IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 17, 2021, accepted October 5, 2021, date of publication October 8, 2021, date of current version October 15, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3118786

Adversarial Networks With Circular Attention Mechanism for Fine-Grained Domain Adaptation

NINGYU HE AND JIE ZHU
Shanghai Jiao Tong University, Shanghai 200240, China
Corresponding author: Ningyu He (ningyuhe_ruby@163.com)

This work was supported by the Shanghai Frontier Science Research Center for Gravitational Wave Detection.

ABSTRACT Fine-grained Image Analysis (FGIA) as a branch of the image analysis tasks has received more and more attention in recent years. Compared with ordinary image analysis tasks, FGIA requires more detailed human data annotation, which not only requires the annotator to have professional knowledge, but also requires greater labor costs. An effective solution is to apply the domain adaptation (DA) method to transfer knowledge from existing fine-grained image datasets to massive unlabeled data. This paper presents the circular attention mechanism to cyclically extract deep-level image features to match the label hierarchy from coarse to fine. What is more, the networks effectively improve the distinguishability and transferability of fine-grained features based on the adversarial learning framework. Experimental results show that our proposed method achieves excellent transfer performance on three fine-grained recognition benchmarks.

INDEX TERMS Fine-grained, domain adaptation, image recognition, attention, adversarial learning.

I. INTRODUCTION

Fine-grained Image Analysis (FGIA) is called sub-category image analysis which aims to categorize an object among a large number of subordinate categories within the same meta-category. In previous FGIA tasks, the dataset required manual annotation by professionals, which required a great time cost and manpower. Therefore, people try to use machine learning models as a substitute for fine-grained image recognition and annotation. However, different from general image analysis tasks, different sub-category images in FGIA tasks may have the similar shape, size and even textures. The huge intra-class differences and subtle inter-class differences in FGIA tasks bring challenges to mainstream machine learning models.

To address this issue, people have made many efforts and achieved great advance in fine-grained recognition tasks in recent years. On one hand, many researches [1], [2] are dedicated to extracting local discriminative features to improve the ability of deep networks for identifying subtle differences between similar fine-grained image samples. On the other hand, the number of fine-grained image datasets has increased significantly in recent years, which includes different sample types such as birds [3] [4], flowers [5], [6], cars [7]–[10], dogs [11], [12], etc.

The associate editor coordinating the review of this manuscript and approving it for publication was Shagufta Henna.

Still, it is unrealistic to allow the labels of fine-grained images to cover all the datasets on demand. Therefore, scholars try to use computers to replace human experts for fine-grained annotation of images in large-scale data sets. One promising way is to apply the domain adaptation approaches [13] to fine-grained recognition tasks. For example, we may transfer the knowledge from existing labeled birds datasets to massive unlabeled birds images in the wild to save the tedious fine-grained annotation work.

However, fine-grained domain adaptation algorithms face great challenges in many aspects. In fine-grained domain adaptation tasks, we not only have to face the common problem of inter-domain distribution differences in domain adaptation algorithm, but also have to solve the problems of huge intra-class differences and subtle inter-class differences that are unique in fine-grained domain adaptation tasks. The traditional image domain adaptation algorithm [14]–[16] usually establishes the connection between the two domains by finding the correlation between the source domain and the target domain in the feature space, and thus achieves the purpose of reducing the inter-domain distribution differences. But 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. As shown in Figure 1, birds under different fine labels may have similar characteristics, such as similar feather colors and bird beak shapes. This makes it difficult

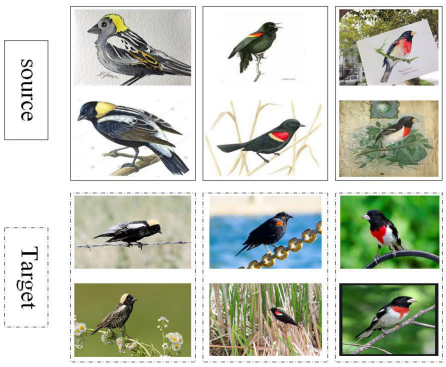


FIGURE 1. Birds under different fine labels in different datasets.

for feature-based domain adaptation algorithms to achieve satisfactory results.

