# Fine-grained Recognition in the Wild: A Multi-Task Domain Adaptation Approach

Timnit Gebru Judy Hoffman Li Fei-Fei
CS Department Stanford University
{tgebru, jhoffman, feifeili}@cs.stanford.edu

## Abstract

*While fine-grained object recognition is an important problem in computer vision, current models are unlikely to accurately classify objects in the wild. These fully supervised models need additional annotated images to classify objects in every new scenario, a task that is infeasible. However, sources such as e-commerce websites and field guides provide annotated images for many classes. In this work, we study fine-grained domain adaptation as a step towards overcoming the dataset shift between easily acquired annotated images and the real world. Adaptation has not been studied in the fine-grained setting where annotations such as attributes could be used to increase performance. Our work uses an attribute based multi-task adaptation loss to increase accuracy from a baseline of* $4.1\%$ *to* $19.1\%$ *in the semi-supervised adaptation case. Prior domain adaptation works have been benchmarked on small datasets such as [46] with a total of* $795$ *images for some domains, or simplistic datasets such as [41] consisting of digits. We perform experiments on a subset of a new challenging fine-grained dataset consisting of* $1,095,021$ *images of* $2,657$ *car categories drawn from e-commerce websites and Google Street View.*

## 1. Introduction

The ultimate goal of image recognition is to recognize all objects in the world, as they appear in their natural environments. An even more difficult task, fine-grained recognition, aims to distinguish between objects in the same category (e.g. different bird species or car brands). Current state-of-the-art fine-grained classification methods [2, 5, 8, 33] focus on fully supervised learning regimes: a setting where human annotated images are available for all object categories of interest. To enable these methods, datasets have been proposed to train models recognizing all categories and scenes [15, 35, 54], or focus on the fine-grained recognition task [51, 30, 52, 42, 34].


Figure 1. We aim to recognize fine-grained objects in the real world without requiring large amounts of expensive expert annotated images. Instead, we propose training fine-grained models using cheaper annotated data such as field guides or e-commerce web sources (see *top row*). We adapt the learned models to our task using only a sparse set of annotations in the real world.

Models trained on these datasets are capable of outperforming humans when evaluated on benchmark tasks such as [15, 44]. However, this evaluation paradigm ignores a key challenge towards the development of real world object classification models. Namely, fixed datasets such as ImageNet or Birds offer a sparse and biased sample of the world [48]. Thus, to achieve comparable performance in real-world settings, fully supervised models trained with these datasets need additional annotated data from each new scenario. However, collecting images capturing all possible appearances of an object in a constantly changing real world environment is infeasible. The large number of possible images makes it prohibitively expensive to obtain labeled examples for every object category in the real world. Moreover, this annotation burden is amplified when we consider recognition for fine-grained categories. In this setting, only experts are able to provide our algorithms with labeled data.

Fortunately, freely available sources of paired images and category labels exist for many objects we may want to recognize. For example, images and annotations from a field guide can be used to train a model recognizing various bird species in the wild (Fig. 1 *(top row)*). Similarly, annotated car images on e-commerce websites can be used to train a model distinguishing between different types of cars

in unstructured urban environments (Fig. 1 *(middle row)*). However, images from these sources have different statistics from those we may encounter in the real world. And this statistical difference can cause significant degradation of model performance [48, 46, 4].

In this work, we study fine-grained domain adaptation as a step towards overcoming the dataset shift between easily acquired annotated images and the real world. To our knowledge, adaptation has not been studied in the fine-grained setting where it is especially expensive to obtain image annotations. In this scenario, many of our categories may be related to one another in some known hierarchical way. For example, multiple distinct car varieties may share the same body type or the same make.

Our contributions are two fold: first, we propose a new multi-task adaptation approach which explicitly benefits from these known cross-category relationships. Our model consists of a multi-task adaptation objective which simultaneously learns and adapts recognition at the attribute and category level. We first show that our objective effectively regularizes the source training and hence improves the generalization of the source model to the target domain. Then, for the task of semi-supervised adaptation (i.e. when category labels are only available from a subset of the classes in the target domain), we exploit the fact that labels will often exist for all attributes. For example, while annotated target images for a 1998 Honda Accord sedan may not be available, some images of other Hondas and sedans are likely in our dataset. In this way, we are able to apply different adaptation techniques at the class and attribute levels.

Our second contribution characterizes a large scale fine-grained car dataset for domain adaptation. While this dataset was introduced by [27] in the context of fine-grained detection, it has not been used in adaptation. We perform experiments on a subset of 170 out of $2,657$ classes (a total of $71,030$ images) and show significantly improved performance using our method. While visual domain adaptation has been well studied [46, 3, 28, 49, 24], most approaches focus on adapting between relatively small data sources consisting of tens of object categories and hundreds of images in total [46, 20, 41]. The use of such small datasets in developing adaptation algorithms makes it difficult to reliably benchmark these algorithms. To our knowledge, our work is the first to study this important problem on a large scale, real-world dataset and in the fine-grained scenario.

## 2. Related Work

**Fine-Grained object recognition.** While fine-grained image recognition is a well studied problem [2, 5, 8, 10, 11, 9, 16, 17, 19, 26], its real world applicability is hampered by limited available data. Works such as [33] have used large-scale noisy data to train state-of-the-art fine-grained recognition models. However, these models are unlikely to generalize to real world photos because they are trained with images derived from field guides or product shots. Similarly, standard fine-grained datasets such as [51] and [6] are derived from a single domain. Due to the large variation in object appearance between the real world and these datasets, models trained on these images are unlikely to generalize well to real world objects.

