Progressive Adversarial Networks for Fine-Grained Domain Adaptation

Sinan Wang, Xinyang Chen, Yunbo Wang, Mingsheng Long (✉), and Jianmin Wang
School of Software, BNRist, Tsinghua University, China
Research Center for Big Data, Tsinghua University, China
Beijing Key Laboratory for Industrial Big Data System and Application
{wang-sn17,chenxiny17,wangyb15}@mails.tsinghua.edu.cn, {mingsheng,jimwang}@tsinghua.edu.cn

Abstract

Fine-grained visual categorization has long been considered as an important problem, however, its real application is still restricted, since precisely annotating a large fine-grained image dataset is a laborious task and requires expert-level human knowledge. A solution to this problem is applying domain adaptation approaches to fine-grained scenarios, where the key idea is to discover the commonality between existing fine-grained image datasets and massive unlabeled data in the wild. The main technical bottleneck lies in that the large inter-domain variation will deteriorate the subtle boundaries of small inter-class variation during domain alignment. This paper presents the ***Progressive Adversarial Networks (PAN)*** *to align fine-grained categories across domains with a curriculum-based adversarial learning framework. In particular, throughout the learning process, domain adaptation is carried out through all multi-grained features, progressively exploiting the label hierarchy from coarse to fine. The progressive learning is applied upon both category classification and domain alignment, boosting both the discriminability and the transferability of the fine-grained features. Our method is evaluated on three benchmarks, two of which are proposed by us, and it outperforms the state-of-the-art domain adaptation methods.*

1. Introduction

Fine-grained recognition aims to categorize an object among a large number of subordinate categories within the same root category. It is a valuable problem in the sense that it could potentially endow machine learning models with strong cognitive abilities approaching human experts on some tasks. For example, we might be interested in distinguishing subordinate species of birds such as *pacific gull* or *black-tailed gull*. In recent years, there has been great advance in some fundamental problems of fine-grained recognition. On one hand, the ability of deep networks for identifying subtle differences between highly similar objects has been greatly improved [23, 43, 45, 16, 46, 25, 13]. On the other hand, an increasing number of fine-grained image datasets have been collected, including a variety of root categories such as birds [55, 54, 50], dogs [18, 27], flowers [36, 1], aircrafts [52, 32], cars [44, 22, 26, 58], and food [4].

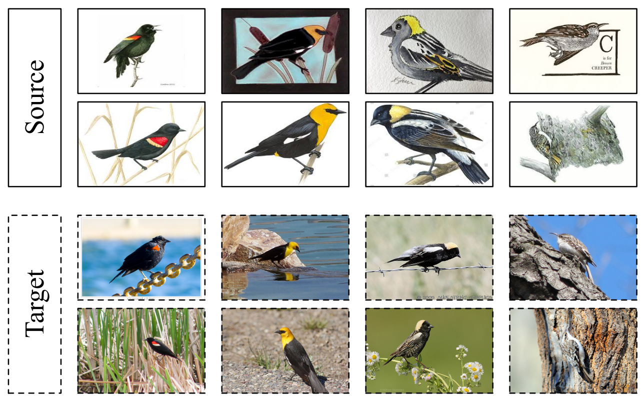


Figure 1. The fine-grained domain adaptation problem is characterized by the entanglement of **large inter-domain variations, small inter-class variations, and large intra-class variations**, which differs from classic domain adaptation scenario in granularity. From left to right: *red-winged blackbirds, yellow-headed blackbirds, bobolinks*, and *brown creepers*.

Still, it is unrealistic to cover all subordinate categories, and the limited size of the existing datasets still hampers the scalability of the fine-grained recognition algorithms. Annotating large-scale image datasets with fine-grained labels is time-consuming and requires strong expertise, especially for some particular application domains. To solve this problem, a promising idea is to apply the domain adaptation approaches [38] to fine-grained recognition tasks. For example, learning to categorize birds in the field guide may help recognize bird species in the wild, as illustrated in Figure 1. Thus, we may transfer the common knowledge from existing labeled datasets to massive unlabeled data, and save the efforts of dense fine-grained annotations.

However, there are new challenges in the context of **fine-grained domain adaptation**: the concurrence of large

Table 1. Comparisons among different fine-grained transfer methods (fewer requirements are better) (−: partially).

Method	require image-level labels?		require additional annotations?		
	source domain	target domain	attributes	object bounding boxes	part landmarks
Gebru *et al.* [14]	✓	−	✓	✗	✗
Xu *et al.* [57]	✓	−	✗	✓	✓
Cui *et al.* [9]	✓	✓	✗	✗	✗
Ours	✓ (hierarchical)	✗	✗	✗	✗

inter-domain variations, small inter-class variations, and large intra-class variations. The classic domain adaptation algorithms overcome the **inter-domain** variations by making images from different domains have similar distributions in the feature space [40, 28, 47, 12]. When it comes to the fine-grained domain adaptation, the situation becomes more complicated in that we have to confront tough issues brought by the fine-grained categorization. A combination of large **intra-class** variations and small **inter-class** variations may deteriorate the inter-class boundaries, and thus make the classic domain adaptation algorithms fail in respectively mapping objects of neighboring categories from the source domain to the target domain. As in Figure 1, *yellow-headed blackbirds* and *bobolinks* are perceptually similar and may be mismatched across domains.

This paper aims to address these challenges by designing a new fine-grained domain adaptation method, and presents **Progressive Adversarial Networks (PAN)**. In fine-grained scenarios, natural objects have taxonomic ranks in biology, and man-made objects also have reasonable hierarchical labels. The general idea is integrating curriculum learning [3] and adversarial learning [15] to enable domain adaptation progressively from coarse-grained categories (easy) to fine-grained categories (difficult). This disentangles the difficulties by large inter-domain variations, small inter-class variations, and large intra-class variations. The training process of our model only depends on *hierarchical category labels* on the source domain. We evaluate our method on three benchmarks, two of which are proposed by us, based on several existing datasets for fine-grained visual categorization and one brand-new dataset we collect from the web and filter manually. We demonstrate that the proposed approach outperforms the state-of-the-art domain adaption methods.