This paper aims to address these challenges by designing adversarial networks with circular attention mechanism for fine-grained domain adaptation. We use the attention mechanism to locate the most discriminative regions in images. Furthermore, the circular attention mechanism is designed to locate multi discriminative regions for fine-grained image analysis tasks by recursively dropping the previous discriminative region and adopting the attention mechanism again. The general idea of our domain adaptation method is to extract the fine features in the fine-grained images from the multi discriminative regions learned in the circular attention mechanism, and use the adversarial learning network to enable domain adaptation progressively from coarse-grained categories to fine-grained categories.

We evaluate our method on three benchmarks. Two of them are based on the domain adaptation of bird images, including the CUB-200-2011 [17], CUB-200-Painting [18], NABirds [4] and iNaturalist2017 [19] datasets, and the other is based on the domain adaptation of vehicle images, including the Stanford [8] dataset. The extensive experimental results show that the proposed adversarial networks with circular attention mechanism achieve excellent performance in fine-grained domain adaptation tasks.

The rest of this paper is organized as follow. Section II gives a brief description on related work. In Section III, the adversarial networks with circular attention mechanism are introduced in details. Section IV provides the comparison experiments and ablation experiments on three different benchmarks. Finally, we conclude the paper in Section V.

II. RELATED WORK

A. FINE-GRAINED IMAGE CLASSIFICATION

In recent years, fine-grained image classification as the basis of fine-grained image tasks, has received more and more attention in the field of computer vision. Since the differences between the fine-grained categories are subtle, traditional CNN networks are difficult to obtain features that are sufficient to support fine-grained image classification.

To address this issue, researchers have proposed three solutions. The first solution is to enhance the fine-grained classification ability by introducing additional labels such as partial annotations and visual attributes to the images [20]–[23]. Another solution is to improve the feature representation ability of the network. For example, Lin *et al.* [24] proposed a bilinear model to fuse features in different dimensions of images to obtain features that are more suitable for fine-grained image recognition tasks. On the basis of this article, Gao *et al.* [25] proposed a compact bilinear pooling method, which reduces the computational complexity. The third and which is the most mainstream method is to locate the position of the object to be classified in the image, so that the CNN networks can provide more refined features. Hu *et al.* [26] first proposed attention mechanisms to locate the object. Similarly, Yang *et al.* [27] proposed the Region Proposal Network (RPN) which concatenates original features and partial features together to do the object location. The above methods have achieved fairly good performance in fine-grained image classification tasks.

B. DOMAIN ADAPTATION

Domain adaptation problem is a representation method in transfer learning, which aims to use the labeled data (source domain) to learn the classifier and use it to predict the label of the unlabeled data (target domain). The most commonly used method for domain adaptation is to transform the data features of the source domain and the target domain into a unified feature space through feature transformation, so as to reduce the discrepancy between two domains [15], [28] [29]. Pan *et al.* [30] proposed Transfer Component Analysis (TCA) method which uses the Maximum Mean Discrepancy (MMD) [31] as a metric to minimize the distribution discrepancy between the source and the target domain. In recent years, feature-based domain adaptation methods are usually combined with neural networks. Long *et al.* [32] integrated the idea of adversarial learning into the domain migration algorithm and proposed Conditional Adversarial Domain Adaptation (CADN) method. The above methods have made great contributions to domain adaptation algorithms, but unfortunately, none of them are aimed at fine-grained images adaptation tasks. Due to the ignorance of the hierarchical labeling of fine-grained images, these methods are difficult to achieve satisfactory results in the task of fine-grained domain adaptation.

III. PROPOSED METHOD

In this section, our proposed adversarial networks with circular attention mechanism is introduced in details. Some mathematical notation is set to interpret our method. In the fine-grained domain adaptation task, a source domain is given as $S = \left\{\left(x, y_f, y_c^k \,|_{k=1}^K\right)\right\}$ with both fine label y_f and coarse label $\left\{y_c^k\right\}_{k=1}^K$ in a K-layer class hierarchy. On contrast, a target domain T is consisted of n_t unlabeled examples. The joint

distributions on the source and target domains are denoted as $P(x, y)$ and $Q(x, y)$ respectively.

