An Adversarial Domain Adaptation Network for Cross-Domain Fine-Grained Recognition

Yimu Wang[1] Ren-Jie Song[2] Xiu-Shen Wei[2] Lijun Zhang[1]
[1]Nanjing University [2]Megvii Technology
{wangym, zhanglj}@lamda.nju.edu.cn, {songrenjie, weixiushen}@megvii.com

Abstract

In this paper, we tackle a valuable yet very challenging visual recognition task, where the instances are within a subordinate category, and the target domain undergoes a shift with the source domain. This task, termed as cross-domain fine-grained recognition, relates closely to many real-life scenarios, e.g.*, recognizing retail products in storage racks by models trained with images collected in controlled environments. To deal with this problem, we design a new algorithm and propose a corresponding fine-grained domain adaptation dataset. Firstly, we propose a novel end-to-end CNN architecture that integrates two specialized modules: an adversarial module for domain alignment and a self-attention module for fine-grained recognition. The adversarial module is used to handle domain shift by gradually aligning the different domains with domain-level and class-level alignments, and strive to help the classifier learn with domain-invariant features generated by nets. The self-attention module is designed to capture discriminative image regions which are crucial for fine-grained visual recognition. Secondly, we collect a large-scale fine-grained domain adaptation dataset of retail products, which contains 52,011 images of 263 classes from 3 domains. Thirdly, we validate the effectiveness of our method on three datasets, showing that the proposed method can yield significant improvements over baseline methods on fine-grained datasets. Besides, we also evaluate the effectiveness of the self-attention module by performing visualization, which can capture the discriminative image regions in both source and target domains.*

1. Introduction

As a fundamental and challenging problem in computer vision, fine-grained image analysis (FGIA) [32] has attracted extensive research attention for several decades, especially in fine-grained image recognition [18, 33, 34], which aims to distinguish categories that are similar to each other, while different categories can only be distinguished by slight and subtle differences. In many real-life tasks, *e.g.*, instance retrieval [11, 31], vehicle identification [7], and retail production recognition [30], fine-grained image recognition is widely applied.

In the literature, to push the accuracy of fine-grained image recognition, some works focus on designing effective networks to learn more discriminative fine-grained representations [1, 3, 4, 5, 18, 31]. However, these methods ignore the challenge that there exists domain shift between the source domain and the target domain. As shown in Fig. 1, in the retail industrial scenario, a retail product recognition system is trained on the images taken in a controlled environment, where an ideal background and different views of products can be collected. However, the test environment (a.k.a. target domain) is usually the realistic storage racks scenario, where the background is noisy, and random orientations, different lightings, and complex clutters of products are also common.

In this paper, we study the problem of fine-grained image recognition with domain shift and propose a domain adaptation fine-grained network. Specifically, our network consists of two main components: a adversarial domain adaptation (DA) module and a self-attention (SA) module. The adversarial module is designed to progressively align source and target domains by domain-level alignment with adversarial discriminators and category-level alignment with the cosine metric (cf. Sec. 3.2). The adversarial discriminator in domain-level alignment first globally aligns the different domains in course-grained. Then, employing the cosine metric in category-level alignment locally aligns clusters of different domains category by category which is feasible for the fine-grained task. The second component is self-attention, which is tailed for the fine-grained task. It focuses on the subtle discriminative parts of fine-grained objects, and thus can further boost the accuracy.

Furthermore, considering the significant practical and research value of cross-domain fine-grained recognition, we collect, label images and construct a dataset named *DA-Retail*. It contains 52,011 images of 263 classes, enabling researchers to further study on domain adaptation fine-grained

978-1-7281-6553-0/20/$31.00 ©2020 IEEE

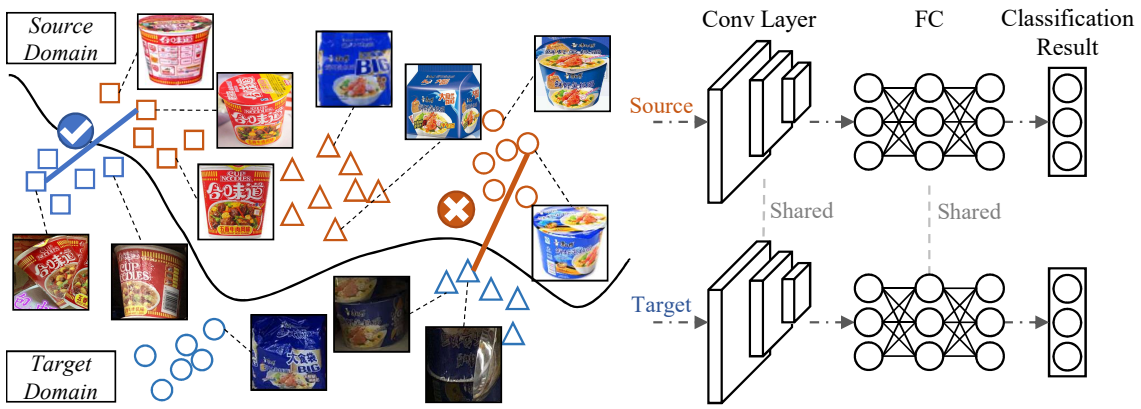

Figure 1. This task is to recognize fine-grained retail products in real-life shelves, while training data is *not* from the same/similar environment. This is an example where a domain shift exists between the source domain and the target domain. Features may be falsely aligned across domains, making classification wrong, as shown with a red cross in the left part. The domain shift plus the fine-grained nature of objects will bring extra challenges to the problem. Images from the source and target domains are classified through the same network.

recognition task. Also, it benefits the robust training with huge discrepancies among 3 different domains.