2. Related Work

2.1. Fine-Grained Visual Categorization

In recent years, fine-grained visual categorization has become a prevalent problem in computer vision. As it requires expertise to recognize the subtle differences between the subordinate categories within the same root category, some methods introduced additional labels such as part-annotations and visual attributes to enhance fine-grained recognition [5, 59, 60, 53, 14].

Instead of using the cost-prohibitive part-annotations or additional attributes, some work attempted to improve the fine-grained recognition performance in other ways. Krause *et al.* [20] tried to solve the fine-grained recognition problem by generating parts using co-segmentation and alignment. Lin *et al.* [25] proposed a two-stream CNN model based on the bilinear pooling, which is also trained with category labels. Gao *et al.* [13] presented the compact bilinear pooling method as an extension of [25] to lower the computation complexity while retaining comparable accuracy. Other variants of the original bilinear pooling method were soon proposed and applied on the neural network models for fine-grained recognition [24, 19]. Dubey *et al.* introduced confusion in the activations [10] and revisited Maximum-Entropy [11]. To further alleviate the difficulty of collecting expert-level annotations manually, some methods were proposed to make the fine-grained recognition models benefit from the large-scale but noisy web data [21, 14, 57].

The above methods achieved fairly good performance even without part-annotations, but yet, their scalability is limited by the lack of fine-grained annotations for the vast subordinate categories in the real world.

2.2. Domain Adaptation

Domain adaptation is to transfer knowledge from the source domain to the target domain, which saves the cost of manual annotations [38]. The discrepancy between the source domain and the target domain causes the main difficulties for knowledge transfer. To learn domain-invariant, transferable features, some work proposed different adaptation layers based on deep networks [49, 28, 30]. Some more recent work studied the domain-adversarial methods, which incorporated the adversarial learning [15] into the domain adaptation framework [47, 12]. These models aligned the feature distributions of different domains by trying to fool the domain discriminator. Through further conditioning the adversarial adaptation models on the discriminative information in class predictions, CDAN [29] sheds light into the direction in addressing the problem of fine-grained cross-domain recognition. PFAN [7] adopts an "Easy-to-Hard Transfer Strategy" to select easy samples from the target domain and align these pseudo-labeled samples with their corresponding source categories. A basic distinction from our work is that PFAN learns from progressive samples without exploring the granularity information, while our method learns from progressive granularity diametrically.

These methods above are insightful. Unfortunately, all of them were not specifically designed for fine-grained cross-domain visual categorization and did not explore the label hierarchy specific in fine-grained recognition scenarios.

2.3. Fine-Grained Domain Adaptation

Fine-grained domain adaptation was first studied by Gebru *et al.* [14]. They proposed a model that was trained with annotated web images and evaluated with real-world data, using the domain adaptation approach proposed in [47] and requiring additional annotations of *attributes*. Only when labeled images on the target domain are available, can its *semi-supervised* adaptive loss be performed, which is only a tailored design for fine-grained domain adaptation.

With a unique design of exploiting strong supervision, Xu *et al.* [57] utilized detailed annotations including object *bounding boxes* and *part landmarks*, in addition to standard image-level labels. As much knowledge as possible was transferred from existing strongly supervised datasets to weakly supervised web images.

Additionally, Cui *et al.* [9] achieved obvious improvements on several fine-grained visual benchmarks, by fine-tuning well-performing CNNs pre-trained on the large-scale iNaturalist2017 dataset [51]. They proposed a measure to estimate domain similarity and selected a subset from the source domain that is more similar to target domain.

All above methods obtained encouraging performance. However, the problem setups are completely different from ours, as in Table 1. Our approach does not require attributes, bounding boxes or part landmarks but relies on the hierarchical labels that are easier to obtain in fine-grained tasks. To our knowledge, our work is the first method designed for *unsupervised* fine-grained domain adaptation only depending on hierarchical image-level labels from source domain.

3. Method

In the fine-grained domain adaptation problem, we are given a source domain $\mathcal{S} = \{(x, y_f, y_c^k|_{k=1}^K)\}$ of n_s examples with both fine-grained label y_f and coarse-grained labels $\{y_c^k\}_{k=1}^K$ in a K-layer class hierarchy, and a target domain $\mathcal{T} = \{(x, ?, ?)\}$ of n_t unlabeled examples. There is a large discrepancy between the joint distributions $P(x, y)$ and $Q(x, y)$ of source domain and target domain respectively. Due to the distribution shift, a fine-grained recognition model trained on $\mathcal{S}$ cannot perform accurately on $\mathcal{T}$.

The domain adversarial networks [12] are performant domain adaptation models. They usually consist of three network modules: the feature extractor F, the domain discriminator D and the label predictor Y. A combination of F and Y is trained with recognition objectives (only with labels from the source domain). Simultaneously, to extract domain transferable features, F and D work together and play an adversarial game. The domain discriminator D is trained to distinguish the source domain from the target domain, while the feature extractor F is trained to confuse D, keeping it away from making correct judgments.

The **Progressive Adversarial Network (PAN)** exploits *hierarchical labels* of fine-grained objects. As opposed to fine-grained labels in the bottom layer of the label hierarchy, we refer to higher-level labels in the label hierarchy as coarse-grained labels. Fine-grained domain adaptation is very *difficult* due to its large inter-domain variations, small inter-class variations, and large intra-class variations. In contrast, coarse-grained domain adaptation is *easy*. Inspired by curriculum learning [3], we begin with the easy granularity, and then progressively move to the difficult granularity. An accurate sup-class alignment across domains works as a solid foundation for sub-class alignment.