**Domain adaptation.** Domain adaptation works enhance the performance of models trained on one domain (such as product shot images) and applied to a different domain (such as real world photos). Since the theoretical framework provided by [4], many computer vision works have published algorithms for unsupervised domain adaptation: i.e. a task where no labeled target images are available during training [53, 39, 23, 1, 7, 50]. Most methods strive to learn a classifier with domain invariant features [49, 25, 38] . Long et al. relax the assumption of a single classifier for both source and target images and instead use 2 classifiers with a residual connection [39]. While these works focus on unsupervised domain adaptation, [49] performs semi-supervised adaptation, transferring knowledge from classes with labeled target images to those without. To our knowledge, there have been no studies of visual adaptation in the fine-grained setting. Our work builds on [49]'s method to show that attribute level softlabel transfer and domain confusion significantly boost performance in this scenario.

**Attributes, structured data and multitask learning.** Attributes have been used to improve object classification in [47] and perform zero shot learning in [43, 37]. Kodirov et al. [31] uses sparse coding and subspace alignment techniques to perform zero shot learning when images are sourced from multiple domains. We draw inspiration from these works and leverage attributes to improve performance in unsupervised and semi-supervised domain adaptation. In contrast to [32]'s adaptation of user specified attributes, we use labels shared between different fine-grained categories to facilitate class level transfer. While prior works such as [45, 21, 18] focus on attribute learning, our goal is to improve adaptation using ground truth attribute labels.

Our method to enforce consistency between attribute and class predictions is similar in spirit to a number of works exploiting label structure [14, 12]. [14] uses Hierarchy and Exclusion (HEX) graphs to encapsulate semantic relations between pairs of labels. We use a KL divergence loss between predicted label distributions instead of hard constraints.

Finally, some prior works have shown that learning multiple tasks can improve generalization for each task. For example, [13] found that a multi-task network for segmentation improves object detection results as a bi-product. Similarly, [22] showed that a machine learning to translate multiple languages performs better on each language. We observe similar results where a multi-task adaptation approach using attributes improves class level performance.

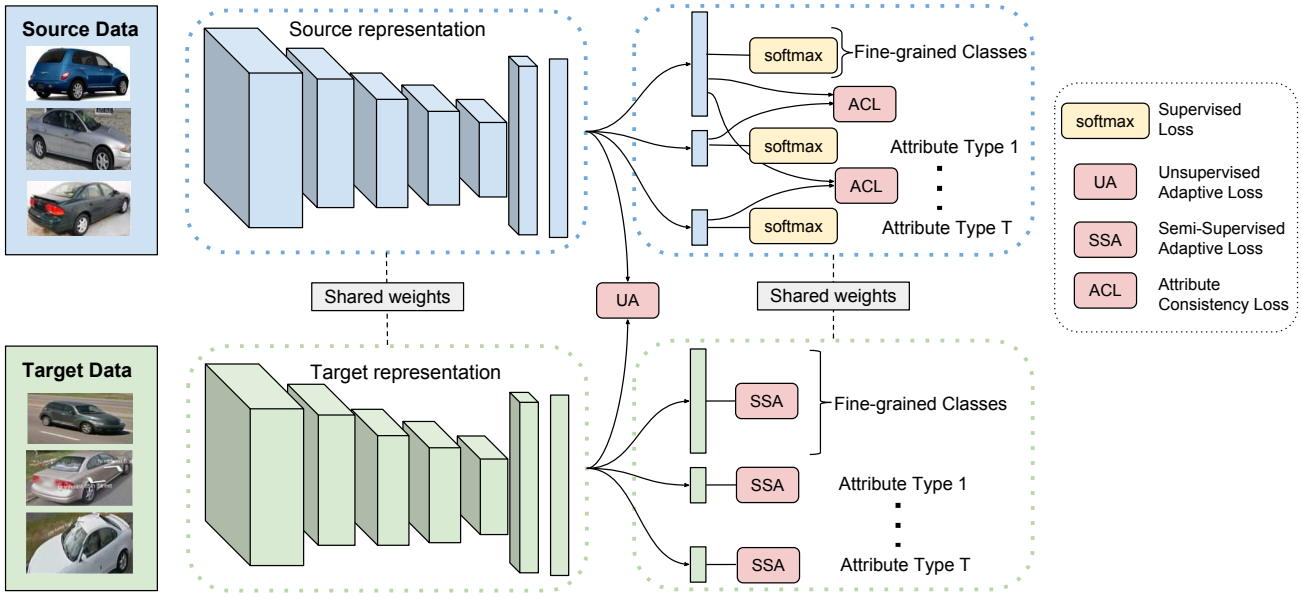


Figure 2. Our architecture for unsupervised and semi-supervised domain adaptation. Two CNNs based on [36] with shared weights classify source and target images. The fc7 feature maps of labeled source and target images are input into independent softmax classifiers classifying each attribute and fine-grained class of the image. Any unsupervised adaptive loss such as domain confusion (denoted as UA) [49] can be used to further improve adaptation. When labeled target images are available, semi-supervised adaptive loss (denoted as SSA) such as the soft label loss of [49] can be performed at the attribute, as well as fine-grained level. An attribute consistency loss (denoted as ACL) encourages the fine-grained and attribute classifiers to predict consistent labels.