In empirical studies, we conduct comprehensive experiments on two different fine-grained datasets, *i.e.*, *GSV Cars* [7] and our *DA-Retail* (cf. Sec. 4) and evaluate the methods in both unsupervised and semi-supervised settings. The results of *GSV Cars* demonstrate that our model can greatly improve accuracy by 3.56% and 1.02% in both two settings. Besides, our model performs better than the state-of-the-art [6] of fine-grained domain adaptation, improving by 4.57% and 4.74% on the proposed dataset *DA-Retail*. To show the priority of our method, we also evaluate the performances of state-of-the-art [6] and our method on the generic dataset, *i.e.*, *Office* [25]. Our model outperforms state-of-the-art [6] by 2.1% in the unsupervised setting and 0.8% in the semi-supervised setting, respectively.

We summarize the main contributions below:

- We propose a novel domain adaptation fine-grained network consisting of two main components: an adversarial module designed for deriving domain-invariant features and a self-attention module for capturing the subtle discriminative parts of fine-grained objects.
- We collect, label and propose a domain adaptation fine-grained dataset, named *DA-Retail*. Our dataset contains 52,011 images of 263 classes in 3 different domains for domain adaptation performance evaluation.
- We conduct comprehensive experiments on three domain adaptation datasets (including our *DA-Retail*). Empirical results show that our proposed network outperforms state-of-the-art methods [6, 27] of the domain adaptation fine-grained task on three datasets.

2. Related work

In this section, we briefly review related work on fine-grained image recognition and domain adaptation and compare the difference among existing methods.

2.1. Fine-grained image recognition

Research on fine-grained recognition has been working in two ways recently. The first is leveraging features to get the high-order information. A symmetric two-stream network is proposed as Bilinear [18] by leveraging the second-order information. Then, the idea is quickly extended to bilinear pooling [4], which can achieve equal performance compared with Bilinear only using one-stream. The other way is bringing or predicted extra information into training, as text [24, 38], attribute [6] and part annotation [2, 8, 35].

In the literature, a recent work Multi-Task [6] also focuses on cross-domain fine-grained recognition, which is related to ours. Compared with this study, the major differences are: (i) they build the model in a multi-task framework, while we solve the cross-domain problem directly; (ii) their model requires fine-grained attributes for recognition, which is more expensive than image-level labels, while we only use image-level supervisions; and (iii) our model achieve better results than Multi-Task [6].

2.2. Domain adaptation

The research on domain adaptation has been working on reducing the discrepancy between different domains. Maximum Mean Discrepancy (MMD) [20] and Coral [26] design two different metrics to measure and minimize the distance. Batch Normalization (BN) [17] is also capable of this task by forcing the distributions of different domains to be closer. Nevertheless, recently, some researches strive to learn a classifier with domain-invariant features. Simultaneous Deep

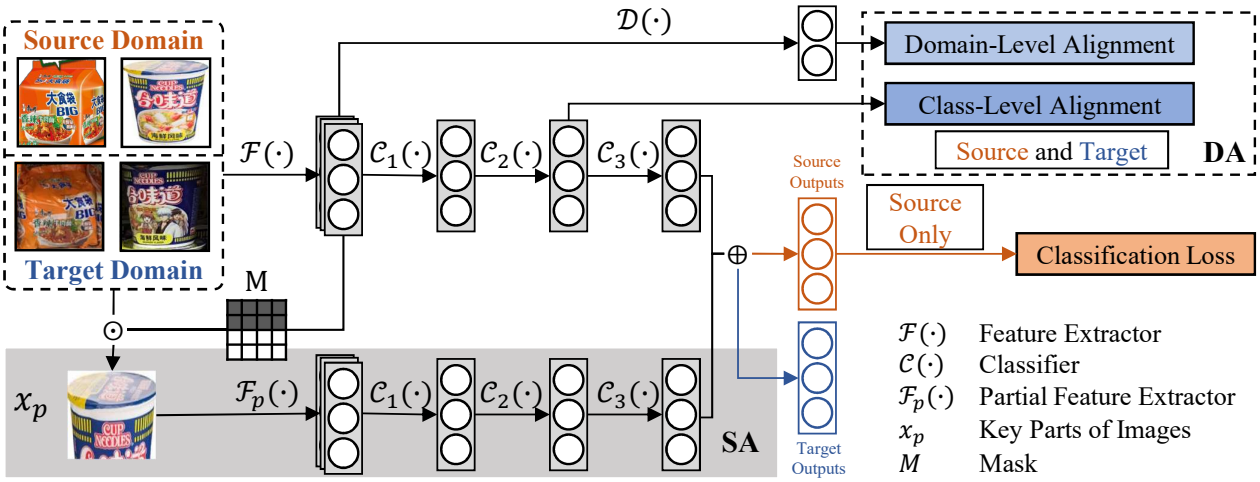

Figure 2. The structure of our proposed model. Two symmetric streams take the images from source domain and target domain as the input. Here we illustrate only one stream for clear presentations. The weights of the feature extractor $\mathcal{F}(\cdot)$, classifier $\mathcal{C}(\cdot)$ and discriminator $\mathcal{D}(\cdot)$ are shared. The adversarial module for domain alignment contains domain-level and class-level alignments. The self-attention module for fine-grained recognition is presented in the gray background in this figure. The network can be trained with only image-level supervisions.

Transfer Nets (DC) [27] employs a fully-connected layer to indicate which domain the input is and adds a soft label constraint to further force the model outputting domain-invariant features. Also, [16] proposes to minimize the intra-class dispersion for solving the misalignments. [12] utilizes the feature map to align the different domains. Further, with Generative Adversarial Nets (GAN), the researchers enhance the accuracy through extracting more domain-invariant features, *i.e.*, CoGAN [19].

Our method is related to existing methods [16, 19, 22, 27] in some aspects. However, there are some differences between our method and others. First, our model sequentially aligns the different domains by domain-level and category-level alignments which are lighter than [16] as we employ a simple but powerful metric. Second, our model can be simply extended by adding more domain-discriminators to handle the multi-domain scenario. Third, the attention map in our model is utilized to find the discriminative parts for the fine-grained task, while [12] used the attention map to align the different domains.

3. Our method for cross-domain fine-grained recognition

In this section, we present our cross-domain fine-grained recognition model. We use $\mathcal{S}$ and $\mathcal{T}$ to denote the domain of training data and testing data as the *source domain* and the *target domain*. Also, we denote the number of classes by N_{class}. The input image and its corresponding class label are presented by $\mathbf{x}$ and a one-hot vector $\mathbf{y} = (y_1, y_2, \ldots, y_{N_{class}})^\top$, respectively.