3.1. Progressive Granularity Learning

In progressive granularity learning (PGL), we progressively change the granularity of supervision for the recognition task on the source domain from coarse-grained to fine-grained during training. We replace fine-grained ground-truth labels with dynamic mixing of fine-grained ground-truth labels and coarse-grained *predicted distributions* given by the recognition model trained at the corresponding granularity, denoted as *progressive labels*, as shown in Figure 2. Predicted distributions convey information of relationships between classes, which are considered beneficial to domain adaptation [47]. The coarse-grained labels could be more than one levels, say K ($K \geq 1$). The *coarse-grained CNN*, an auxiliary network effective at training and will be removed at inference, is introduced with a feature extractor G and K label predictors $C^k, k = 1, ..., K$. The data point x with coarse-grained labels $y_c^k, k = 1, ..., K$ is fed into the coarse-grained CNN. The coarse-grained CNN is trained on the source domain by minimizing the recognition objective:

$$\sum_{k=1}^{K} L_y\left(\widehat{y}_c^k, y_c^k\right), \tag{1}$$

where $\widehat{y}_c^k = C^k\left(G\left(x\right)\right)$ is the k-th coarse-grained predicted distribution and L_y is the cross-entropy (CE) loss.

The fine-grained labels of all images are explored by the *fine-grained CNN*, which is trained by minimizing a novel **coarse-fine hybrid loss** we propose in this paper:

$$L_h(\widehat{y}_c^k|_{k=1}^K, \widehat{y}_f, y_f) = D_{\mathrm{KL}}\left(\varepsilon y_f + (1-\varepsilon)\sum_{k=1}^{K}\frac{\widehat{y}_c^k}{K}\middle\|\, \widehat{y}_f\right), \tag{2}$$

where D_{KL} is the Kullback-Leibler divergence, and $\widehat{y}_f = Y\left(F\left(x_i\right)\right)$ is the fine-grained predicted distribution, y_f is the corresponding ground-truth label. Besides, $\widehat{y}_c^k$ has been extended to the same dimension as $\widehat{y}_f$, according to a class subordination strategy as illustrated in Figure 3.

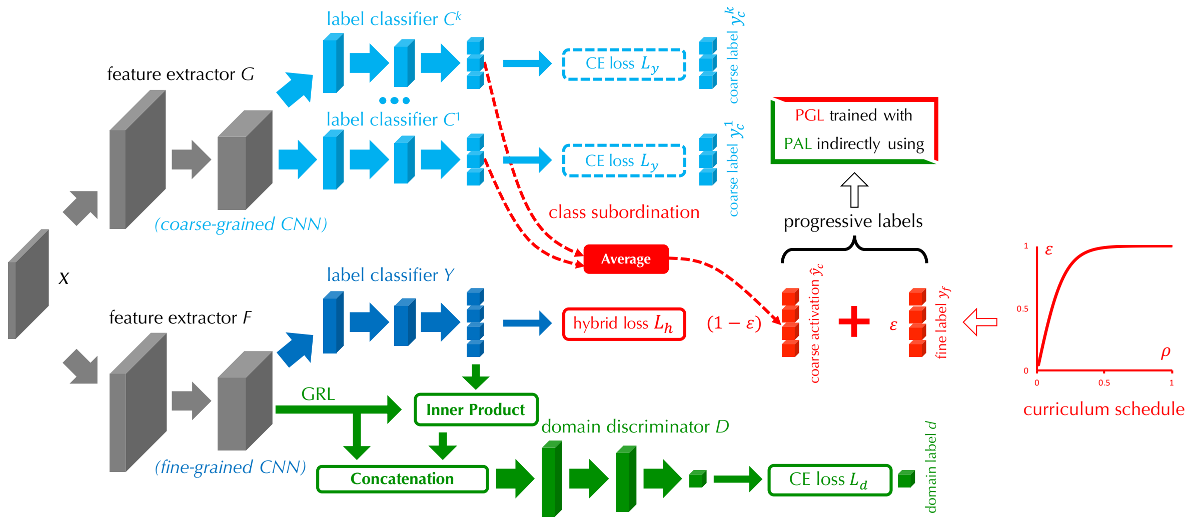


Figure 2. Progressive Adversarial Networks (PAN). A shared feature extractor G and K label predictors $C^k|k=1,...,K$ together form the coarse-grained CNN (top) for coarse-grained recognition. Similarly, the fine-grained CNN (bottom) contains a feature extractor F and a label predictor Y. Progressive Granularity Learning (PGL, red): the fine-grained ground-truth label y_f and the coarse-grained predicted distributions $\widehat{y}_c$ are mixed by ratio ε following a well-established schedule [12] for curriculum learning. We combine the coarse-grained and fine-grained labels into progressive labels via class subordination in label hierarchy (details in Figure 3). Progressive Adversarial Learning (PAL, green): the coarse-to-fine classifier predictions (supervised by progressive labels) are combined with the feature representations by inner product and residual connection, and fed into a domain discriminator D. GRL is the gradient reversal layer [12].

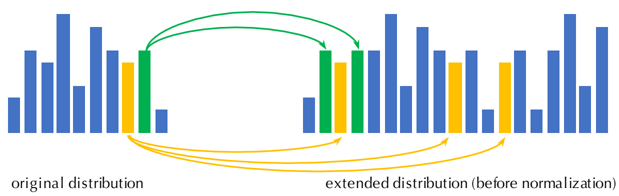


Figure 3. The probability of each coarse-grained class $\widehat{y}_c^k$ is duplicated to the corresponding positions of its subordinate fine-grained classes. The extended elements are normalized into a probability.

During training, ε is progressively changed from 0 to 1, following the common curriculum schedule studied in [12]:

$$\varepsilon = \frac{1-\exp\left(-10\rho\right)}{1+\exp\left(-10\rho\right)}, \tag{3}$$

where ρ is the ratio of the current number to the maximum number of iterations. Eventually, with the decay of ε, the influence of coarse-grained labels will vanish and the *coarse-fine hybrid loss* will converge to the *fine-grained loss*:

$$L_h(\widehat{y}_c^k|_{k=1}^K, \widehat{y}_f, y_f) = D_{\mathrm{KL}}\left(y_f \| \widehat{y}_f\right), \tag{4}$$

which plays the same role as the cross-entropy (CE) loss.