A. CIRCULAR ATTENTION MECHANISM

The overall framework of our networks is shown in Figure 2. Networks with circular attention mechanism are designed to extract the fine features in the fine-grained images from the multi discriminative regions learned in the circular attention mechanism.

We first introduce the circular attention mechanism which is shown in Figure 3. In our work, we adopt attention mechanism on bilinear pooling to train the attention maps. The bilinear pooling was first adopted by Lin *et al.* [24] to improve the performance on fine-grained image classification tasks. The algorithm flow of circular attention mechanism is summarized in Algorithm 1.

Algorithm 1: Circular Attention Mechanism

Input: Input image $I = R, i = 1, K$(K-layer class hierarchy), threshold δ

Output: Attention areas $\{L_1, L_2, \ldots, L_n\}$

while $i < K$ **do**

1. Generate attention maps A with spatail attentional bilinear pooling;
2. Generate mA by averaging the attention maps on channels;
3. Binarize mA according the threshold δ:

$$M = \begin{cases} 0 & if\ mA < \delta \\ 1 & if\ mA > \delta \end{cases}$$

4. Locate the discriminative region and sample local image L_i from the raw image R;
5. Generate the drop image D by dropping the discriminative region in the input image I;
6. $i \leftarrow i + 1, I \leftarrow D$

By iteratively dropping the previous discriminative region from the raw image, the circular attention mechanism could propose a set of local images $\{L_1, L_2, \ldots, L_n\}$ from high to low information. It is natural to associate these local images with the class hierarchy of fine-grained labels. This is also in line with human's cognitive habits. In order to distinguish fine-grained differences in images, people may focus more attention on the details of the object. The circular attention mechanism filters out the background of the images and selects local images that have received more attention in the raw image to assist the fine-grained classification.

B. PROGRESSIVE GRANULARITY LEARNING

After extracting the local images from the circular attention mechanism, progressive granularity learning method [18] is used to complete the training from coarse-grained to fine-grained for the recognition tasks. As shown in Figure 2, the coarse labels are divided into K levels. CNN is introduced with a coarse feature extractor G and K label predictors $C^k, k = 1, 2, \cdots, K$. The image data x with coarse labels $y_c^k, k = 1, 2, \cdots, K$ is fed into the coarse-grained CNN and trained on the source domain by minimizing the cross-entropy loss as follows:

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

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

On the other hand, the fine labels of the images are explored by the fine feature extractor F and fine label predictor Y, which is trained by minimizing a cross-fine hybird loss:

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

where D_{KL} is the Kullback-Leibler divergence, $\hat{y}_f = Y(F(x))$ is the fine predicted distribution and y_f is the corresponding truth label. During training, ε follows the formula to change from 0 to 1 [33]:

$$\varepsilon = \frac{1 - e^{-10\rho}}{1 + e^{-10\rho}}$$

where ρ is the ratio of the training iteration progress. As the training progresses, ε gradually approaches to 1. Thus the influence of coarse labels disappears and the cross-fine hybird loss converges to the fine-grained loss:

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

which plays the same role as the CE loss.

C. ADVERSARIAL LEARNING

After progressive granularity learning, the domain adversarial networks are used for domain adaptation. We first establish the relationship between the predicted distribution $\hat{y}$ and the fine feature $f = F(x)$. In this paper, we employ a bilinear transformation with residual connection [34] to combine the $\hat{y}$ and the feature f. Embedding the feature with the predicted class information can enhance discriminative. What is more, the residual connection retains the subtle differences between the features in the fine-grained images. The bilinear transformation is expressed as follows:

$$Bi\left(\hat{y}, f\right) = \left(\hat{y}^T A F + b\right) \oplus f$$

where A and b are the weight and bias of the bilinear transformation, $\oplus$ represents the residual connection. Parameters A and b are both learned from the following adversarial learning.