A real-world domain adaptation solution should utilize labeled source or target images which are easy to collect and improve the classification performance on target images whose label are hard to obtain. In the following, we investigate both unsupervised and semi-supervised settings. In the unsupervised setting, all the source images are labeled while all the target images are unlabeled. In the semi-supervised setting, the source images and a subset of target images are labeled and available in the training procedure.

3.1. Overview structure

Fig. 2 shows the framework of our end-to-end cross-domain fine-grained recognition model. Following Multi-Task [6], we employ CaffeNet [10] as the base model, while further employ ResNet [9] for generalizing our methods to more base model. For clear presentations, we separate it into two parts, *i.e.*, the feature extractor $\mathcal{F}(\cdot)$ and the category classifier $\mathcal{C}(\cdot)$. Concretely, $\mathcal{F}(\cdot)$ and $\mathcal{C}(\cdot)$ correspond to the convolution component and the fully-connected layers of CaffeNet. For an input image $\mathbf{x}$, the convolution representation and the classification prediction are calculated by $\mathcal{F}(\mathbf{x}) \in \mathbb{R}^{h\times w\times d}$ and $\hat{\mathbf{y}} = \mathcal{C}(\mathcal{F}(\mathbf{x})) = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{N_{class}})^\top$, respectively.

As shown in Fig. 2, we develop two modules to conquer the challenges in this task. Firstly, to reduce the performance drop caused by domain shift, we propose a module for domain alignment, *i.e.*, domain alignment module (DA). This module is employed to derive the domain-invariant features $\mathcal{F}(\mathbf{x})$. Details of this module will be elaborated shortly. Secondly, as objects in fine-grained recognition usually differ in subtle image regions, we propose a self-attention module (SA) to generate a corresponding self-attention map by compressing $\mathcal{F}(\mathbf{x})$ using channel-wise average pooling. Be-

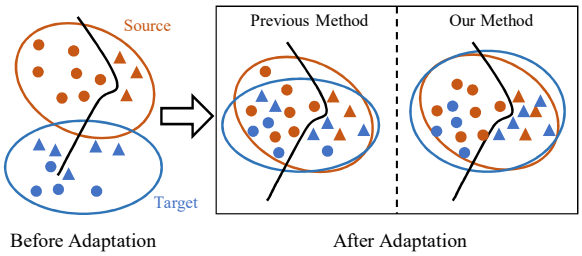

Figure 3. Comparing with most of the previous work (only domain-level alignment) and our work (domain-level and category-level alignments). Employing only Domain-level alignment may misalign features of the different classes from different domains while employing category-level alignment can fix this problem.

cause the model is trained for recognition, the intensity of each pixel in the self-attention map is proportional to the discriminative power. In this way, we can approximate the spatial distribution of the most discriminative part efficiently by our proposed mask mechanism. SA has a similar structure with the mainstream, including a mask producer $\mathcal{M}(\cdot)$, a part-level feature extractor $\mathcal{F}_p(\cdot)$ and a classifier $\mathcal{C}(\cdot)$.

3.2. Domain Alignment (DA) module

We continuously align different domains in two levels essential for domain alignment, *i.e.*, domain-level and class-level. Specifically, we first globally reduce the discrepancy between different domains, and further align the same class of different domains in the following fully-connected layers $\mathcal{C}(\cdot)$. Most of the previous work focuses on global alignment (domain-level alignment) ignoring the class-level alignment, which is the problem shown in Fig. 3.

3.2.1 Domain-level alignment

The features $\mathcal{F}(\mathbf{x})$ have domain shift between the source domain and the target domain. We here leverage the adversarial learning to align different domains in the domain-level alignment. Concretely, we employ discriminators $\mathcal{D}(\cdot)$ to output the probabilities of images belonging to the source domain. The domain-level alignment loss is defined as:

$$\mathcal{L}_{domain} = \mathbb{E}_{\mathbf{x}\sim\mathcal{T}}\left[\mathcal{D}(\mathcal{F}(\mathbf{x}))\right] + \mathbb{E}_{\mathbf{x}\sim\mathcal{S}}\left[1-\mathcal{D}(\mathcal{F}(\mathbf{x}))\right]. \tag{1}$$

In domain-level alignment, the feature extractor $\mathcal{F}(\cdot)$ is treated as a generator. The task of discriminator $\mathcal{D}(\cdot)$ is to distinguish the representation generated from images in the source domain or the target domain, while the generator (a.k.a. the feature extractor $\mathcal{F}(\cdot)$) tries to fool the discriminator $\mathcal{D}(\cdot)$ by deriving domain-invariant features. They are playing a zero-sum game and can be modeled by a min-max optimization. The feature extractor $\mathcal{F}(\cdot)$ tries to minimize Eq. (1), while the domain discriminator $\mathcal{D}(\cdot)$ leans to maximize it. Ideally, after convergence, domain-invariant image representations can be obtained. To ease the training procedure, we employ multiple discriminators and each discriminator treats one specific domain as the source domain. In the scenario with two domains, we have two discriminators, where one takes the source as the source and another takes the target as the source. Basically, our method can be easily extended to multiple domains by simply equipping with multiple domain discriminators.