The reason why the *coarse-fine hybrid loss* works can be analyzed from another perspective. **First**, if the label predictor is too confident in its outputs during training, it may be over-fitting. Since the inter-class variations are small and intra-class variations are large, it is unreasonable for the label predictor to assign full probability to the ground-truth label. **Second**, intemperate fitting of ground-truth labels may disrupt the transferability of the features, as subtle relationships between classes are destroyed. These subtle relationships are expected to play an important role in domain adaptation when small inter-class variations greatly heighten the uncertainty of domain alignment.

3.2. Progressive Adversarial Learning

While progressive granularity learning (PGL) enables the supervised task on the source domain to progressively move from coarse to fine, *it does not necessarily guarantee that the hierarchical classes are aligned across domains progressively, first sup-classes and then sub-classes*. Fortunately, the predicted distribution $\widehat{y} = Y(F(x))$ conveys important discriminative information. And it is progressive as Y and F are trained to minimize *coarse-fine hybrid loss*. Inspired by incorporating the conditional information into the discriminator of GANs [35], we feed the progressive label $\widehat{y}$ with the feature $f = F(x)$ into the domain discriminator D to enable progressive adversarial learning (PAL).

While it is natural to choose the concatenation of $\widehat{y}$ and f as the input to the domain discriminator D, as adopted by conditional GANs [34, 42, 17, 37], such a fusion strategy is not expressive enough to model the complex relationship between $\widehat{y}$ and f. Ingeniously, CDAN [29] employs outer product to replace the concatenation. However, the exploding feature dimension requires excessive memory.

To enable progressive adversarial learning, we employ a different *bilinear* transformation to combine the predicted distribution $\widehat{y}$ and the feature f. However, though feature embeddings with predicted class information can enhance discriminability, it has side effect that may completely destroy the subtle differences between features in fine-grained scenarios, especially when the sample is misclassified. And misclassification is more likely to occur in fine-grained tasks. Thus it is necessary to additionally introduce a *residual connection* by concatenating the features. Finally, the fusion result is fed to the domain discriminator D:

$$\text{Bi}\left(\widehat{y}, f\right) = \left(\widehat{y}^T A f + b\right) \oplus f, \tag{5}$$

where A and b are the learnable weight and bias of the bilinear transformation, and $\oplus$ is the residual concatenation.

3.3. Progressive Adversarial Network

The architecture of **Progressive Adversarial Network (PAN)** is shown in Figure 2. The coarse-grained predictors $C^k|_{k=1}^K$, the fine-grained label predictor Y, and the domain discriminator D are jointly trained by unifying progressive granularity learning and progressive adversarial learning:

$$\begin{aligned} & O\left(G, C^k|_{k=1}^K, F, Y, D\right) \\ & \quad = \frac{1}{n_s} \sum_{x \in \mathcal{S}} \sum_{k=1}^{K} L_y\left(C^k\left(G\left(x\right)\right), y_c^k\right) \\ & \quad + \frac{1}{n_s} \sum_{x \in \mathcal{S}} L_h\big(C^k\left(G\left(x\right)\right)|_{k=1}^K, Y\left(F\left(x\right)\right), y_f\big) \\ & \quad - \frac{\lambda}{n} \sum_{x \in \mathcal{S} \cup \mathcal{T}} L_d\left(D\left(\text{Bi}\left(Y\left(F\left(x\right)\right), F\left(x\right)\right)\right), d\right), \end{aligned} \tag{6}$$

where d is the domain label of x, λ is a hyperparameter, and $n = n_s + n_t$. L_y is the cross-entropy loss for coarse-grained recognition, which is minimized by G and $C^k|_{k=1}^K$. L_h is the proposed *coarse-fine hybrid loss* for fine-grained recognition, minimized by Y and F. L_d is the cross-entropy loss for domain discrimination, minimized by D and maximized by F. Eventually G, $C^k|_{k=1}^K$, F, Y, D converge to:

$$\begin{aligned} (\widehat{G}, \widehat{C}^k|_{k=1}^K) &= \underset{G, C^k|_{k=1}^K}{\arg\min}\, O\left(G, C^k|_{k=1}^K, F, Y, D\right), \\ (\widehat{F}, \widehat{Y}) &= \underset{F, Y}{\arg\min}\, O\left(G, C^k|_{k=1}^K, F, Y, D\right), \\ (\widehat{D}) &= \underset{D}{\arg\max}\, O\left(G, C^k|_{k=1}^K, F, Y, D\right). \end{aligned} \tag{7}$$

Compared with previous domain adversarial networks, PAN enables domain alignment in fine-grained sub-classes by progressively aligning the feature distributions across domains from coarse-grained to fine-grained. Note that a correct coarse-grained alignment of super-classes is the basis of corresponding alignments of fine-grained sub-classes. Though the auxiliary multi-task network seems to make the architecture complicated (in Figure 2), all branches of PAN are closely coordinated and indispensable. In the inference phase, all components are discarded except the feature extractor F and the label classifier Y in the fine-grained CNN.

3.4. Theoretical Explanation

Coarse-grained models have higher generalization performance and domain adaptation on coarse-grained categories is easier. Dubey *et al.* [11] defines the diversity of features as the sum of the eigenvalues of the equivalent covariance matrix. And fine-grained problems are characterized as feature distributions with the following property:

$$\sqrt{\nu(\Phi^F, P_x^F)} \ll \sqrt{\nu(\Phi^G, P_x^G)}, \tag{8}$$

where ν denotes the diversity of the features, P_x^F is the fine-grained data distribution yielded by feature extractor Φ^F, and P_x^G is the generic data distribution yielded by feature extractor Φ^G. *The diversity in fine-grained visual categorization tasks is considered to be smaller than coarse-grained tasks.* The norm of weights Θ of the final classifier layer is lower bounded by [11] (H is the entropy):

$$||\Theta||_2 \geq \frac{\log\left(C\right) - E_{x \sim P_x}\left(H\left(P\left(\cdot|x;\Theta\right)\right)\right)}{2\sqrt{\nu(\Phi, P_x)}}. \tag{9}$$

Unlike coarse-grained categorization tasks, the fine-grained tasks generally have smaller diversity of features, which will enlarge the norm of the classifier weights Θ and make learning much more difficult. Hence, we introduce both the progressive granularity learning and progressive adversarial learning to enable domain adaptation from coarse-grained (easy) to fine-grained (hard). This progressive strategy can lower the difficulty of fine-grained domain adaptation.