Common domain adversarial networks are consist of three network modules: the feature extractor F, the domain discriminator D and the label predictor Y. F and Y is trained to extract transferable features. At the same time, F and D conduct adversarial training. D aims to distinguish the source

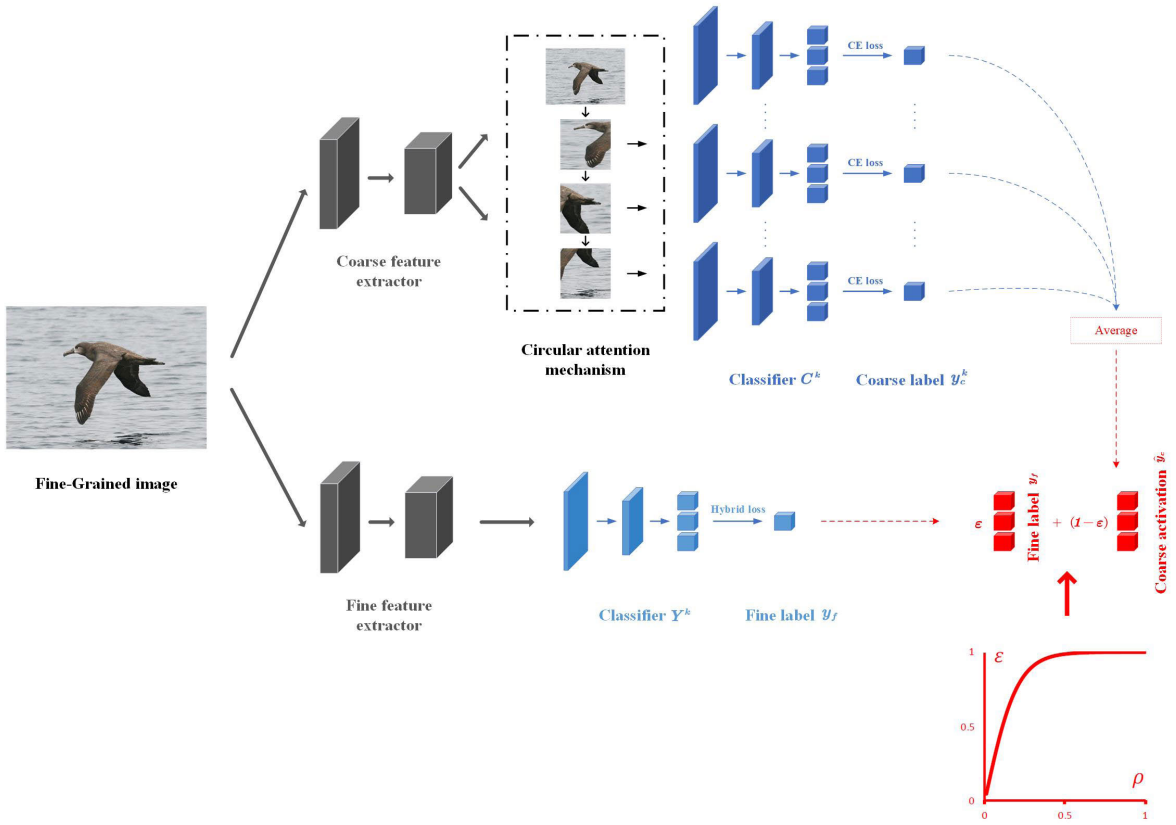


FIGURE 2. Networks with circular attention mechanism.

domain from the target domain, while F is trained to keep the D away from making correct judgments. In our method, the coarse predictors $C^k \mid_{k=1}^{K}$, the fine label predictor Y and the domain discriminator D are trained for adversarial learning. The overall loss of the network is as follows.