## 3. Multi-Task Domain Adaptation for Fine-Grained Recognition

In the fine-grained classification setting, obtaining labels for every single class is infeasible. However, classes often share attributes. For instance, a Beagle and a Jack Russell terrier are both small dogs while a Bearded Collie and Afghan Hound are both shaggy dogs. In the general object classification setting, a taxonomic tree such as WordNet can be used to group categories and obtain labels at multiple levels in the hierarchy. Thus, while the target domain may not have labels for every leaf node class, we are more likely to have images annotated at higher levels in the hierarchy.

We leverage these additional annotations in a multi-task objective, providing regularization and additional supervision. Specifically, we minimize a multi-task objective consisting of softmax classification losses at the fine-grained and attribute level. In our architecture shown in Fig. 2, this is achieved by having multiple independent softmax layers that perform attribute level, in addition to category level, classification. We add an attribute consistency loss to prevent the independent classifiers from predicting conflicting labels. Any unsupervised adaptive loss (denoted as UA) in Fig. 2 can be used in conjunction with our method. Similarly, when target labels are available for some classes, any semi-supervised adaptive loss (denoted as SSA) can be added at the class and attribute levels. Here, we apply our method to [49] to evaluate its efficacy.

### 3.1. CNN Architecture for Multi-Task Domain Transfer

We give an overview of our architecture for semi-supervised domain adaptation shown in Fig. 2. Our model is trained using annotated source images for all classes, which we denote as $\{x_S, y_S\}$, and labeled and unlabeled target images, $\{x_T, y_T\}$. $x_S, x_T$ are source and target image samples respectively and $y_S, y_T$ are their associated labels. Our goal is to train a model classifying images $\{x_T\}$ for fine-grained categories with no labeled target images. We denote the number of target images as $N_T$ and the number of labeled target images as $N_{TL}$. $N_{TL} = 0$ and $N_{TL} = N_T$ in the unsupervised and fully supervised adaptation settings respectively. Only a subset of the target images are labeled in the semi-supervised adaptation setting resulting in $N_{TL} < N_T$.

In addition to class labels $y_S$, $y_T$, we also have attribute level annotations $y_{Sa}$, $y_{Ta}$ for source and target images respectively. There are at least as many labeled source and target images available for each attribute $a$, as each class $c$. This implies that even when no labeled target images are available for class $c$, there are labels for classes with similar attributes to $c$. We optimize a multi-task loss with 3 components: a softmax classification loss at the fine-grained and attribute levels, an attribute consistency loss, and any unsupervised or semisupervised adaptation loss.

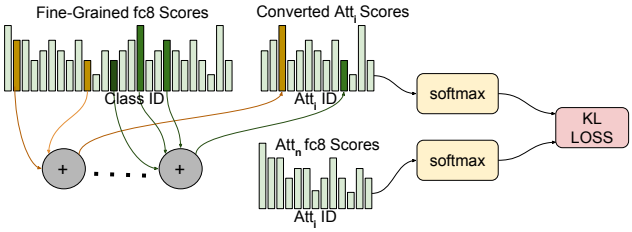


Figure 3. An attribute consistency loss between the fine-grained and attribute classifiers encourages them to predict consistent results. For each of the $i$ attributes $Att_i$, $fc8$ scores from the fine-grained classifier are converted to scores across attributes. We minimize a KL divergence loss between the softmax of these attribute scores and the softmax output of the attribute classifier $fc8_{Att_i}$.

### 3.2. Classification Loss

We start with a CNN following the architecture of [36], taking $\{x_S, y_S\}$ and $\{x_T, y_T\}$ as inputs. We denote the parameters of this classifier as $\theta_{rep}$. Let each attribute $a$ have $a_K$ categories. We have $N_a$ attribute classifiers $f_{a_n}$ parametrized by $\theta a_n, n = 1...N_a$. These classifiers operate on the image feature map $f(x, y; \theta_{rep})$ produced by our CNN. $N_a$ is the number of attributes and $x, y$ are an input image and its associated label respectively. We minimize $N_a$ softmax losses:

$$L_{a_n}(x, y; \theta_{rep}, \theta_{a_n}) = -\sum_{a_k=1}^{a_K} \mathbf{1}[y_a = a_k] \log p_{ak} \quad (1)$$

where $y_a$ is the ground truth label for image $x$ and attribute $a$, and $p_a = [p_{a1}, ...p_{aK}]$ is the softmax of the activations of attribute classifier $f_{a_n}$. I.e., $p = softmax(\theta_{a_n}^T f(x; \theta_{rep}))$.

