + \frac{\mu}{2} \| \mathbf{E}^w - (\mathbf{X}^w - \mathbf{D}^t \mathbf{A}^t w) \|_F^2 + \langle \mathbf{R} \mathbf{E}^w - (\mathbf{X}^w - \mathbf{D}^t \mathbf{A}^t w) \mathbf{E}^w \rangle + \frac{\lambda_2}{2} \| \mathbf{Z}^t \|_F^2. \tag{23}
\]

We set the derivative of (23) w.r.t. \( \mathbf{D}^t \) as zeros and arrive at the following closed-form solution to \( \mathbf{D}^t \):

\[
\mathbf{D}^t = \left[ \mathbf{X}^t \mathbf{A}^t + \lambda_1 \mathbf{D}^q + (\mu \mathbf{X} \mathbf{E}^w - \mathbf{R}) \mathbf{A}^w + \gamma \mathbf{X}^w - \gamma \mathbf{W} \mathbf{X}^w \mathbf{A}^w \right] \left( \mathbf{A}^t \mathbf{A}^t + \mu \mathbf{A}^w \mathbf{E}^w \mathbf{A}^w + \lambda_1 \mathbf{I} + \gamma \mathbf{A}^w \mathbf{A}^w \right)^{-1}. \tag{24}
\]

#### Update \( \mathbf{W} \): The subproblem of (22) w.r.t. \( \mathbf{W} \) is as follows,

\[
\min_{\mathbf{W}} \frac{1}{2} \| \mathbf{X}^w - \mathbf{D}^t \mathbf{A}^w \|_F^2. \tag{25}
\]

We set the derivative of (25) w.r.t. \( \mathbf{W} \) as zeros, and obtain the closed-form solution to \( \mathbf{W} \) as

\[
\hat{\mathbf{W}} = (\mathbf{X}^w \mathbf{X}^w - \mathbf{D}^t \mathbf{A}^w \mathbf{X}^w) (\mathbf{X}^w \mathbf{X}^w)^{-1}. \tag{26}
\]

### References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in Proceedings of the 22nd IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, Jun. 2009, pp. 248–255.

[2] 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 Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, Oct. 2016, pp. 301–320.

[3] C. H. Lampert, H. Nickisch, and S. Harmeling, “Attribute-based classification for zero-shot visual object categorization,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 3, pp. 453–465, 2014.

[4] X. Chen and A. Gupta, “Webly supervised learning of convolutional networks,” in Proceedings of the 15th International Conference on Computer Vision, Santiago, Chile, Dec. 2015, pp. 1431–1439.

[5] 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,” in Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, Jul. 2017.

[6] A. Torralba and A. A. Efros, “Unbiased look at dataset bias,” in Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, Jun. 2011, pp. 1521–1528.

[7] Z. Xu, S. Huang, Y. Zhang, and D. Tao, “Augmenting strong supervision using web data for fine-grained categorization,” in Proceedings of the 15th International Conference on Computer Vision (every 2 years), Santiago, Chile, Dec. 2015, pp. 2524–2532.

[8] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in Proceedings of the 27th Annual Conference on Neural Information Processing Systems, Lake Tahoe, Nevada, United States, Dec. 2013, pp. 3111–3119.

[9] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, Oct. 2014, pp. 1532–1543.

[10] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, “Evaluation of output embeddings for fine-grained image classification,” in Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, Jun. 2015, pp. 2927–2936.

[11] V. Vapnik and A. Vashist, “A new learning paradigm: Learning using privileged information,” Neural Networks, vol. 22, no. 5-6, pp. 544–557, 2009.

[12] L. Niu, A. Veeraraghavan, and A. Sabharwal, “Webly supervised learning meets zero-shot learning: A hybrid approach for fine-grained categorization,” in Proceedings of the 31th IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, Jun. 2018, pp. 7171–7180.

[13] X. Chen, A. Shrivastava, and A. Gupta, “NEIL: Extracting visual knowledge from web data,” in Proceedings of the 28th Annual Conference on Neural Information Processing Systems, Lake Tahoe, Nevada, United States, Dec. 2013, pp. 3111–3119.

[14] W. Li, L. Niu, and D. Xu, “Exploiting privileged information from web data for image categorization,” in Proceedings of the 13th European Conference on Computer Vision, Zurich, Switzerland, Sep. 2014, pp. 437–452.

[15] L. Niu, W. Li, and D. Xu, “Exploiting privileged information from web data for action and event recognition,” International Journal of Computer Vision, vol. 118, no. 2, pp. 130–150, 2016.

[16] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang, “Learning from massive noisy labeled data for image classification,” in Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, Jun. 2015, pp. 2691–2699.

[17] L. Niu, X. Xu, L. Chen, L. Duan, and D. Xu, “Action and event recognition in videos by learning from heterogeneous web sources,” IEEE Transactions on Neural Networks and Learning Systems, vol. 28, no. 6, pp. 1290–1304, 2017.

[18] L. Niu, Q. Tang, A. Veeraraghavan, and A. Sabharwal, “Learning from noisy web data with category-level supervision,” in Proceedings of the 31th IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, Jun. 2018, pp. 7689–7698.

[19] M. Bergamo and L. Torresani, “Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach,” in Proceedings of the 24th Annual Conference on Neural Information Processing Systems, Vancouver, Canada, Dec. 2010, pp. 181–189.
[62] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele, “Latent embeddings for zero-shot classification,” in Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, Jun. 2016, pp. 69–77.

[63] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha, “Synthesized classifiers for zero-shot learning,” in Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, Jun. 2016, pp. 5327–5336.

[64] Z. Zhang and V. Saligrama, “Zero-shot learning via joint latent similarity embedding,” in Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, Jun. 2016, pp. 6034–6042.

[65] S. M. Shojaee and M. S. Baghshah, “Semi-supervised zero-shot learning by a clustering-based approach,” arXiv preprint arXiv:1605.09016, 2016.

[66] H. Zhang and P. Koniusz, “Zero-shot kernel learning,” in Proceedings of the 31th IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, Jun. 2018, pp. 7670–7679.

[67] V. Kumar Verma, G. Arora, A. Mishra, and P. Rai, “Generalized zero-shot learning via synthesized examples,” in Proceedings of the 31th IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, Jun. 2018, pp. 4281–4289.

[68] J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization,” Journal of Machine Learning Research, vol. 13, no. Feb, pp. 281–305, 2012.

[69] 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 Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, Oct. 2016, pp. 52–68.

[70] G. Liu, Z. Lin, and Y. Yu, “Robust subspace segmentation by low-rank representation,” in Proceedings of the 27th IEEE International Conference on Machine Learning, Haifa, Israel, Jun. 2010, pp. 663–670.

[71] J.-F. Cai, E. J. Candès, and Z. Shen, “A singular value thresholding algorithm for matrix completion,” SIAM Journal on Optimization, vol. 20, no. 4, pp. 1956–1982, 2010.

[72] J. Platt, “Sequential minimal optimization: A fast algorithm for training support vector machines,” Microsoft Research, Tech. Rep. MSR-TR-98-14, April 1998.