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

where λ is a hyperparameter, d is the domain label of x and $n = n_s + n_t$ is the total sample size of the source and target domains. The overall loss of the network can be divided into three parts as shown in the formula. First is the L_y, which is the cross-entropy loss for coarse recognition and is minimized by G and $C^k \mid_{k=1}^{K}$. Second is the L_h, which is the coarse-fine hybird loss for fine recognition and is minimized by Y and F. Both two losses are introduced in the previous sections. The last part is L_d, which is the cross-entropy loss for domain discrimination and is minimized by F. The adversarial training of the network is to reduce these three losses at the same time to obtain better fine-grained recognition accuracy. Compared

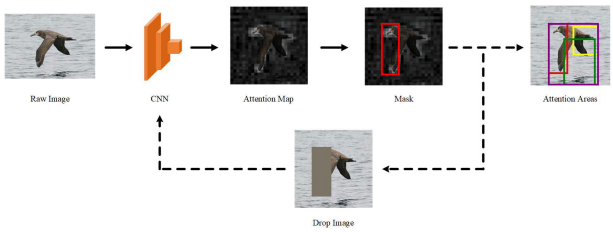


FIGURE 3. Circular attention mechanism.

with previous domain adversarial networks, our networks can gradually align the feature distribution between domains from coarse-grained to fine-grained.

IV. EXPERIMENTS

We evaluate the proposed Adversarial Networks with Circular Attention Mechanism to state-of-the-art domain adaptation models on three benchmarks. Table 1 records the dataset sources of the three benchmarks and the specific number of images. Figure 4, 5 and 6 show example images for fine-grained domain adaptation task in the three benchmarks.

A. DATASETS

1) BENCHMARK I: CompCars

Benchmark I (**CompCars**) is composed of two fine-grained image datasets about cars. One is Stanford (S) [8] dataset introduced by Jonathan Krause *et al.* The other is collected by

TABLE 1. The dataset sources of the three benchmarks.

Benchmarks	Datasets	Images
CompCars	Stanford	8,144
	Web	8,041
CUB-Birds	CUB-200-2011	3,788
	CUB-200-Painting	3,047
Birds-31	NABirds	2,988
	iNaturalist2017	2,857

Stanford

Web

FIGURE 4. Images in Benchmark I: CompCars.

CUB-200-2011

CUB-200-Painting

FIGURE 5. Images in Benchmark II: CUB-Birds.

us from the Web (W). Images in both datasets are organized in three levels. From coarser to finer, the same 42 Make, 72 Model and 20 Year are collected. For example, for a specific image, the label is like 2012 Tesla Model S.

2) BENCHMARK II: CUB-BIRDS

Benchmark II (**CUB-Birds**) contains two domains for fine-grained birds images domain adaptation: CUB-200-2011 (C) [17] and CUB-200-Painting (P) [18]. CUB-200-2011 is the most widely used dataset for fine-grained visual categorization task. It contains 11,788 images of 200 subcategories belonging to birds. Each image has detailed annotations: 15 Part Locations, 312 Binary Attributes, 1 Bounding Box. CUB-200-Painting is a dataset of bird paintings introduced in [18]. The category lists of the two datasets are consistent. However, the CUB-200-Painting dataset only contains 3,047 images. In order to balance the amount and the distribution between the two domains, 3,788 images in the CUB-200-2011 are selected as the domain images.

NABirds

iNaturalist2017

FIGURE 6. Images in Benchmark III: Birds-31.

3) BENCHMARK III: BIRDS-31

Benchmark III (**Birds-31**) can also be split into two domains: NABirds (N) [4] and iNaturalist2017 (I) [19]. Not all of the images from the two datasets are incorporated into the Benchmark. 31 categories with a balanced sample size are selected and the labels are in four levels. Specifically, there are 31 Species, 25 Genera, 16 Families, and 4 Orders.

B. IMPLEMENTATION AND RESULTS

All comparison experiments are carried out on Pytorch. We finetune the ResNet-50 [34] model pretrained on ImageNet. For the fairness of the experiments, the parameters in all domain adaptation tasks are kept consistent and unchanged. Mini-batch SGD with momentum of 0.9 is adopted as the optimization function and batch size is fixed to 36. The learning rate strategy is the same as [33].