In addition to attribute level softmax losses, we minimize a softmax classification loss at the fine-grained level. With $K$ classes, and a fine-grained classifier parametrized by $\theta_C$ operating on feature map $f(x, y; \theta_{rep})$, we minimize the loss:

$$L_C(x, y; \theta_{rep}, \theta_C) = -\sum_{k=1}^{K} \mathbf{1}[y = k] \log p_k \quad (2)$$

Our final multi-task softmax loss is the weighted sum of the attribute and fine-grained softmax losses. Omitting parameters for simplicity of notation,

$$L_{softmax} = \sum_{n=1}^{N_a} \alpha_n L_{a_n} + \alpha_c L_C \quad (3)$$

### 3.3. Attribute Consistency Loss

While our attribute and class classifiers are independently trained using ground truth labels, our pipeline so far poses no restrictions on how these classifications are related to each other. That is, the fine-grained classifier can output a class whose attributes are different from ones predicted by the attribute classifiers. However, we know that the attributes of the fine-grained class should be the same as those predicted by the independent attribute classifiers. To enforce this structure, we add an attribute consistency loss that penalizes differences between attributes predicted by the fine-grained and attribute classifiers. We minimize a symmetric version of the KL divergence between the distribution of attributes predicted by attribute classifier $a_n$ and those inferred by the fine-grained class classifier. Our procedure is visualized in Fig. 3. For each attribute $a$, we first convert scores across classes (fc8 output in [36]) to ones across categories for that attribute. Let $f = [f_1, ...f_k]$ consist of scores for $k$ classes. To obtain scores across attribute categories, we average values in $f$ belonging to classes from the same attribute. Averaging, rather than simply adding, scores mitigates the effect of dataset bias where some classes and attributes appear more frequently. We then compute a softmax distribution across attribute categories for attribute $a$, $\hat{p}_a = [\hat{p}_{a1}, ..., \hat{p}_{aK}]$ using the computed attribute scores.

We define a consistency loss for each attribute $a$ as the symmetric version of the KL divergence between $\hat{p}_a$ and $p_a$:

$$L_{con_{a_n}}(x, \theta_{rep}, \theta_{a_n}, \theta_c) = \frac{1}{2} D_{KL}(p_a||\hat{p}_a) + \frac{1}{2} D_{KL}(\hat{p}_a||p_a) \quad (4)$$

$$D_{KL}(p_a||\hat{p}_a) = \sum_{a_k=1}^{a_K} p_{ak} \log \frac{p_{ak}}{\hat{p}_{ak}} \quad (5)$$

where attribute $a$ has $a_K$ categories as defined in 3.2. Since we are not trying to match a reference distribution and are only minimizing the distance between two distributions, we use a symmetric version of the KL divergence in our loss instead of cross-entropy loss. Omitting parameters for simplicity, the final consistency loss $L_{consistency}$ is a weighted sum of the losses for each attribute:

$$L_{consistency} = \sum_{n=1}^{N_a} \beta_{a_n} L_{con_{a_n}} \quad (6)$$

### 3.4. Augmenting Existing Adaptation Algorithms with Attribute Loss

We can augment any existing adaptation algorithm with our attribute based losses to perform adaptation at the attribute as well as the class level. Here, we describe how we apply our method to [49]. To use our method with [49], we add the domain confusion and softlabel losses introduced in [49]. The softlabel loss is only used in the semi-supervised setting where labeled target images are available for some classes. However, in addition to a softlabel loss $L_{csoft}$ at the fine-grained level, we also minimize the softlabel objective $L_{asoft}$ for each attribute $a$. This allows us

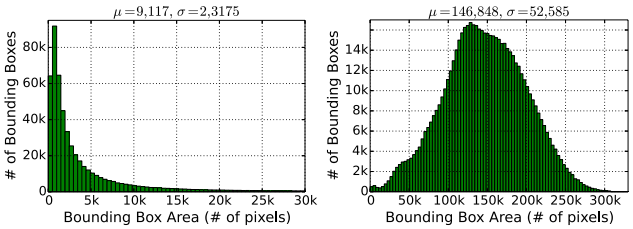


Figure 4. Histogram of *GSV* (left) and *web* (right) bounding box sizes. While cars in *GSV* images are typically small (with an average size of $9,117$ pixels), those in *web* images are much larger, occupying an average of $146,848$ pixels.

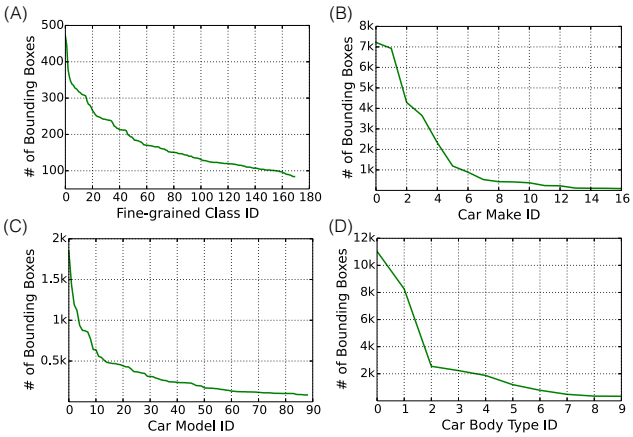


Figure 5. The distribution of *GSV* images for each class (A), each make (B), each model (C) and each body type (D) for the subset of the car dataset used in our evaluation. While each fine-grained class has less than $500$ labeled images, some body types have close to $12,000$ labeled *GSV* images (D).

to leverage attribute level annotations that exist for classes with no labeled target images. Denoting the domain confusion loss as $L_{conf}$, our final objective is a weighted sum of $L_{csoft}$, $L_{asoft}$, $L_{conf}$, $L_{softmax}$ and $L_{consistency}$.

## 4. Evaluation

We evaluate our multi-task adaptation algorithm on two datasets, a recently proposed large scale car dataset [27] and the office dataset [46] augmented with attributes taken from the WordNet [40] hierarchy. To test the efficacy of our attribute level adaptation approach, we modify an existing domain adaptation method, DC [49], by adding our attribute level losses.