3.2.2 Category-level Alignment

Category-level alignment is employed to ensure that features of the same class from different domains are close in the semi-supervised setting. Specifically, we employ the class-level loss at the second fully-connected layer of the base model (denoted as $\mathcal{C}_2(\cdot)$), which is defined as follows:

$$\begin{aligned}\mathcal{L}_{category} =& \mathbb{E}_{\mathbf{x_m},\mathbf{x_n}\sim\mathcal{S}\cup\mathcal{T}_l}[\mathbb{I}(\mathbf{y}_m=\mathbf{y}_n)\cdot d_{mn} \\ &+ \mathbb{I}(\mathbf{y}_m\neq\mathbf{y}_n)\cdot\max(\delta-d_{mn},0)], \\ d_{mn} =& sim(\mathcal{C}_2(\mathbf{x}_m),\mathcal{C}_2(\mathbf{x}_n)), \\ \mathbb{I}(cond) =& \begin{cases}1, & cond \text{ is true},\\ 0, & \text{otherwise}.\end{cases}\end{aligned} \tag{2}$$

where $\mathcal{T}_l$ is the labeled target domain subset available in training, $sim(\cdot,\cdot)$ is the cosine similarity function, $\mathbf{y}_m$ and $\mathbf{y}_n$ are labels of $\mathbf{x}_m$ and $\mathbf{x}_n$, and δ is the target margin.

3.3. Self-Attention (SA) module for capturing fine-grained parts

Tailed for fine-grained recognition, we further propose a self-attention (SA) module for capturing the discriminative fine-grained parts. Specifically, we choose the parts of images, whose activation values are bigger than the average values of the features, as the most informative parts. After choosing the most important parts of images, we use them as the input of our part-level feature extractor $\mathcal{F}_p(\cdot)$ to capture the representation of crucial parts in images. Then, the part-level features will go through the classifier $\mathcal{C}(\cdot)$ outputting a part-level classification result.

Concretely, after the features $\mathcal{F}(\mathbf{x}) \in \mathbb{R}^{h\times w\times d}$ are returned, we employ channel-wise average pooling to obtain $\mathcal{A}(\mathbf{x}) = avg(\mathcal{F}(\mathbf{x})) \in \mathbb{R}^{h\times w\times 1}$. Consequently, we calculate the mean value $\bar{a}$ of all the positions of the $h\times w$ matrix in $\mathcal{A}(\mathbf{x})$ as the adaptive threshold to decide which positions localize key parts. If the activation response of a position is higher than $\bar{a}$, we set the element corresponding to the same position in the mask map $\mathbf{M} \in \mathbb{R}^{h\times w\times 1}$ as 1; otherwise, we set it as 0. Therefore, we can locate the most informative fine-grained object parts based on the positive values (*i.e.*, 1) of the mask. Later, the mask is resized using the bicubic interpolation, such that its size is the same as the input image. Then, by applying the Hadamard product with resized masks and images, we can locate the key parts of images

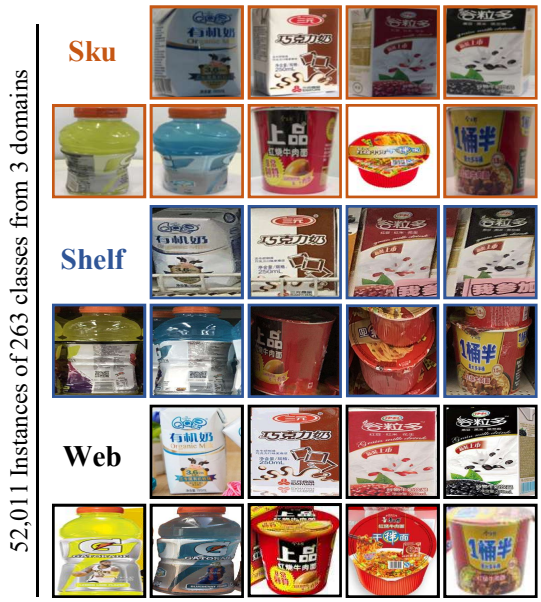

Figure 4. Examples of our dataset. We have 263 fine-grained classes in 3 different domains.

$\mathbf{x}_p = resize(\mathbf{M}) \odot \mathbf{x}$ in such a self-attention way. Based on these parts $\mathbf{x}_p$, the part-level representation learner $\mathcal{F}_p(\cdot)$ will take it as the inputs, and return the part-level features $\mathcal{F}_p(\mathbf{x}_p)$. Finally, the classifier will give the partial classification results by $\hat{\mathbf{y}}_p = \mathcal{C}(\mathcal{F}_p(\mathbf{x}_p))$. The final prediction is a weighted average of $\hat{\mathbf{y}}$ and $\hat{\mathbf{y}}_p$ as

$$\hat{\mathbf{y}} = \gamma \cdot \mathcal{C}(\mathcal{F}(\mathbf{x})) + (1-\gamma) \cdot \mathcal{C}(\mathcal{F}_p(\mathbf{x}_p)) \,, \tag{3}$$

where we set $\gamma = 0.5$ in our experiments.

3.4. Objective loss function

Classification Loss: The classification results with SA are a weighted average of $\mathbf{y}$ and $\hat{\mathbf{y}}$ as Eq. (3). Thus, we minimize the classification loss by

$$\mathcal{L}_{cls} = \mathbb{E}_{\mathbf{x}\sim\mathcal{U}} \, \mathrm{CE}(\mathbf{y}_\mathbf{x}, \hat{\mathbf{y}}_\mathbf{x}) \,, \tag{4}$$

where $\mathcal{U} = \mathcal{S}$ in the unsupervised setting and $\mathcal{U} = \mathcal{S} \cup \mathcal{T}_l$ in the semi-supervised setting, $\mathbf{y}_\mathbf{x}$ and $\hat{\mathbf{y}}_\mathbf{x}$ are the ground-truth and the prediction of $\mathbf{x}$, and $\mathrm{CE}(\cdot,\cdot)$ is the cross-entropy loss.

Therefore, our final objective loss is as follows:

$$\min_{\mathcal{F},\mathcal{F}_p,\mathcal{C}} \max_{\mathcal{D}} \mathcal{L} = \mathcal{L}_{cls} + \alpha \cdot \mathcal{L}_{domain} + \beta \cdot \mathcal{L}_{category} \,, \tag{5}$$

where α and β are trade-off parameters. In the unsupervised setting, we set $\beta = 0$ as the target images are unlabeled.