We evaluate the proposed Adversarial Networks with Circular Attention Mechanism and record the average classification accuracy after the domain adaptation based on three benchmarks. Several fined-grained visual categorization and domain adaptation methods are selected for comparison experiments, including ResNet-50 [34], Inception-v3 [35], Bilinear CNN [24], Deep Adaptation Network (DAN) [15], Domain Adversarial Neural Network (DANN) [33], joint Adap-tation Network (JAN) [36], Adversarial Discriminative Domain Adaptation (ADDA) [37], Multi-Adversarial DomainAdaptation (MADA) [38], Conditional Adversarial Domain Adaptation (CDAN) [32] and Batch Spectral Penalization (BSP) [39].

As shown is Table 2, our method performs best across both transfer tasks on **CompCars**. It outperforms the second best method CDAN+BSP by more than 1.5% on average accuracy. We raise average accuracy from the baseline DANN of 73.03% to 80.66%, an increase of 7 percent. Similarly, the experimental results on **CUB-Birds** and **Birds-31** are recorded in Table 3 and Table 4. On **CUB-Birds**, our method achieves the best performance among all domain adaptation methods. The accuracy is improved by more than 8% compared to the baseline DANN. Our algorithm also achieved the best performance on **Birds-31** with an accuracy rate of 6.3% higher than the baseline DANN and 2.2% higher than the second best method.

TABLE 2. Accuracy(%) on Benchmark I: CompCars.

Methods	$S \to W$	$W \to S$	Avg
ResNet-50	74.22 ± 0.20	65.93 ± 0.22	70.08
Inception-v3	69.74 ± 0.17	64.58 ± 0.31	67.16
Bilinear CNN	76.51 ± 0.23	66.74 ± 0.35	71.63
DAN	73.73 ± 0.29	71.70 ± 0.24	72.72
DANN	73.67 ± 0.32	72.38 ± 0.12	73.03
JAN	84.16 ± 0.18	71.01 ± 0.26	77.59
ADDA	74.01 ± 0.27	72.96 ± 0.30	73.49
MADA	81.77 ± 0.20	71.89 ± 0.29	76.83
CDAN	82.37 ± 0.21	74.56 ± 0.17	78.47
CDAN + BSP	83.35 ± 0.34	74.91 ± 0.15	79.13
Our method	$\mathbf{84.40 \pm 0.02}$	$\mathbf{76.92 \pm 0.26}$	**80.66**

TABLE 3. Accuracy(%) on Benchmark II: CUB-Birds.

Methods	$C \to P$	$P \to C$	Avg
ResNet-50	47.88 ± 0.31	36.62 ± 0.23	42.25
Inception-v3	51.59 ± 0.21	40.72 ± 0.15	45.88
Bilinear CNN	54.09 ± 0.35	41.59 ± 0.57	47.84
DAN	58.95 ± 0.43	39.33 ± 0.35	49.14
DANN	57.54 ± 0.38	43.01 ± 0.29	50.28
JAN	62.42 ± 0.29	40.37 ± 0.39	51.40
ADDA	60.12 ± 0.31	40.65 ± 0.17	50.36
MADA	63.67 ± 0.23	44.28 ± 0.30	53.98
CDAN	63.18 ± 0.16	45.42 ± 0.25	54.30
CDAN + BSP	63.27 ± 0.19	46.62 ± 0.39	54.95
Our method	$\mathbf{67.05 \pm 0.12}$	$\mathbf{49.57 \pm 0.23}$	**58.31**

TABLE 4. Accuracy(%) on Benchmark III: Birds-31.

Methods	$N \to I$	$I \to N$	Avg
ResNet-50	71.08 ± 0.23	82.46 ± 0.45	76.77
Inception-v3	68.00 ± 0.16	79.88 ± 0.17	73.94
Bilinear CNN	71.37 ± 0.48	83.37 ± 0.43	77.37
DAN	70.67 ± 0.33	82.91 ± 0.60	76.79
DANN	71.00 ± 0.24	80.53 ± 0.25	75.77
JAN	71.09 ± 0.48	83.34 ± 0.20	77.22
ADDA	72.39 ± 0.31	84.36 ± 0.47	78.38
MADA	70.99 ± 0.17	87.05 ± 0.25	79.02
CDAN	73.80 ± 0.17	86.17 ± 0.26	79.99
CDAN + BSP	74.11 ± 0.16	85.72 ± 0.32	79.92
Our method	$\mathbf{76.18 \pm 0.26}$	$\mathbf{88.01 \pm 0.18}$	**82.10**