We use Caffe [29] in all of our experiments. Our source only models are initialized with ImageNet weights using the released CaffeNet model [29]. For experiments on the car dataset, we use equal weights across all our losses and a temperature of 2 while calculating softlabel losses. We set the learning rate to $0.0001$ for all experiments and will release our custom layers for optimizing $KL$ divergence loss. For experiments on the office dataset, we set all loss weights to 1 except for domain confusion loss whose weight was set to $0.1$.

| Train | Test | Accuracy (%) Class | Make | Model | Body |
|---|---|---|---|---|---|
| S | S | 73.9 | 85.0 | 82.2 | 92.0 |
| S | T | 8.5 | 36.2 | 18.2 | 59.7 |
| T | T | 18.9 | 51.9 | 31.6 | 73.9 |
| S+T | T | 27.9 | 56.4 | 41.1 | 75.8 |

Table 1. We quantify the amount of domain shift between the *web* source domain (S) and *GSV* target domain (T). Training on source and evaluating on target shows a significant performance drop. Accuracies are shown for models trained at the fine-grained class, make, model and body-type level. There are 170 fine-grained classes, 89 models, 17 makes and 10 body-types in our dataset.

| Model | Adapt | Attr | Consist | Acc (%) |
|---|---|---|---|---|
| Source CNN | | | | 9.28 |
| Source CNN w/att | | ✓ | | 10.80 |
| Source CNN w/att+ACL | | ✓ | ✓ | 14.37 |
| DC [49] | ✓ | | | 14.98 |
| DC [49] w/att+ACL | ✓ | ✓ | ✓ | **19.05** |

Table 2. **Cars→GSV Unsupervised Adaptation:** We report multi-class accuracy for all classes in the GSV validation set and demonstrate the effectiveness of incorporating our attributes and consistency loss into the baseline and adaptive methods.

| Model | Adapt | Attr | Consist | Acc (%) |
|---|---|---|---|---|
| S+T CNN | | | | 4.12 |
| S+T CNN w/att+ACL | | ✓ | ✓ | 7.45 |
| DC [49] | ✓ | | | 12.34 |
| DC [49] w/att+ACL | ✓ | ✓ | ✓ | **19.11** |

Table 3. **Cars→GSV Semi-supervised Adaptation:** We report multi-class accuracy for the held-out unlabeled classes in the GSV validation set and demonstrate the effectiveness of incorporating our attributes and consistency loss into the baseline and adaptive methods.

| Model | Adapt | Attr | Consist | Acc (%) |
|---|---|---|---|---|
| Source CNN | | | | 60.9 |
| Source CNN w/att | | ✓ | | 59.5 |
| Source CNN w/att+ACL | | ✓ | ✓ | 61.2 |
| DC [49] | ✓ | | | 61.1 |
| DC [49] w/att+ACL | ✓ | ✓ | ✓ | **62.4** |

Table 4. **Amazon→Webcam Unsupervised Adaptation:** We report multi-class accuracy for the full Webcam dataset and demonstrate the effectiveness of incorporating our attributes and consistency loss into the baseline and adaptive methods.


Figure 6. Examples of *web* and *GSV* images for one type of car in our dataset. *Web* images are typically un-occluded with a high resolution while *GSV* images are blurry and occluded.

| Model | Adapt | Attr | Consist | Acc (%) |
|---|---|---|---|---|
| S+T CNN | | | | 45.5 |
| S+T CNN w/att+ACL | | ✓ | ✓ | 45.3 |
| DC [49] | ✓ | | | 47.0 |
| DC [49] w/att+ACL | ✓ | ✓ | ✓ | **51.8** |

Table 5. **Amazon→Webcam Semi-supervised Adaptation:** We report multi-class accuracy for the held-out unlabeled classes in the Webcam dataset and demonstrate the effectiveness of incorporating our attributes and consistency loss into the baseline and adaptive methods.

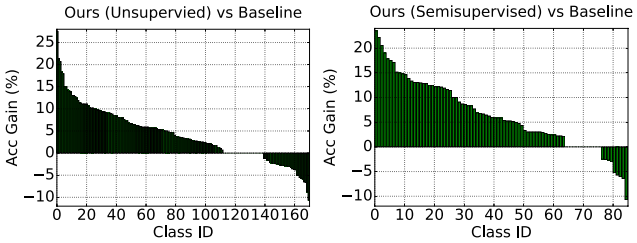


Figure 7. The difference in accuracy per class between our models and baselines on the car dataset. 66% of all fine-grained categories see a gain in accuracy in the unsupervised setting (*left*). Similarly, in the semi-supervised setting, our model improves classification accuracy on 75% of the held-out classes (*right*).

### 4.1. Large scale car dataset

The car dataset introduced in [27] consists of $1,095,021$ images of $2,657$ categories of cars from $4$ sources: craigslist.com, cars.com, edmunds.com and Google Street View. We refer to images from craigslist.com, cars.com and edmunds.com as *web* images and those from Google Street View as *GSV* images. As shown in Fig. 6, cars in *web* images are large and typically un-occluded whereas those in *GSV* are small, blurry and occluded. The difference in image size is apparent in Fig. 4 which shows a histogram of bounding box sizes in *GSV* and *web* images. These large variations in pose, viewpoint, occlusion and resolution make this dataset ideal for a study of domain adaptation, especially in the fine-grained setting. In addition to the category labels, each class is accompanied by metadata such as the make, model body type, and manufacturing country of the car.