4. Experiments

We evaluate the proposed PAN with state-of-the-art image classification and domain adaptation models based on deep learning architectures. Experiments are conducted on an existing benchmark **CompCars** [58] and two brand-new benchmarks we construct, **CUB-Paintings** and **Birds-31**.

Table 2. Number of images in the three benchmarks.

Dataset	Domains	#Images
CompCars [58]	Web Surveillance	33,780 44,481
CUB-Paintings	CUB-200-2011 CUB-200-Painting	11,788 3,047
Birds-31	CUB-200-2011 NABirds iNaturalist2017	1,848 2,988 2,857

4.1. Datasets

4.1.1 Benchmark I: CompCars

We evaluate PAN on **CompCars** [58] which can be split into two domains: Web (**W**) and Surveillance (**S**), as shown in Table 2. Only two levels of classes are available: 281

Figure 4. Examples for the first 31 categories from *CompCars*: *Web* (top) and *Surveillance* (bottom).

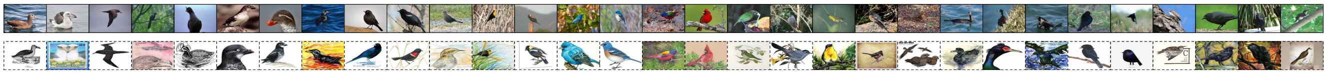


Figure 5. Examples for the first 31 categories from *CUB-Paintings*: *CUB-200-2011* (top) and *CUB-200-Paintings* (bottom).

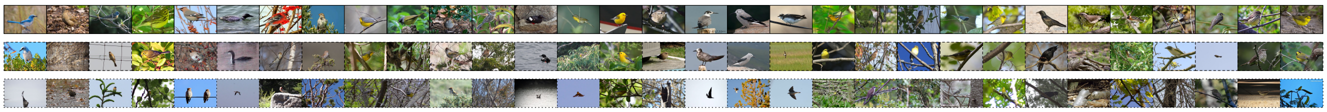


Figure 6. Examples for the 31 categories from *Birds-31*: *CUB-200-2011* (top), *NABirds* (middle) and *iNaturalist2017* (bottom).

`models` (finer) and 68 `Makes` (coarser). Figure 4 are example images of the first 31 categories (models).

Notice that the car surveillance images are all in front views, with various environment conditions such as foggy, and at night, which are quite different from web images, indicating that the transfer task **S** $\rightarrow$**W** is extremely challenging. This may not be conducive to the evaluation of methods. So we construct another two novel benchmarks.

4.1.2 Benchmark II: CUB-Paintings

CUB-Paintings contains two domains: CUB-200-2011 (**C**) and CUB-200-Paintings (**P**), as shown in Table 2. Figure 5 are example images of the first 31 categories, with obvious visual domain gap. Images are organized in a four-levels hierarchy. From finer to coarser, there are 200 `Species`, 122 `Genera`, 38 `Families`, and 14 `Orders`.

CUB-200-2011 [54] is a fine-grained visual categorization benchmark with $11,788$ bird images in 200 species.

CUB-200-Paintings is a dataset of bird paintings we collect from the web and filter manually. The class lists of CUB-200-Paintings and CUB-200-2011 are identical. We search Internet to collect candidate images for a total of 200 classes. Both the English common name and binomial name are adopted as retrieval keywords. Watercolors, oil paintings, pencil drawings, stamps, and cartoons are all within the scope of being selected. Then candidate images are further filtered manually. Only paintings with obvious species characteristics or with reliable labels are retained. However, this dataset needs further polishing. $3,047$ images are insufficient for training very deep models, considering there are 200 categories. Potential label noise needs to be eliminated.

4.1.3 Benchmark III: Birds-31

There are three domains in Birds-31: CUB-200-2011 (**C**), NABirds (**N**) and iNaturalist2017 (**I**). Not all of the images from the original datasets are incorporated into Birds-31. The numbers of images selected are $1,848$, $2,988$ and $2,857$ respectively. Figure 6 shows example images of all 31 categories from Birds-31. In contrast to CUB-Paintings, inter-domain variations of Birds-31 are relatively smaller. Labels are in four levels. Specifically, there are 31 `Species`, 25 `Genera`, 16 `Families`, and 4 `Orders`.

NABirds [50] is a fine-grained visual categorization dataset, composed of $48,000$ images in 400 species.

iNaturalist2017 [51] is a benchmark for iNaturalist 2017 competition.There are $5,089$ categories in it, with $579,184$ training images and $95,986$ validation images.

We employ binomial nomenclature to categorize objects from these three datasets, and then get the intersection, 123 categories. As Benchmark II contains up to 200 categories and numbers of samples vary greatly across domains, 31 categories with a balanced sample size are selected finally.

4.2. Implementation

We implement all deep methods in PyTorch and we use NVIDIA Titan RTX for training. We fine-tune ResNet-50 [16] model pre-trained on ImageNet. The classifier layers are trained from scratch, and their learning rate is set 10 times that of the other layers. We adopt mini-batch SGD with momentum of 0.9. Batch size is fixed to 36. The learning rate strategy is the same as [12]. Consistent with [12], hyperparameter λ changes from 0 to 1 following a schedule of $\lambda = \frac{1-\exp(-10\rho)}{1+\exp(-10\rho)}$ in all experiments. For fair comparison, all parameters are not changed across all transfer tasks.