4. DA-Retail dataset

In the past, there are several benchmark datasets for traditional fine-grained recognition, to name a few: CUB [29], Dog [13] and Aircraft [21], etc. The traditional fine-grained setting does not consider domain shift between source and target domains. Also, some excellent generic benchmarks for the domain adaptation task are proposed, *i.e.*, OpenMic [14] and DomainNet [23]. OpenMic contains photos taken in 10 distinct exhibition spaces of several museums which is more likely to be a generic dataset as the differences among the 866 identities are huge. DomainNet is a generic dataset with 6 different domains as the categories are different, *e.g.*, airplane, axe, and clock. Recently, *GSV Cars* [7] is proposed for fine-grained domain adaptation. It contains 1,095,021 images of 2,657 categories of cars in two domains, while only a small subset of *GSV Cars* is available in experiments when following the protocol proposed in [6].

In order to further facilitate the research of cross-domain fine-grained recognition, we collect, label images and construct a dataset under the retail application, termed *DA-Retail*. *DA-Retail* consists of 52,011 images of 263 fine-grained classes from 3 domains. Abundant images from multiple sources make our dataset more challenging. The collected fine-grained products are from the retail scenario, *e.g.*, instant noodles, fruit juice, mineral water, yogurt, and milk. The data were collected under different domains/conditions:

SKU: All the images taken under standard studio lights are shot under a stable condition and in an ideal environment with ideal resolutions and qualities. **SKU** has 1,870 images taken in the same environment. Each object is shot from 8 different angles.

Shelf: The second is taken on the supermarket shelves. The images are hard to be classified with low resolution and complex backgrounds, while the instances on the supermarket shelves may block with each other. **Shelf** has 1,631 images from different shelves. Obviously, the resolutions of these images are low compared with images from **SKU**.

Web: **Web** is the biggest domain consisting of 23,024 images crawled from the Internet with different resolutions and qualities. Each image may contain several instances if different products are set in a picture. The features of this domain make it a perfect multi-scale dataset, which is harder to train and evaluate on comparing with **SKU** and **Shelf**. Nevertheless, due to the rapidly updated package of per SKU, we found that the instances of most sub-categories in instant noodles are noisy, while the instances of most sub-categories in fruit juice, mineral water, yogurt, and milk are not.

The dataset corresponds to the real-life application in retail, which is equipped with fine-grained and cross-domain natures. It has different domains, enabling us to accomplish several tasks. First, it allows us to research on adaptive models learned on the different domains, which also improves the generalization of models. Second, we can test multi-scale recognition accuracy on the **Web** domain, since it contains images of different resolutions. Also, *DA-Retail* has more

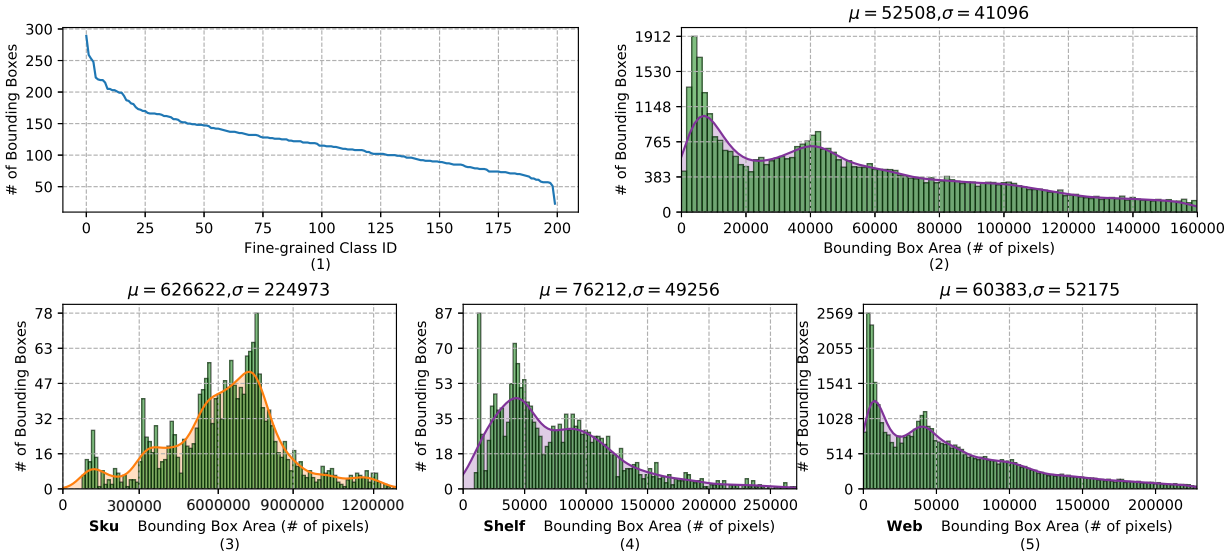

Figure 5. The distribution of *DA-Retail* images for each class (Fig. 5 (1)) used in our evaluation. The subset used in training is balanced. Histogram of *DA-Retail* bounding box sizes is presented in (Fig. 5 (2)). Histogram of bounding box sizes in *SKU*, *Shelf* and *Web* are presented in Fig. 5 (3), (4) and (5). While instances in *Web* and *Shelf* images are typically small (with an average size of 60,383 and 76,212 pixels), those in *SKU* images are larger, occupying an average of 626,622 pixels.

Table 1. Comparisons of datasets **used in experiments**. We evaluate datasets on different perspectives. BBox represents that the dataset has element-wise labels. Multi-label refers that each image may contain instances from different categories. We only report the number of classes, images, and domains available in training and the evaluation.

Datasets	BBox	Multi-label	Fine-grained	♯ classes	♯ images	♯ domains
Office				31	4,110	3
GSV Cars	✓		✓	170	22,344	2
DA-Retail (Ours)	✓	✓	✓	200	24,395	3

domains than the other fine-grained datasets.

Here we present some statistics of out proposed dataset. As shown in Fig. 5, instances of **SKU** are large and typically un-occluded whereas those of **Shelf** are small, blurry and occluded. The number of images per class is presented in Fig. 5 implying that the images in evaluations are balanced. Also, the difference of image size in Fig. 5 shows a histogram of bounding box sizes of **SKU** , **Shelf** 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.