### 4.2. Quantifying Domain Shift on the Car Dataset

In any adaptation experiment, it is crucial to first understand the nature of the discrepancy between the different sources of data. Following the standard set by [46], we quantify this shift in the car dataset by training a sequence of models and evaluating both within and across domains. We perform all of our experiments on a subset consisting of 170 of the most common classes in the dataset, particularly those with at least 100 target images per class. This ensures that we have enough images to reliably evaluate our algorithm.

In particular, we train a source only model and find that while accuracy is relatively high when evaluating within the source *web* domain (73.9%), performance catastrophically drops when evaluating within the *GSV* target domain. To aid in analyzing our semi-supervised domain adaptation experiments, we train a target only model using all available *GSV* labels. This model serves as an oracle for our adaptation experiments which use a reduced set of labeled images. As shown in Tab. 1, the target only model significantly outperforms our source only model indicating a large shift between the two domains. Finally, we train a joint fully supervised source and target model to test whether *web* and *GSV* data are complementary. Indeed, the joint model outperforms even the fully supervised target only model. This indicates that annotated images from the source domain will be a useful resource to train models classifying target images. Thus, in the next set of experiments, we evaluate our adaptation solutions.

For each of these experiments, we train models using fine-grained class as well as make, model and body type labels. There are 10 body types, 17 makes and 89 models in the subset of the dataset used for our experiments.


Figure 8. Example images for classes resulting in the highest accuracy gain with our method (*top*), and the highest accuracy drop with our method (*bottom*) on the car dataset. The class with the highest accuracy gain is the 2008-2010 Dodge Grand Caravan while the 2005-2011 Toyota Tacoma sees the highest accuracy loss.

### 4.3. Multi-Task Adaptation on the Car Dataset

A real world domain adaptation pipeline should leverage the availability of labeled target images for popular fine-grained objects, to improve classification performance on classes whose labels are difficult to obtain. With this motivation, we partition the target data into labeled and unlabeled sets to perform semi-supervised domain adaptation experiments. We first sort the fine-grained classes by the number of target images they have. We then use images for the top $50\%$ of target classes (85 classes) with the highest number of labels in conjunction with source images for all classes as labeled training data. Our test data comprises of images for the $50\%$ of classes with the least number of labels. Thus, no labeled target images from the held-out classes are used in training the models used in semi-supervised adaptation experiments.

Table 2 shows classification accuracies for various baseline methods as well as our architecture. Our baselines are source only and DC [49] adaptive models. We also compare our full model to one without attribute consistency loss. In all cases, our attribute level adaptation mechanism drastically improves performance. For example, in the unsupervised adaptation scenario, we see a $\sim 10\%$ gain. To ensure that our attribute loss indeed aids adaptation and does not solely improve the baseline classifier, we also train a CNN that solely incorporates non-adaptation based components of our loss: i.e, $L_{softmax}$ and $L_{consistency}$. While attributes indeed improve the baseline model (accuracy jumps from $9.28\%$ to $14.37\%$), they also improve adaptation. For example, domain confusion increases accuracy by $\sim 5\%$ without attributes but this improvement jumps to $\sim 10\%$ with attributes.

We see similar gains with our method in the semi-supervised adaptation setting. Training with a labeled subset of *GSV* classes in addition to *web* images generally reduces performance on the held-out *GSV* classes; the model overfits to the labeled *GSV* classes and becomes less generalizable. While domain confusion and softlabel loss combat this problem, we see the most significant improvement when these methods are used in conjunction with attribute level transfer: accuracy increases from $12.34\%$ to $19.11\%$. This confirms our intuition that using attribute labels helps our classifier learn domain invariant features.

### 4.4. Multi-Task Adaptation on the Office Dataset

While our attribute level adaptation approach is most suitable in the fine-grained setting, we also tested its efficacy on the office dataset [46] since there are no other fine-grained adaptation datasets. The office dataset consists of 31 classes of objects found around the office (such as backpacks, computers, desk lamps and scissors). For each of these objects, images are available from 3 domains: Amazon, WebCam and DSLR. While this dataset, introduced in 2010, is still the standard adaptation benchmark used today, its size is much smaller than the car dataset used in our experiments. For example, the WebCam domain consists of 785 images in total (across 31 classes).

Since the office dataset does not have attribute level annotations, we use class labels with varying degrees of granularity to evaluate our multi-Task adaptation approach. We annotate each image with the class name of its parent's, grandparent's and great grand parent's node in the WordNet hierarchy [40]. Thus, each image has 3 labels consisting of 3, 7 and 19 categories respectively in addition to its class label. We use these additional labels in place of attributes in our multi-task adaptation approach. Our source domain is Amazon and the target is WebCam.

Tab. 4 and Tab. 5 show our results for the unsupervised and semi-supervised scenarios respectively. Augmenting both baselines with our multi-task adaptation approach improves performance in the unsupervised as well as semi-supervised settings. This shows that our multi-task approach is not simply limited to attributes, and can be used in any scenario with a hierarchy of labels.