From the experimental results on the three benchmarks, we notice that the improvement of our methon on **CUB-Birds** is larger than that on **CompCars** and **Birds-31**. There are two reasons. First, the basic recognition accuracy is relatively low on **CUB-Birds**, which leaves to the domain adaptation algorithm a larger room for improvement. Second, it can be seen from Figure 5 that the inter-domain variations of **CUB-Birds** are much larger than **CompCars** and **Birds-31**. Unlike **CompCars** and **Birds-31**, the images in two domains are all real photos. The images in CUB-200-Painting (P) dataset include watercolor, oil painting, cartoon, etc. This shows that the circular attention mechanism in our algorithm locates the details of the object itself so as to reduce the influence of image style and background on domain adaptation task.

C. ABLATION STUDY

We design ablation experiments by removing the circular attention mechanism. The results of the ablation experiments on the three benchmarks are recorded in Table 5, 6 and 7.

TABLE 5. Ablation study: Accuracy(%) on Benchmark I: CompCars.

Methods	$N \to I$	$I \to N$	Avg
Without attention	80.05 ± 0.19	71.04 ± 0.29	75.55
With attention	$\mathbf{84.40 \pm 0.02}$	$\mathbf{76.92 \pm 0.26}$	**80.66**

TABLE 6. Ablation study: Accuracy(%) on Benchmark II: CUB-Painting.

Methods	$N \to I$	$I \to N$	Avg
Without attention	61.46 ± 0.30	43.32 ± 0.37	52.39
With attention	$\mathbf{67.05 \pm 0.12}$	$\mathbf{49.57 \pm 0.23}$	**58.31**

TABLE 7. Ablation study: Accuracy(%) on Benchmark III: Birds-31.

Methods	$N \to I$	$I \to N$	Avg
Without attention	72.02 ± 0.16	83.92 ± 0.35	77.97
With attention	$\mathbf{76.18 \pm 0.26}$	$\mathbf{88.01 \pm 0.18}$	**82.10**

It can be seen from the Table 5, 6 and 7 that circular attention mechanism improves the accuracy by about 5% on the three benchmarks. This demonstrates that the circular attention mechanism works well to the positioning of the fine-grained features. With the gradual learning of labels from coarse to fine, the attention mechanism effectively reduces the inter-domain variations in the datasets, thereby achieving better domain adaptation accuracy.

V. CONCLUSION

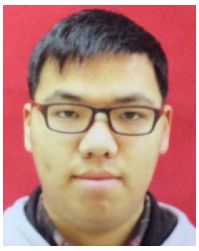
In this paper, we propose the adversarial networks with circular attention mechanism to solve the fine-grained domain adaptation problem. The key idea of our model is to locate multiple discriminative areas in the image through the circular attention mechanism and gradually align them with multiple levels in the fine-grained image label. On this basis, we design an adversarial training network to complete the domain adaptation task of fine-grained images. We compare our method with other state-of-the-art methods on three benchmarks for fine-grained domain adaptation. The experimental results show that the proposed method is effective and achieves the best performance in all three benchmarks.

REFERENCES

[1] Y. Wang, J. Choi, V. I. Morariu, and L. S. Davis, "Mining discriminative triplets of patches for fine-grained classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1163–1172.

[2] X.-S. Wei, C.-W. Xie, J. Wu, and C. Shen, "Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization," *Pattern Recognit.*, vol. 76, pp. 704–714, Apr. 2018.