Following the protocols proposed in [6, 25], we choose a subset consisting of the most 200 common classes in the dataset, which enables us to conduct evaluations of our model with enough images.

In unsupervised evaluations, only the labeled images of 200 classes from the source domain are available in training. The test dataset is composed of the labeled images of 200 classes from the target domain.

In semi-supervised evaluations, we split the target data into labeled and unlabeled subsets. The fine-grained classes are sorted in descending order by the number of target images they have. Then, the images of top 50% classes (100 classes) from the target domain, and all the images from the source domain are labeled and used as training data. Evaluation is conducted on the images of the rest classes from the target domain with the least number of labels. Briefly, the labeled images in the semi-supervised training procedure are the top 100 classes containing most of the images from the target domain, and all the images from the source domain.

5. Experiments

In this section, we evaluate the performance of our proposed method with two specific modules on cross-domain fine-grained recognition. We conduct experiments on two fine-grained datasets, *i.e.*, *GSV Cars* [7] and our *DA-Retail* dataset proposed in Sec. 4. Also, we evaluate the performance on the generic image dataset *Office* [25]. The differences among these datasets are presented in Tab. 1.

As aforementioned, the feature extractor and the classi-

Table 2. **Results on our *DA-Retail* in unsupervised adaptation:** "S", "R" and "W" refer the **SKU**, **Shelf** and **Web** domain in *DA-Retail*. "Adapt" and "Attention" mean domain adaptation and self-attention. All the labeled source domain images are used in training.

Method	Adapt	Attention	Acc (%)						
			S→R	S→W	R→S	R→W	W→S	W→R	**Average**
Baseline (CaffeNet) [10]			40.43	13.45	30.30	13.30	28.91	32.26	26.44
DC (CaffeNet) [27]	✓		41.98	14.86	33.33	15.31	33.33	36.76	29.26
Multi-Task (CaffeNet) [6]	✓		46.84	15.38	37.54	15.66	32.64	38.13	31.03
DDC (AlexNet) [28]	✓		43.26	15.43	36.48	15.08	34.22	42.81	31.21
DeepCoral (AlexNet) [26]	✓		47.57	16.77	30.43	16.99	29.43	34.93	29.35
Ours (DA) (CaffeNet)	✓		48.85	16.38	39.27	15.27	34.22	46.56	33.56
Ours (DA+SA) (CaffeNet)	✓	✓	**53.44**	**17.37**	**42.69**	**17.20**	**35.07**	**47.83**	**35.60**
CoGAN [19]	✓		45.72	13.66	45.40	14.21	36.45	39.64	30.61
CMD (VGG16) [36]	✓		47.66	16.59	45.18	16.43	35.06	45.65	34.43
MADA (ResNet) [22]	✓		49.65	19.25	52.91	21.26	41.34	52.93	39.56
iCAN (ResNet) [37]	✓		52.00	20.18	53.38	17.69	**41.60**	47.81	38.78
CADA (ResNet) [15]	✓		51.28	21.41	51.77	18.05	41.55	52.81	39.48
Ours (DA) (ResNet)	✓		50.33	20.59	50.39	17.43	40.06	51.65	38.41
Ours (DA+SA) (ResNet)	✓	✓	**56.10**	**23.69**	**55.18**	**23.69**	**41.60**	**53.16**	**42.24**

Table 3. **Results on our *DA-Retail* in semi-supervised adaptation:** The images from the source domain and the most popular 100 classes in the target domain are available in training.

Method	Adapt	Attention	Acc (%)						
			S→R	S→W	R→S	R→W	W→S	W→R	**Average**
Baseline [10]			24.14	4.65	20.50	4.92	15.38	26.85	16.07
DC [27]	✓		28.83	5.13	23.46	4.47	20.69	29.01	18.60
Multi-Task [6]	✓		37.21	9.84	28.41	8.06	24.05	37.16	24.12
Ours (DA) (CaffeNet)	✓		40.72	9.55	33.75	8.97	28.75	36.94	26.45
Ours (DA+SA) (CaffeNet)	✓	✓	**47.39**	**10.88**	**35.12**	**9.74**	**30.75**	**40.72**	**28.86**

Table 4. **Results on *GSV Cars* in unsupervised and semi-supervised settings:** "DA" and "SA" represent our proposed modules, *i.e.*, domain alignment, and self-attention modules. The best accuracies are presented in bold. The data with "*" refer to the accuracies in the original papers.

Method	Attention	Acc (%)	
		Unsupervised	Semi-supervised
Baseline (CaffeNet) [10]		9.28*	4.72*
DC (CaffeNet) [27]		14.98*	12.34*
Multi-Task (CaffeNet) [6]		19.05*	19.11*
DDC (AlexNet) [28]		15.86	–
DeepCoral (AlexNet) [26]		16.62	–
Ours (DA) (CaffeNet)		20.99	17.36
Ours (DA+SA) (CaffeNet)	✓	**22.61**	**20.13**
CoGAN [19]		19.19	–
CMD (VGG16) [36]		21.81	–
MADA (ResNet) [22]		27.34	–
iCAN (ResNet) [37]		26.61	–
CADA (ResNet) [15]		26.43	–
Ours (DA) (ResNet)		25.55	–
Ours (DA+SA) (ResNet)	✓	**29.71**	–

fier used in all of our experiments are parts of and initialized by CaffeNet and ResNet. We compare our work with Baseline [10], eight domain adaptation methods (DC [27], MADA [22], iCAN [37], CADA [15], DeepCoral [26], CMD [36], DDC [28] and CoGAN [19]) and a fine-grained domain adaptation method (Multi-Task [6]). Among them, Multi-Task [6] is a state-of-the-art of fine-grained domain adaptation method. We set $\alpha = 1.0$ and $\beta = 0$ in the unsupervised setting and $\alpha = 1.0$ and $\beta = 0.1$ in the semi-supervised setting.