Nevertheless, our method's performance gain on the office dataset is much less than on cars. While car attributes are visually informative, WordNet labels might not be. For example, bike and backpack both share the node "con-

tainer" although their visual appearance is very different. Our future work plans to explore additional methods for obtaining visually distinctive attribute labels.

### 4.5. Analysis

Our model results in a significant increase in performance on most fine-grained categories. As shown in Fig. 7, 75% of held-out categories see a gain in accuracy over [49]. Similarly, in the unsupervised setting, our model improves performance on 66% of the target classes. There is no change in accuracy on 14% and 16% of classes while 10% and 18% of classes see a performance drop in the two regimes respectively.

While, as shown in Fig. 5, there is a maximum of 500 labeled target images per fine-grained class, Figs. 5(B), (C), and (D) show that some attributes have as many as 12,000 labeled images. Although we do not use any target images with fine-grained labels for our 85 held out classes as training data, there are classes in the training data with shared attributes as the test data. Thus, we expect our method to improve accuracy on classes with many attribute labels. Fig. 8 *top* shows example images for the top 3 classes with an accuracy gain in the semi-supervised setting on the car dataset. These classes have body type minivan, extended cab and SUV: 3 out of the top 4 body types with the highest number of labeled target training images.

Conversely, the class resulting in the highest accuracy loss with our method is a crew cab: there are only 57 labeled *GSV* images of crew cabs in our training set. Surprisingly, 2 out of the 3 classes with the highest accuracy loss are sedans. Although sedans have the most number of labeled *GSV* images in our training set (and thus expected to see an accuracy gain), one of these 2 classes has 243 source training images. Fig. 10 plots relative accuracy gain (compared to [49]) vs. the number of labeled source training examples per class. Our approach results in higher accuracy gain on classes with few labeled training data. We measure a correlation of $-0.29$ between the number of labels per class and the accuracy gain.

Finally, Fig. 9 shows example images in the $GSV$ test set and their corresponding nearest neighbors in the training set in the unsupervised setting. For each example image, we compute its feature activations using a baseline model trained with [49], and our multi-task approach. We retrieve images in the training set whose fc7 activations minimize the $||L_2||$ distance to fc7 activations of the example image. While our attribute based classifier retrieves images in the same class as the target image, the baseline adapted model returns a nearest neighbor in the wrong class.

## 5. Conclusion

We have presented a multi-task CNN architecture for semi-supervised domain adaptation. Our pipeline leverages


Figure 9. Source training images nearest to example target images according to [49] and our multi-task model. Nearest neighbors are computed with $||L_2||$ distance in the feature activation space. First column is the test example, second column shows results of models trained with [49] to compute the feature activations, and the last column shows results retrieved by our model.

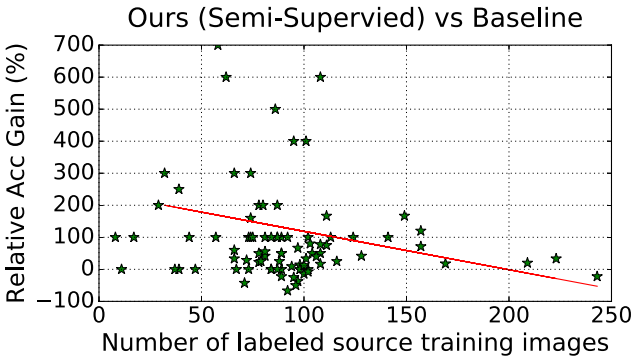


Figure 10. The number of labeled images per class vs our relative accuracy gain on the target held-out classes. We see an increase in accuracy gain with decreasing labeled training data.

the fact that fine-grained classes share attributes which can help transfer knowledge from classes seen in training to those that are not. We evaluated our method on a subset of a large-scale fine-grained dataset consisting of $\sim 1M$ images and 2,657 car categories. The large number of labeled images from multiple domains makes this dataset ideal for adaptation studies. We also evaluated on the standard office dataset using additional labels from WordNet. In the future, we plan to refine our methodology for incorporating attributes in adaptation, and perform hierarchical adaptation in settings where attribute labels are not available.

## 6. Acknowledgments

We thank Kenji Hata, and Oliver Groth for their valuable feedback. This research is partially supported by an ONR MURI grant, the Stanford DARE fellowship (to T.G.) and by NVIDIA (through donated GPUs).

## References


[1] R. Aljundi and T. Tuytelaars. Lightweight unsupervised domain adaptation by convolutional filter reconstruction. *arXiv preprint arXiv:1603.07234*, 2016. 2

[2] A. Angelova, S. Zhu, and Y. Lin. Image segmentation for large-scale subcategory flower recognition. In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, pages 39–45. IEEE, 2013. 1, 2

[3] Y. Aytar and A. Zisserman. Tabula rasa: Model transfer for object category detection. In *Proc. ICCV*, 2011. 2

[4] S. Ben-David, J. Blitzer, K. Crammer, F. Pereira, et al. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137, 2007. 2

[5] T. Berg and P. Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 955–962, 2013. 1, 2

[6] T. Berg, J. Liu, S. W. Lee, M. L. Alexander, D. W. Jacobs, and P. N. Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2019–2026. IEEE, 2014. 2

[7] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan. Domain separation networks. *arXiv preprint arXiv:1608.06019*, 2016. 2

[8] S. Branson, G. Van Horn, S. Belongie, and P. Perona. Bird species categorization using pose normalized deep convolutional nets. *arXiv preprint arXiv:1406.2952*, 2014. 1, 2