[3] P. Welinder, S. Branson, C. Wah, F. Schroff, S. Belongi, and P. Perona, "Caltech-UCSD birds 200," California Inst. Technol., Pasadena, CA, USA, Tech. Rep., 2010.
[4] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, and S. Belongie, "Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 595–604.
[5] A. Angelova, S. Zhu, and Y. Lin, "Image segmentation for large-scale subcategory flower recognition," in *Proc. IEEE Workshop Appl. Comput. Vis. (WACV)*, Jan. 2013, pp. 39–45.
[6] M.-E. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 1447–1454.
[7] M. Stark, J. Krause, B. Pepik, D. Meger, J. J. Little, B. Schiele, and D. Koller, "Fine-grained categorization for 3D scene understanding," *Int. J. Robot. Res.*, vol. 30, no. 13, pp. 1543–1552, 2011.
[8] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 554–561.
[9] Y.-L. Lin, V. I. Morariu, W. Hsu, and L. S. Davis, "Jointly optimizing 3D model fitting and fine-grained classification," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 466–480.
[10] L. Yang, P. Luo, C. C. Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3973–3981.
[11] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *Proc. CVPR Workshop Fine-Grained Vis. Categorization (FGVC)*, 2011, vol. 2, no. 1, pp. 1–2.
[12] J. Liu, A. Kanazawa, D. Jacobs, and P. Belhumeur, "Dog breed classification using part localization," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2012, pp. 172–185.
[13] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
[14] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2010, pp. 213–226.
[15] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 97–105.
[16] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4068–4076.
[17] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200-2011 dataset," Tech. Rep., 2011.
[18] S. Wang, X. Chen, Y. Wang, M. Long, and J. Wang, "Progressive adversarial networks for fine-grained domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9213–9222.
[19] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The iNaturalist species classification and detection dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8769–8778.
[20] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based R-CNNs for fine-grained category detection," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 834–849.
[21] A. Vedaldi, S. Mahendran, S. Tsogkas, S. Maji, R. Girshick, J. Kannala, E. Rahtu, I. Kokkinos, M. B. Blaschko, D. Weiss, B. Taskar, K. Simonyan, N. Saphra, and S. Mohamed, "Understanding objects in detail with fine-grained attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3622–3629.
[22] T. Gebru, J. Hoffman, and L. Fei-Fei, "Fine-grained recognition in the wild: A multi-task domain adaptation approach," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1349–1358.
[23] N. Zhang, E. Shelhamer, Y. Gao, and T. Darrell, "Fine-grained pose prediction, normalization, and recognition," 2015, *arXiv:1511.07063*. [Online]. Available: http://arxiv.org/abs/1511.07063
[24] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1449–1457.
[25] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 317–326.
[26] T. Hu, H. Qi, Q. Huang, and Y. Lu, "See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification," 2019, *arXiv:1901.09891*. [Online]. Available: http://arxiv.org/abs/1901.09891
[27] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, and L. Wang, "Learning to navigate for fine-grained classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 420–435.
[28] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," 2014, *arXiv:1412.3474*. [Online]. Available: http://arxiv.org/abs/1412.3474
[29] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Unsupervised domain adaptation with residual transfer networks," 2016, *arXiv:1602.04433*. [Online]. Available: http://arxiv.org/abs/1602.04433
[30] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.
[31] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Scholkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, Jul. 2006.
[32] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," 2017, *arXiv:1705.10667*. [Online]. Available: http://arxiv.org/abs/1705.10667
[33] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, May 2015.
[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[35] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[36] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2208–2217.
[37] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7167–7176.
[38] Z. Pei, Z. Cao, M. Long, and J. Wang, "Multi-adversarial domain adaptation," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018.
[39] X. Chen, S. Wang, M. Long, and J. Wang, "Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 1081–1090.



NINGYU HE received the B.S. degree from the School of Electronic Engineering, Xidian University, China, in 2017. He is currently pursuing the Ph.D. degree with Shanghai Jiao Tong University, China. He also works with the Department of Electronic Engineering, Shanghai Jiao Tong University. His research interests include audio signal processing, image processing, and deep learning.

JIE ZHU received the Ph.D. degree in communications and information systems from Shanghai Jiao Tong University. He went to Bell Labs, Murray Hill, NJ, USA, in 1997, for cooperative scientific research. He was a Senior Visiting Scholar with the Dresden University of Technology, Germany, in 2000, for visiting research. He is currently a Professor with the Department of Electronic Engineering and a Ph.D. Supervisor in electronic science and technology. He went to the USA, Europe, Japan, South Korea, and other countries to participate in international conferences and academic exchanges for many times.