4.3. Results

We evaluate Progressive Adversarial Networks (**PAN**) and report the average classification accuracy based on three random experiments. In addition to the widely-used baseline Domain Adversarial Neural Network (**DANN**) [12], we also compare PAN with generic visual categorization, fined-grained visual categorization and domain adaptation methods: **ResNet-50** [16], **Inception-v3** [46], **Bilinear CNN** [25], Deep Adaptation Network (**DAN**) [28], Joint Adaptation Network (**JAN**) [31], Adversarial Discriminative Domain Adaptation (**ADDA**) [48], Multi-Adversarial Domain Adaptation (**MADA**) [39], Maximum Classifier Discrepancy (**MCD**) [41], Conditional Adversarial Domain Adaptation (**CDAN**) [29], Batch Spectral Penalization (**BSP**) [8], and Stepwise Adaptive Feature Norm (**SAFN**) [56].

Table 3. Classification accuracy (%) on *Birds-31* (ResNet-50).

Method	C → I	I → C	I → N	N → I	C → N	N → C	Avg
ResNet-50 [16]	64.25±0.28	87.19±0.15	82.46±0.45	71.08±0.23	79.92±0.21	89.96±0.29	79.14
Inception-v3 [46]	62.09±0.49	86.20±0.52	79.88±0.17	68.00±0.16	76.79±0.22	90.42±0.22	77.23
Bilinear CNN [25]	64.82±0.39	88.43±0.30	83.37±0.43	71.37±0.48	79.86±0.25	91.22±0.37	79.85
DAN [28]	63.90±0.49	85.86±0.66	82.91±0.60	70.67±0.33	80.64±0.48	89.40±0.23	78.90
DANN [12]	64.59±0.34	85.64±0.29	80.53±0.25	71.00±0.24	79.37±0.24	89.53±0.19	78.44
JAN [31]	63.69±0.99	86.29±0.25	83.34±0.20	71.09±0.48	81.06±0.39	89.55±0.38	79.17
ADDA [48]	63.03±0.42	87.26±0.25	84.36±0.47	72.39±0.31	79.69±0.11	89.28±0.26	79.33
MADA [39]	62.03±0.37	89.99±0.21	87.05±0.29	70.99±0.17	81.36±0.40	92.09±0.25	80.50
MCD [41]	66.43±0.44	88.02±0.28	85.57±0.25	73.06±0.43	82.37±0.19	90.99±0.17	81.07
CDAN [29]	68.67±0.25	89.74±0.45	86.17±0.26	73.80±0.17	83.18±0.28	91.56±0.24	82.18
CDAN+BSP [8]	68.64±0.37	89.71±0.26	85.72±0.32	74.11±0.16	83.22±0.33	91.42±0.45	82.13
SAFN [56]	65.23±0.26	90.18±0.32	84.71±0.35	73.00±0.40	81.65±0.21	91.47±0.08	81.08
PAN (Proposed)	**69.79±0.10**	**90.46±0.35**	**88.10±0.08**	**75.03±0.18**	**84.19±0.15**	**92.51±0.31**	**83.34**

Table 4. Accuracy (%) on *CompCars* (ResNet-50).

Method	W → S	S → W	Avg
ResNet-50 [16]	34.22±0.20	5.93±0.22	20.08
Inception-v3 [46]	29.74±0.17	4.58±0.31	17.16
Bilinear CNN [25]	36.51±0.23	6.74±0.35	21.63
DAN [28]	33.73±0.29	11.70±0.24	22.72
DANN [12]	33.67±0.32	12.38±0.12	23.02
JAN [31]	44.16±0.18	11.01±0.26	27.59
ADDA [48]	34.01±0.27	12.96±0.30	23.49
MADA [39]	41.77±0.20	11.89±0.29	26.83
MCD [41]	40.25±0.37	13.66±0.42	26.96
CDAN [29]	42.37±0.21	14.56±0.17	28.47
CDAN+BSP [8]	43.35±0.34	14.91±0.15	29.13
SAFN [56]	41.75±0.36	14.29±0.25	28.02
PAN	**47.05±0.12**	**15.57±0.23**	**31.31**

Table 5. Accuracy (%) on *Cub-Paintings* (ResNet-50).

Method	C → P	P → C	Avg
ResNet-50 [16]	47.88±0.31	36.62±0.23	42.25
Inception-v3 [46]	51.59±0.21	40.72±0.15	45.88
Bilinear CNN [25]	54.09±0.35	41.59±0.57	47.84
DAN [28]	58.95±0.43	39.33±0.35	49.14
DANN [12]	57.54±0.38	43.01±0.29	50.28
JAN [31]	62.42±0.29	40.37±0.39	51.40
ADDA [48]	60.12±0.31	40.65±0.17	50.36
MADA [39]	63.67±0.23	44.28±0.30	53.98
MCD [41]	63.40±0.65	43.63±0.77	53.52
CDAN [29]	63.18±0.16	45.42±0.25	54.30
CDAN+BSP [8]	63.27±0.19	46.62±0.39	54.95
SAFN [56]	61.38±0.33	48.86±0.35	55.12
PAN	**67.40±0.02**	**50.92±0.26**	**59.16**

On CompCars as in Table 4, our method performs best across both transfer tasks. It outperforms CDAN+BSP, the second best method, by 2.1 percent. **On CUB-Paintings** as in Table 5, our method achieves the best performance across all two transfer tasks. We raise average accuracy from the baseline DANN of 50.28% to 59.16%, a boost of more than 8 percent. **On Birds-31** as in Table 3, our method achieves the highest average accuracy and the best performance across all six tasks. The accuracy is improved by about 5 percent compared to DANN.

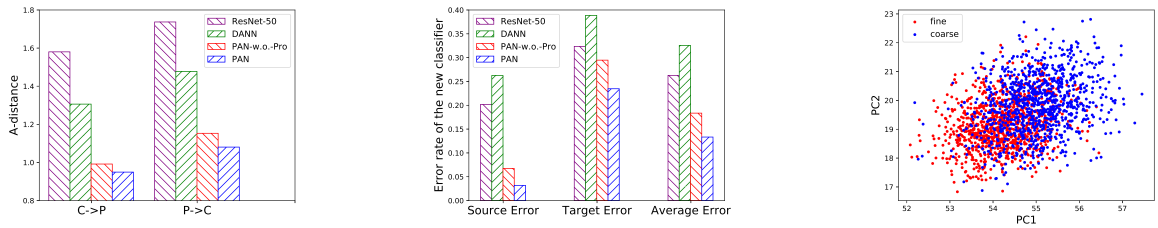
Note that PAN yields larger boosts on CompCars and CUB-Paintings than on Birds-31. There are two reasons. First, the inter-domain variations of the former are much larger than the later, as shown in Figures 4, 5, and 6. Small inter-domain variations imply less gain by bridging the domain gap. Second, the classification accuracy of Birds-31 is generally higher, leaving us with a relatively smaller room for improvement. For example, in task **N** → **C**, the accuracy of most methods is about 90%. And, as some neighboring categories are visually indistinguishable, the performance of expert annotators is only 93% [6].