[9] Y. Chai, V. Lempitsky, and A. Zisserman. Symbiotic segmentation and part localization for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 321–328, 2013. 2

[10] Y. Chai, V. S. Lempitsky, and A. Zisserman. Bicos: A bi-level co-segmentation method for image classification. In *ICCV*, pages 2579–2586, 2011. 2

[11] Y. Chai, E. Rahtu, V. Lempitsky, L. Van Gool, and A. Zisserman. Tricos: A tri-level class-discriminative co-segmentation method for image classification. In *European conference on computer vision*, pages 794–807. Springer, 2012. 2

[12] L.-C. Chen, A. G. Schwing, A. L. Yuille, and R. Urtasun. Learning deep structured models. In *ICML*, pages 1785–1794, 2015. 2

[13] J. Dai, K. He, and J. Sun. Instance-aware semantic segmentation via multi-task network cascades. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3150–3158, 2016. 2

[14] J. Deng, N. Ding, Y. Jia, A. Frome, K. Murphy, S. Bengio, Y. Li, H. Neven, and H. Adam. Large-scale object classification using label relation graphs. In *European Conference on Computer Vision*, pages 48–64. Springer, 2014. 2

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 1

[16] J. Deng, J. Krause, and L. Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2013. 2

[17] K. Duan, D. Parikh, D. Crandall, and K. Grauman. Discovering localized attributes for fine-grained recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3474–3481. IEEE, 2012. 2

[18] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1778–1785. IEEE, 2009. 2

[19] R. Farrell, O. Oza, N. Zhang, V. I. Morariu, T. Darrell, and L. S. Davis. Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance. In *2011 International Conference on Computer Vision*, pages 161–168. IEEE, 2011. 2

[20] B. Fernando, T. Tommasi, and T. Tuytelaars. Joint cross-domain classification and subspace learning for unsupervised adaptation. *Pattern Recognition Letters*, 65:60 – 66, 2015. 2

[21] V. Ferrari and A. Zisserman. Learning visual attributes. In *Advances in Neural Information Processing Systems*, pages 433–440, 2007. 2

[22] O. Firat, K. Cho, and Y. Bengio. Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint arXiv:1601.01073*, 2016. 2

[23] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*, 2014. 2

[24] Y. Ganin and V. Lempitsky. Unsupervised Domain Adaptation by Backpropagation. In *International Conference on Machine Learning (ICML)*, 2015. 2

[25] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *arXiv preprint arXiv:1505.07818*, 2015. 2

[26] E. Gavves, B. Fernando, C. G. Snoek, A. W. Smeulders, and T. Tuytelaars. Fine-grained categorization by alignments. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1713–1720, 2013. 2

[27] T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng, and L. Fei-Fei. Fine-grained car detection for visual census estimation. AAAI, 2017. 2, 5, 6

[28] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Proc. CVPR*, 2012. 2

[29] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 5

[30] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, volume 2, 2011. 1

[31] E. Kodirov, T. Xiang, Z. Fu, and S. Gong. Unsupervised domain adaptation for zero-shot learning. In *Proceedings*

*of the IEEE International Conference on Computer Vision*, pages 2452–2460, 2015. 2

[32] A. Kovashka and K. Grauman. Attribute adaptation for personalized image search. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3432–3439, 2013. 2

[33] 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. *arXiv preprint arXiv:1511.06789*, 2015. 1, 2

[34] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. 1

[35] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. 1

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 3, 4

[37] 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*, 36(3):453–465, 2014. 2

[38] M. Long and J. Wang. Learning transferable features with deep adaptation networks. *CoRR, abs/1502.02791*, 1:2, 2015. 2

[39] M. Long, J. Wang, and M. I. Jordan. Unsupervised domain adaptation with residual transfer networks. *arXiv preprint arXiv:1602.04433*, 2016. 2

[40] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 5, 7

[41] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. 2011. 1, 2

[42] M.-E. Nilsback and A. Zisserman. A visual vocabulary for flower classification. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1447–1454. IEEE, 2006. 1

[43] B. Romera-Paredes and P. Torr. An embarrassingly simple approach to zero-shot learning. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 2152–2161, 2015. 2

[44] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 1

[45] O. Russakovsky and L. Fei-Fei. Attribute learning in large-scale datasets. In *European Conference on Computer Vision*, pages 1–14. Springer, 2010. 2

[46] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *European Conference on Computer Vision (ECCV)*, pages 213–226. Springer, 2010. 1, 2, 5, 6, 7

[47] Y. Su and F. Jurie. Improving image classification using semantic attributes. *International journal of computer vision*, 100(1):59–77, 2012. 2

[48] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1521–1528. IEEE, 2011. 1, 2

[49] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *International Conference on Computer Vision (ICCV)*, 2015. 2, 3, 4, 5, 6, 7, 8

[50] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014. 2

[51] 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. 1, 2

[52] L. Yang, P. Luo, C. Change Loy, and X. Tang. A large-scale car dataset for fine-grained categorization and verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3973–3981, 2015. 1

[53] D. Yoo, N. Kim, S. Park, A. S. Paek, and I. S. Kweon. Pixel-level domain transfer. *arXiv preprint arXiv:1603.07442*, 2016. 2

[54] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Neural Information Processing Systems (NIPS)*, 2014. 1