5.1. Performance on fine-grained datasets

DA-Retail: The results of unsupervised and semi-supervised evaluations are presented in Tab. 2 and Tab. 3. We can observe similar dramatical improvements in both settings as our model outperforms other methods. Our method increases the average accuracy by 4.57% and 2.58% compared with Multi-task [6] and MADA [22] in the unsupervised setting with CaffeNet [10] and ResNet [9]. In the semi-supervised setting, our model outperforms Baseline [10] and Multi-Task [6] by 12.79% and 4.74%. Some methods [6, 10, 27] have overfitted in the source domain, which aggravates the domain shift and decreases the accuracy, while other domain adaptation methods [15, 22, 37] basically can not capture the key parts essential for the fine-grained task. Our model without SA (Ours(DA)) is only compatible with some domain adaptation methods, which is reasonable as they have more complex training strategies for domain adaptation.

GSV Cars: *GSV Cars* is proposed by [7], consisting of two different domains, Cars and GSV. We follow the protocols in Multi-Task [6] to conduct experiments. The results of unsupervised and semi-supervised experiments on *GSV Cars* are presented in Tab. 4. In the unsupervised setting, DA can increase the accuracy by 1.94%, while SA further boosts the accuracy by 1.62%. Our proposed method improves the accuracy by 3.56% in total compared with Multi-Task [6]. We also conduct semi-supervised experiments. As mentioned in Multi-Task [6], we only use the labeled images of the source domain and half of the target domain (top 85 classes sorted by the number of image). The evaluation is conducted on the rest images of the target domain. Similarly, DA and SA can make progress on the evaluation of *GSV Cars*, with an improvement of 1.02% compared with Multi-Task [6].

Table 5. **Results on the generic image dataset, *i.e.*, *Office* in unsupervised and semi-supervised settings:** "Attention" refers to self-attention. "DA" and "SA" represent our proposed modules, *i.e.*, domain alignment, and self-attention. The best accuracies are presented in bold. The data with "*" refer to the accuracies in the papers.

Method	Attention	Acc (%)	
		Unsupervised	Semi-supervised
Baseline (CaffeNet) [10]		60.9*	45.5*
DC (CaffeNet) [27]		61.1*	47.0*
Multi-Task (CaffeNet) [6]		62.4*	51.8*
DDC (AlexNet) [28]		59.4*	–
DeepCoral (AlexNet) [26]		**66.8***	–
Ours (DA) (CaffeNet)		63.2	51.7
Ours (DA+SA) (CaffeNet)	✓	**64.5**	**52.6**
CoGAN [19]		74.5	–
CMD (VGG16) [36]		77.0*	–
MADA (ResNet) [22]		90.0*	–
iCAN (ResNet) [37]		92.5*	–
CADA (ResNet) [15]		**97.0***	–
Ours (DA) (ResNet)		84.3	74.0
Ours (DA+SA) (ResNet)	✓	**85.0**	**74.2**

5.2. Performance on generic datasets

While our adversarial approach is most suitable in the fine-grained setting, we also conduct the experiments on the generic dataset to show compatible performance. *Office* [25] is a typical generic dataset consisting of 4,110 images from 3 domains, *i.e.*, Amazon, Dslr and Webcam. We investigate unsupervised and semi-supervised settings and follow protocols proposed with the *Office* dataset [25]. The results are presented in Tab. 5. In the unsupervised setting, our method improves performance by 2.1% comparing with the fine-grained domain adaptation state of the art [6]. In the semi-supervised setting, only a subset of source domain are available. As for the images of the target domain, each class from the top 15 of 31 classes has 10 labeled images available in training, while the rest 16 classes are used for evaluation. Our methods still can improve the performance by 0.8% compared with Multi-task [6]. While our methods may not achieve better performances than some domain adaptation methods, our method still outperforms the fine-grained domain adaptation state-of-the-art in all the tasks.

5.3. Evaluating self-attention module

In this section, we use the model trained in three different semi-supervised settings on *DA-Retail* to visualize the masks $\mathbf{M}$ related to the image $\mathbf{x}$ for showing the superiority of our method.

In the cross-domain fine-grained recognition scenario, it requires us to not only capture the slight subtle features to distinguish the instances, but also capture it in both source and target domains, which makes this problem more challenging. Self-attention (SA) perfectly cooperates with domain alignment (DA) and focuses on similar subtle features of the specific category across domains, while it can not work well individually to capture discriminative parts across domains,

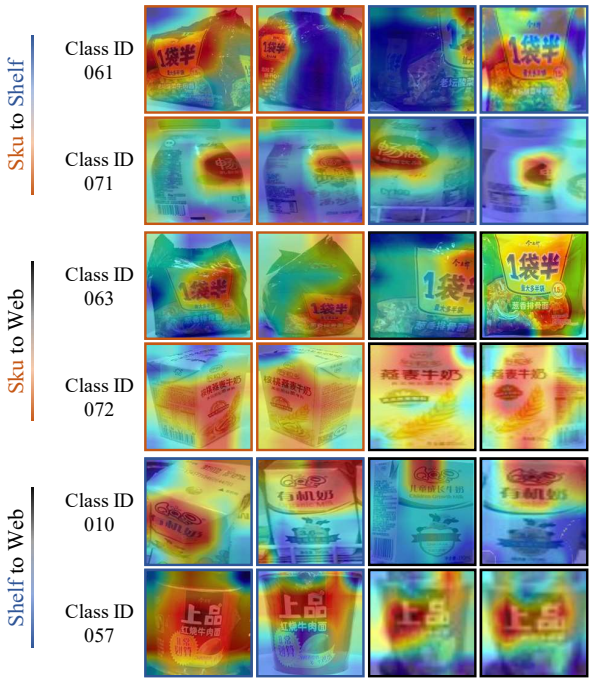

Figure 6. Visualization of the self-attention module. Each column is the images with masks of the same category in different domains. The images surrounded by red, blue and black lines refer to the images in SKU, Shelf and Web. Self-attention module can perfectly capture the same feature across domains cooperating with the domain alignment module.

as shown in Fig. 6. The key slight subtle differences of each categories are activated the most, *e.g.*, the bands, graphs and texts. Also, empirical experiments in Tab. 2 and Tab. 3 imply that SA is essential for the fine-grained recognition task.

