Neural Computing and Applications (2020) 32:14613–14622
https://doi.org/10.1007/s00521-020-05148-3

S.I. : DEEP LEARNING APPROACHES FOR REALTIME IMAGE SUPER RESOLUTION (DLRSR)

A self-attention-based destruction and construction learning fine-grained image classification method for retail product recognition

Wenyong Wang[1] · Yongcheng Cui[1] · Guangshun Li[2] · Chuntao Jiang[3] · Song Deng[4]

Received: 11 January 2020 / Accepted: 17 June 2020 / Published online: 10 July 2020

Abstract
Retail products belonging to the same category usually have extremely similar appearance characteristics such as colors, shapes, and sizes, which cannot be distinguished by conventional classification methods. Currently, the most effective way to solve this problem is fine-grained classification methods, which utilize machine vision + scene to perform fine feature representations on a target local region, thereby achieving fine-grained classification. Fine-grained classification methods have been widely used for recognizing birds, cars, airplanes, and many others. However, the existing fine-grained classification methods still have some drawbacks. In this paper, we propose an improved fine-grained classification method based on self-attention destruction and construction learning (SADCL) for retail product recognition. Specifically, the proposed method utilizes a self-attention mechanism in the destruction and construction of image information in an end-to-end fashion so that to calculate a precise fine-grained classification prediction and large information areas in the reasoning process. We test the proposed method on the Retail Product Checkout (RPC) dataset. Experimental results demonstrate that the proposed method achieved an accuracy above 80% in retail commodity recognition reasoning, which is much higher than the results of other fine-grained classification methods.

Keywords Fine-grained classification · Multi-attribute recognition · Self-attention learning

1 Introduction

Currently, hundreds of millions of retail products traffic every day, both online and offline. Commodity retailing is a labor-intensive industry with a very high cost in manual checkouts. With the development of artificial intelligence and deep neural network, it is an irresistible trend to achieve automatic checkouts of retail products using various image classification methods.

Over the past decade, object recognition methods have made steady progress in large-scale data annotation and complex model design. However, there are still a series of problems in recognizing the categories of fine-grained objects (such as birds, butterflies, and vehicles) [1–18]. Therefore, effective fine-grained image classification/recognition methods need to be developed to accurately classify retail products.

Various image classification methods have developed rapidly in recent years, and the classification results also changed from two-category classification to multi-category classification, from coarse-grained classification to fine-grained classification. Relying on the development of deep learning techniques, deep neural network-based image classification techniques make the image classification easier, and the requirements for image classification results are also constantly increasing. Fine-grained classification

✉ Guangshun Li
guangshunli@qfnu.edu.cn

[1] College of Information Science and Technology, Northeast Normal University, Changchun, China

[2] School of Information Science and Engineering, Qufu Normal University, Rizhao, China

[3] School of Mathematics and Big Data, Foshan University, Foshan, China

[4] Institute of Advanced Technology, Nanjing University of Post and Telecommunication, Nanjing, China

methods are widely used in recognizing the variety of flowers, the types of birds, the classification of vehicles, face recognition, and so on.

The research on fine-grained image analysis can usually be divided into three categories: (i) fine-grained computer vision, (ii) fine-grained image retrieval (by graph search), and (iii) fine-grained image generation. Figure 1 shows the results of conventional image classification, i.e., coarse-grained image classification, in which the differences between categories are obvious. Figure 2 shows the results of fine-grained image classification, which can be regarded as a further classification process of the coarse-grained classification results. With fine-grained image classification methods, we can not only classify the taste of beverages or chips, but also distinguish sub-categories of the same milk powder brand. Therefore, only fine-grained classification methods can divide retail products into their true categories.

Automatic checkout (ACO) is a core scene of fine-grained classification to recognize retail products, which is integrated with computer vision techniques to produce settlement lists created by cashier scene images. For this reason, in early 2019, MEGVII's Nanjing Research Institute created the largest product identification dataset at present, the Retail Product Checkout (RPC) dataset [19], to promote research and technological advancement in the new retail automatic cashier scene. The dataset contains up to 200 product categories, a total of 830,000 images, the real simulation of retail scene, and the fidelity exceeds the existing similar datasets, in the same time fully demonstrating the fine-grained nature of the ACO problem. The release of the RPC dataset provides a powerful dataset support for the fine-grained image classification on retail commodities, laying a data foundation for further refinement of retail product recognition. The RPC dataset not only affects the process of ACO, but also has broad application scenarios in the application of e-commerce and retail products, object information collection and target recognition of household service robots.

In this paper, we propose an improved fine-grained image classification method, which utilizes a self-attention mechanism in the destruction and construction of image information in an end-to-end fashion. The proposed method calculates a precise fine-grained classification prediction and large information areas in the reasoning process. We implement the proposed method on the RPC dataset. The experimental results demonstrate that the proposed method has achieved a higher classification accuracy than other conventional fine-grained classification methods.