4.4. Analyses

Ablation Study. Removing PGL and preserving PAL, we denote the remainder of PAN by PAN-w.o.-Pro. or PAL Only. Note that without PGL, PAL Only is not *progressive* any more. The accuracy decreases sharply with PAL Only (Table 6). We also testify the concatenation operator in PAL. PAL outperforms PAL (w/o concat), proving that the concatenation operator can prevent the model from destroying subtle differences between features. PGL Only still outperforms the baseline DANN by 4 percents.

Hierarchy Selection. PAN exploits labels at all levels. On the dataset CUB-Paintings, coarse-grained labels are at three levels: `genus`, `family`, and `order`. We analyzed the variants of PAN with coarse-grained labels at only one level, with improved results by PAN shown in Table 7.

Curriculum Schedule. The curriculum schedule that



(a) Distribution Discrepancy (b) Ideal Joint Hypothesis (c) Feature Diversity

Figure 7. Analyses of cross-domain distribution discrepancy, error of ideal joint hypothesis, and diversity of fine-grained features.

the ε in Equation (2) follows is the same as that of λ in Equation (6). This simple and commonly-used strategy [12] outperforms the others, as shown in Table 8.

Table 6. Ablation Study: Accuracy (%) on *CUB-Paintings*.

Method	C $\rightarrow$ P	P $\rightarrow$ C	Avg
PAL (w/o concat)	62.46±0.30	45.32±0.37	53.89
PAL Only	63.05±0.19	45.83±0.33	54.44
PGL Only	61.04±0.29	46.69±0.12	53.87
PAN (PGL+PAL)	**67.40±0.02**	**50.92±0.26**	**59.16**

Table 7. Accuracy (%) of PAN with different coarse-grained label levels on *CUB-Paintings*.

Level	Num	C $\rightarrow$ P	P $\rightarrow$ C	Avg
Genus	122	65.37±0.46	48.33±0.35	56.85
Family	38	65.51±0.37	48.02±0.16	56.76
Order	14	66.32±0.34	49.43±0.23	57.88
Class	1	64.68±0.23	46.92±0.30	55.80
G+F+O	**–**	**67.40±0.02**	**50.92±0.26**	**59.16**

Table 8. Accuracy (%) of PAN with different curriculum strategies on *CUB-Paintings*.

Schedule	**Ours**	Linear	Step	Exponential
Avg	**59.16**	54.67	54.88	55.19

Distribution Discrepancy. In domain adaptation theory [2, 33], A-distance is a measure of inter-domain variation:

$$d_A = 2(1 - 2\mathrm{err}), \tag{10}$$

where err is the error rate of a classifier that is trained to discriminate the source domain and the target domain. Figure 7(a) depicts d_A on transfer tasks **C** $\rightarrow$**P** and **P** $\rightarrow$**C**, with features extracted by ResNet-50, DANN, PAN-w.o.-Pro. and PAN. It is notable that d_A on features extracted by PAN are the smallest on both transfer tasks, which implies that these features are more transferable across domains.

Ideal Joint Hypothesis. The expected error $\mathcal{E}_{\mathcal{T}}(h)$ of a hypothesis h on the target domain can be bounded as [2]

$$\mathcal{E}_{\mathcal{T}}(h) \leq \mathcal{E}_{\mathcal{S}}(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T}) + \mathcal{E}_{ideal}, \tag{11}$$

where $\mathcal{E}_{\mathcal{S}}(h)$ is the source error, $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T})$ is the $\mathcal{H}\Delta\mathcal{H}$-distance measuring the domain shift, and $\mathcal{E}_{ideal}$ is the error of an ideal joint hypothesis $h^* = \min_h \mathcal{E}_{\mathcal{S}}(h) + \mathcal{E}_{\mathcal{T}}(h)$ on labeled source and target domains. $\mathcal{E}_{ideal}$ is defined as

$$\mathcal{E}_{ideal} = \mathcal{E}_{\mathcal{S}}(h^*) + \mathcal{E}_{\mathcal{T}}(h^*), \tag{12}$$

which measures the discriminability of features. For further analysis of our method, we investigate this indicator of discriminability. The average error rate of the new classifier trained on the labeled data of source and target domains is half of $\mathcal{E}_{ideal}$. The results are shown in Figure 7(b). As expected, PAN enhances the discriminability of features.

Feature Diversity. Figure 7(c) is a plot of the top 2 principal components (PCs) of features of ResNet-50 trained from fine-grained (red) and coarse-grained (blue) labels on CUB-200-2011, following the experiment in [11]. Fine-grained features are concentrated with less diversity, in accordance with the theoretical analysis in Section 3.4.

Weight Sharing. The feature extractors should not share weights. Differences between fine-grained features are subtle and sharing weights destroys the subtle differences crucial for discriminability. Using weight sharing, the average accuracy on CUB-Paintings drops from 59.16% to 51.48%.

5. Conclusion

In this paper, we proposed the Progressive Adversarial Networks (PAN) to solve the fine-grained domain adaptation problem with only hierarchical image-level labels. The key idea of our model is to align the corresponding classes across domains from coarse-grained to fine-grained, first sup-classes and then sub-classes. We also theoretically explained the proposed approach from the perspective of feature diversity. We compared PAN with prior works on three benchmarks for fine-grained domain adaptation. And experimental results testified the effectiveness of our method.