6. Conclusions

In this paper, we presented a novel model for cross-domain fine-grained recognition, outperforming existing methods [6, 27] on three different datasets. Our model minimized the discrepancy between different domains, making it more robust under different application views. Furthermore, we proposed a novel dataset for the research on fine-grained domain adaptation. The proposed dataset has 52,011 images of 263 classes from 3 different domains. The huge discrepancy among domains makes it a suitable dataset for this challenge cross-domain fine-grained recognition task.

In the future, it is promising to exploit multiple discriminative fine-grained parts in cross-domain scenarios to further boost the recognition performance. Besides, our *DA-Retail* dataset, source codes and pre-trained models are available at `https://yimuwang96.github.io/DA-Retail/index.html`.

References

[1] 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.

[2] Y. Chai, V. Lempitsky, and A. Zisserman. Symbiotic segmentation and part localization for fine-grained categorization. In *ICCV*, pages 321–328, 2013.

[3] G. Chen, J. Yang, H. Jin, E. Shechtman, J. Brandt, and T. X. Han. Selective pooling vector for fine-grained recognition. In *WACV*, pages 860–867, 2015.

[4] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell. Compact bilinear pooling. In *CVPR*, pages 317–326, 2016.

[5] Z. Ge, A. Bewley, C. McCool, P. Corke, B. Upcroft, and C. Sanderson. Fine-grained classification via mixture of deep convolutional neural networks. In *WACV*, pages 1–6, 2016.

[6] T. Gebru, J. Hoffman, and L. Fei-Fei. Fine-grained recognition in the wild: A multi-task domain adaptation approach. In *ICCV*, pages 1358–1367, 2017.

[7] T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng, and L. Fei-Fei. Fine-grained car detection for visual census estimation. In *AAAI*, pages 4502–4508, 2017.

[8] P. Guo and R. Farrell. Aligned to the object, not to the image: A unified pose-aligned representation for fine-grained recognition. In *WACV*, pages 1876–1885, 2018.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[10] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, pages 675–678, 2014.

[11] C. Kanan. Fine-grained object recognition with gnostic fields. In *WACV*, pages 23–30, 2014.

[12] G. Kang, L. Zheng, Y. Yan, and Y. Yang. Deep adversarial attention alignment for unsupervised domain adaptation: the benefit of target expectation maximization. In *ECCV*, pages 401–416, 2018.

[13] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei. Novel dataset for fine-grained image categorization. In *CVPRW*, pages 1–2, 2011.

[14] P. Koniusz, Y. Tas, H. Zhang, M. Harandi, F. Porikli, and R. Zhang. Museum exhibit identification challenge for the supervised domain adaptation and beyond. In *ECCV*, pages 788–804, 2018.

[15] V. K. Kurmi, S. Kumar, and V. P. Namboodiri. Attending to discriminative certainty for domain adaptation. In *CVPR*, pages 491–500, 2019.

[16] S. Li, S. Song, G. Huang, Z. Ding, and C. Wu. Domain invariant and class discriminative feature learning for visual domain adaptation. *TIP*, 27(9):4260–4273, 2018.

[17] Y. Li, N. Wang, J. Shi, J. Liu, and X. Hou. Revisiting batch normalization for practical domain adaptation. In *ICLRW*, pages 1–12, 2017.

[18] T.-Y. Lin, A. RoyChowdhury, and S. Maji. Bilinear CNN models for fine-grained visual recognition. In *ICCV*, pages 1449–1457, 2015.

[19] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In *NeurIPS*, pages 469–477, 2016.

[20] M. Long, Y. Cao, J. Wang, and M. I. Jordan. Learning transferable features with deep adaptation networks. In *ICML*, pages 97–105, 2015.

[21] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

[22] Z. Pei, Z. Cao, M. Long, and J. Wang. Multi-adversarial domain adaptation. In *AAAI*, pages 3934–3941, 2018.

[23] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang. Moment matching for multi-source domain adaptation. In *ICCV*, pages 1406–1415, 2019.

[24] S. Reed, Z. Akata, H. Lee, and B. Schiele. Learning deep representations of fine-grained visual descriptions. In *CVPR*, pages 49–58, 2016.

[25] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, pages 213–226, 2010.

[26] B. Sun and K. Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, pages 443–450, 2016.

[27] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *ICCV*, pages 4068–4076, 2015.

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

[29] C. Wah, S. Branson, P. Welinder, and S. B. Pietro Perona. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[30] X.-S. Wei, Q. Cui, L. Yang, P. Wang, and L. Liu. RPC: A large-scale retail product checkout dataset. *arXiv preprint arXiv:1901.07249*, 2019.

[31] X.-S. Wei, J.-H. Luo, J. Wu, and Z.-H. Zhou. Selective convolutional descriptor aggregation for fine-grained image retrieval. *IEEE TIP*, 9(3):303–315, 2017.

[32] X.-S. Wei, J. Wu, and Q. Cui. Deep learning for fine-grained image analysis: A survey. *arXiv preprint arXiv:1907.03069*, 2019.

[33] X.-S. Wei, C.-W. Xie, and J. Wu. Mask-CNN: Localizing parts and selecting descriptors for fine-grained image recognition. *PR*, 76:704–714, 2018.

[34] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *CVPR*, pages 842–850, 2015.

[35] S. Yang, L. Bo, J. Wang, and L. G. Shapiro. Unsupervised template learning for fine-grained object recognition. In *NeurIPS*, pages 3122–3130, 2012.

[36] W. Zellinger, T. Grubinger, E. Lughofer, T. Natschläger, and S. Saminger-Platz. Central moment discrepancy (CMD) for domain-invariant representation learning. In *ICLR*, 2017.

[37] W. Zhang, W. Ouyang, W. Li, and D. Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *CVPR*, pages 3801–3809, 2018.

[38] X. Zhang, F. Zhou, Y. Lin, and S. Zhang. Embedding label structures for fine-grained feature representation. In *CVPR*, pages 1114–1123, 2016.