2 Related work

Fine-grained image classification aims to divide coarse-grained classes into detailed sub-classes, such as differentiating the type of birds, which is a challenging task in computer vision. Compared to the results of coarse-grained classification, the accuracy of fine-grained classification is more meticulous since the differences between categories are more subtle. In other words, different classes only can be distinguished based on small local differences in images. Thus, compared to coarse-grained classification tasks, such as face recognition and action recognition, there are particularly a lot of uncertain factors (e.g., angle, light, occlusion, noise), making fine-grained classification more challenging. Conventional coarse-grained classification methods usually depend on a large amount of manually labeled information to achieve fine-grained image classification.

An idea for solving the above problem is to utilize both local information and global information in fine-grained

Fig. 1 The results of coarse-grained image classification

Fig. 2 The results of fine-grained image classification

images. The existing fine-grained image classification methods can be divided into three categories according to the local information used in a fine-grained image:

(i) Fine-grained image classification with strong supervision, which relies on manual annotation information (Annotations), e.g., bounding box [17].
(ii) Fine-grained image classification with weakly supervision, which relies on category labels (Labels) [20].
(iii) Multi-attribute fine-grained image classification, using both multi-attribute text information and category labels to achieve, the complexity of which is between the above two [21–23].

Considering the cost of the datasets, the fine-grained image classification with weakly supervision is the development trend [24, 25]. When the identification of multiple attributes of retail products is involved, the multi-attribute fine-grained image classification can perform more accurate and efficient than the other two ways.

Lin et al. [26] proposed a bilinear CNN model (B-CNN) for weak supervised fine-grained image recognition. Specifically, B-CNN consisted of two feature extractors, and the extracted features were fused for image description. This bilinear form simplified gradient calculations, simulated local pairwise feature interactions in a translation-invariant manner, and allowed an end-to-end training only using image labels. However, this method is incapable of recognizing the subtle traits/features which can fully characterize the objects.

To solve this problem, Yang et al. [2] developed a self-supervision mechanism named navigator-teacher-scrutinizer network (NTS-Net) for fine-grained image recognition, which focus on the informative regions in images without using fine-grained bounding-box/part annotations. Three agents (i.e., a navigator agent, a teacher agent and a scrutinizer agent) were contained in NTS-Net, in which the navigator agent was used to capture/detect the informative regions with the guidance of the teacher agent, and then the scrutinizer agent was used to scrutinize the detected regions and make description.

Similarly, a trilinear attention sampling network (TASN) [3] was developed for fine-grained image recognition, which learned subtle but discriminative features from images in a teacher–student manner. The TASN method consisted of three modules: a trilinear attention module, an attention-based sampler and a feature distiller. Specifically, the trilinear attention module modeled the inter-channel relationships to generate attention maps, and then the attention-based sampler was developed to highlight the attended parts with high resolution. In this way, the feature distiller can use weight sharing and feature preserving strategies to get object-level features from images. This method achieved significant performance on rigid and non-rigid datasets (e.g., Caltech-UCSD Birds (CUB-200-2011) dataset [8, 27, 28] and Stanford Cars dataset [1]). However, for the objects between rigid and non-rigid such as retail products, the performance of TASN was not well.

To solve this problem, Chen et al. [4] proposed a destruction and construction learning method (DCL) for

fine-grained image recognition. Specifically, a destruction-construction stream was developed to destruct and reconstruct images. The destruction enhanced recognition robustness, while the construction simulated the semantic correlation between image regions. Thus, DCL can learn discriminative regions and features from images. Compared to other fine-grained image recognition methods, DCL was lightweight, easy to train, fast in reasoning, and practical.

Table 1 shows the accuracy of the above fine-grained classification methods on CUB-200-2011 dataset and Stanford Cars dataset. Recently, the CUB-200-2011 and Stanford Cars datasets have the most extensive and authoritative datasets in the field of fine-grained image classification. The experimental results on these two datasets are of reference value.

The process of retail product recognition is quite similar to the fine-grained classification process of birds and vehicles. We tested the above fine-grained image classification methods (i.e., B-CNN, NTS-NET, TASN and DCL) on the RPC retail products dataset. Table 2 shows the experimental results.

As shown in Table 2, B-CNN, NTS-NET, and TASN have achieved unsatisfactory accuracies in recognizing retail products, which were much lower than the recognition accuracy of DCL on the RPC dataset. When using DCL with ResNet50 as the backbone, the recognition accuracy on the RPC dataset was 71.8%. This indicates that DCL has good generalization and enhance ability. In theory, the characteristics of the RPC dataset are between rigid and non-rigid datasets, so that the feature extraction of retail commodity details is fundamental to improve recognition accuracy. Therefore, this paper aims to develop an improved DCL fine-grained image classification method for retail product recognition. Specifically, a self-attention mechanism is introduced to DCL to destruct and construct image information in an end-to-end fashion.

Table 2 Experimental results on the RPC dataset

Method	Base Model	Acc on RPC
B-CNN [26]	VGGnet	53.2%
NTS-NET [2] (k = 4)	ResNet-50	54.7%
TASN [3]	VGG-19	61.4%
	ResNet-50	63.9%
DCL [4]	VGG-16	69.5%
	ResNet-50	71.8%

3 The proposed method

In this section, we will present our proposed retail product recognition method in detail.