6. Acknowledgments

This work was supported in part by the Natural Science Foundation of China (61772299, 71690231), and the MOE Strategic Research Project on Artificial Intelligence Algorithms for Big Data Analysis.

References

[1] Anelia Angelova, Shenghuo Zhu, and Yuanqing Lin. Image segmentation for large-scale subcategory flower recognition. In *WACV Workshops*, pages 39–45, 2013.

[2] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

[3] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, pages 41–48, 2009.

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, pages 446–461, 2014.

[5] Steve Branson, Grant Van Horn, Serge Belongie, and Pietro Perona. Bird species categorization using pose normalized deep convolutional nets. *arXiv preprint arXiv:1406.2952*, 2014.

[6] Steve Branson, Grant Van Horn, Catherine Wah, Pietro Perona, and Serge Belongie. The ignorant led by the blind: A hybrid human–machine vision system for fine-grained categorization. *International Journal of Computer Vision*, 108(1-2):3–29, 2014.

[7] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. In *CVPR*, 2019.

[8] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *ICML*, 2019.

[9] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *CVPR*, pages 4109–4118, 2018.

[10] Abhimanyu Dubey, Otkrist Gupta, Pei Guo, Ramesh Raskar, Ryan Farrell, and Nikhil Naik. Pairwise confusion for fine-grained visual classification. In *ECCV*, pages 71–88, 2018.

[11] Abhimanyu Dubey, Otkrist Gupta, Ramesh Raskar, and Nikhil Naik. Maximum-entropy fine grained classification. In *NeurIPS*, pages 635–645, 2018.

[12] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[13] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *CVPR*, pages 317–326, 2016.

[14] Timnit Gebru, Judy Hoffman, and Li Fei-Fei. Fine-grained recognition in the wild: A multi-task domain adaptation approach. In *ICCV*, pages 1358–1367, 2017.

[15] Ian J. Goodfellow, Jean Pougetabadie, Mehdi Mirza, Bing Xu, David Wardefarley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *NeurIPS*, pages 2672–2680, 2014.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[17] Phillip Isola, Junyan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 5967–5976, 2017.

[18] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *CVPR Workshops*, volume 2, page 1, 2011.

[19] Shu Kong and Charless Fowlkes. Low-rank bilinear pooling for fine-grained classification. In *CVPR*, pages 7025–7034, 2017.

[20] Jonathan Krause, Hailin Jin, Jianchao Yang, and Li Fei-Fei. Fine-grained recognition without part annotations. In *CVPR*, pages 5546–5555, 2015.

[21] Jonathan Krause, Benjamin Sapp, Andrew Howard, Howard Zhou, Alexander Toshev, Tom Duerig, James Philbin, and Li Fei-Fei. The unreasonable effectiveness of noisy data for fine-grained recognition. In *ECCV*, pages 301–320, 2016.

[22] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *CVPR Workshops*, pages 554–561, 2013.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1097–1105, 2012.

[24] Tsungyu Lin and Subhransu Maji. Improved bilinear pooling with cnns. In *BMVC*, 2017.

[25] Tsung-Yu Lin, Aruni RoyChowdhury, Subhransu Maji, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *CVPR*, pages 1449–1457, 2015.

[26] Yen-Liang Lin, Vlad I Morariu, Winston Hsu, and Larry S Davis. Jointly optimizing 3d model fitting and fine-grained classification. In *ECCV*, pages 466–480, 2014.

[27] Jiongxin Liu, Angjoo Kanazawa, David Jacobs, and Peter Belhumeur. Dog breed classification using part localization. In *ECCV*, pages 172–185, 2012.

[28] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, pages 97–105, 2015.

[29] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *NeurIPS*, pages 1647–1657, 2018.

[30] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *NeurIPS*, pages 136–144, 2016.

[31] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *ICML*, pages 2208–2217, 2017.

[32] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[33] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *COLT*, 2009.

[34] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[35] Takeru Miyato and Masanori Koyama. cgans with projection discriminator. In *ICLR*, 2018.

[36] M-E Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In *CVPR*, volume 2, pages 1447–1454, 2006.

[37] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *ICML*, pages 2642–2651, 2017.

[38] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[39] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Multi-adversarial domain adaptation. In *AAAI*, 2018.

[40] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, pages 213–226, 2010.

[41] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, pages 3723–3732, 2018.

[42] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In *ICCV*, volume 2, page 5, 2017.

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[44] Michael Stark, Jonathan Krause, Bojan Pepik, David Meger, James J Little, Bernt Schiele, and Daphne Koller. Fine-grained categorization for 3d scene understanding. *International Journal of Robotics Research*, 30(13):1543–1552, 2011.

[45] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.

[46] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016.

[47] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *ICCV*, pages 4068–4076, 2015.

[48] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, pages 2962–2971, 2017.

[49] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[50] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *CVPR*, pages 595–604, 2015.

[51] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, pages 8769–8778, 2018.

[52] Andrea Vedaldi, Siddharth Mahendran, Stavros Tsogkas, Subhransu Maji, Ross Girshick, Juho Kannala, Esa Rahtu, Iasonas Kokkinos, Matthew B Blaschko, David Weiss, et al. Understanding objects in detail with fine-grained attributes. In *CVPR*, pages 3622–3629, 2014.

[53] Andrea Vedaldi, Siddharth Mahendran, Stavros Tsogkas, Subhransu Maji, Ross Girshick, Juho Kannala, Esa Rahtu, Iasonas Kokkinos, Matthew B Blaschko, David Weiss, et al. Understanding objects in detail with fine-grained attributes. In *CVPR*, pages 3622–3629, 2014.

[54] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[55] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010.

[56] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *ICCV*, 2019.

[57] Zhe Xu, Shaoli Huang, Ya Zhang, and Dacheng Tao. Webly-supervised fine-grained visual categorization via deep domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1100–1113, 2018.

[58] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *CVPR*, pages 3973–3981, 2015.

[59] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *ECCV*, pages 834–849, 2014.

[60] Ning Zhang, Evan Shelhamer, Yang Gao, and Trevor Darrell. Fine-grained pose prediction, normalization, and recognition. In *CVPR*, 2015.