3.1 Destruction learning

In the process of fine-grained image classification, it has been proved that the local information of an image plays a more important role than the global information. It is because, in most cases, different fine-grained categories can have the same global structure and differ only in specific local details. The destruction learning is to disrupt the global structure by disrupting the local area, which makes a model easier to find the discriminating area and learning discriminant features. The crucial part of this is named as the Region Confusion Mechanism (RCM) [4].

In addition, in order to prevent the noise emerged by the destruction from negatively affecting network learning, the DCL method introduces the loss of adversarial learning. Experimental results indicate that this prevents the noise pattern caused by RCM from entering the feature space. Through analysis, we found that there is still room for DCL improvement. We removed the traditional adversarial learning and replaced it with a self-attention learning mechanism, which can better eliminate the noise generated by RCM.

Table 1 Experimental results on the CUB-200-2011 dataset and the Stanford Cars dataset

Method	Base Model	Acc on CUB-200-2011	Acc on Stanford Car
B-CNN [26]	VGGnet	84.1%	91.3%
NTS-NET [2] (k = 2)	ResNet-50	87.3%	93.7%
NTS-NET [2] (k = 4)	ResNet-50	87.5%	93.9%
TASN [3]	VGG-19	87.1%	93.2%
	ResNet-50	87.9%	93.8%
DCL [4]	VGG-16	86.9%	94.1%
	ResNet-50	87.8%	94.5%

3.1.1 Region confusion mechanism

RCM is designed to disrupt the spatial distribution of local areas. For any image R, divide it into $N \times N$ sub-regions, each labeled $R_{i,j}$, i and j represent column coordinates and row coordinates. Specifically, a random vector q_j of size N is generated for the jth line of R, where the value of the ith element is $q_{j,i} = i+r$, $r \sim U(-k, k)$, and uniformly distributed random variables within the range of $[-k, k]$. k is an adjustable parameter ($1 < k < N$), and $q_{j,i} = i+r$ defines the neighborhood range. Then a new permutation σ_j^{row} of the region in the jth row can be obtained by sorting q_j, the verification condition is:

$$\forall i \in \{1, \ldots, N\}, \left|\sigma_j^{\text{row}}(i) - i\right| < 2k \tag{1}$$

By using the same way, a new arrangement σ_i^{col} of the region in the ith row can be obtained. Therefore, the coordinates of a new position, where the area at the position (i, j) in the original image, is placed by:

$$\sigma(i,j) = \left(\sigma_j^{\text{row}}(i), \sigma_i^{\text{col}}(j)\right) \tag{2}$$

This confusion method destructs the global structure and ensures that the local area moves within an adjustable range.

Assumed an initial image I, a destructed image $\phi(I)$,and its corresponding one-to-many label l (fine-grained category). These three parts are combined into $(I, \phi(I), l)$ for training the model. Then a classification network maps the input image to a probability distribution vector $C(I, \phi_{cls})$, where ϕ_{cls} represents all learnable parameters in the classification network. Hence, the loss function of the classification networks is:

$$\mathcal{L}_{cls} = -\sum_{I \in \mathcal{I}} l \cdot \log[C(I)C(\phi(I))], \tag{3}$$

where $\mathcal{L}$ is a collection of image training set. Since the global structure is destroyed, to identify these randomly scrambled images, the classification network must look for discriminative areas and then learn the subtle differences between classes.

3.1.2 Improved loss adversarial learning function with self-attention mechanism

Experiments have shown that images destroyed using RCM are not always able to bring useful information for fine-grained classification. When we confuse local areas, RCM generates a unique noise visual pattern. In this pattern, the feature of learning has a negative impact on the classification task. In order to solve this problem, DCL proposed an adversarial loss by the idea from Generative Adversarial Nets (GAN) [29] creatively to prevent the noise pattern caused by overfitting RCM from entering the feature space.

Although the adversarial learning effect is obvious, there is still room for the improvement in DCL. It is because the traditional adversarial model is easy to learn the texture features of a product (such as product packaging and product shape), but it is difficult to learn the specific structure and geometric features of the product (or the internal detail feature of a product, such as trademark and logo).

Since the adversarial network can enhance discriminative local details [4, 30], we intend to replace the traditional adversarial learning model with a Self-Attention GAN (SAGAN) [31]. SAGAN is able to get the global geometric features of a noise image in one step by directly computing/measuring the relationships between any two pixels in the noise image. Its role is to learn the dependencies between global features, and can use spectral normalization [5, 32–34] to improve stability during training.

Figure 3 shows the details of the self-attention module: Firstly, the previously hidden layer feature x is inputed into two feature maps (i.e., f and g) to compute the self-attention parameter. The outputs of these two feature maps will be calculated by a matrix multiplication $\otimes$, and we can get

$$\beta_{j,i} = \frac{\exp(s_{ij})}{\sum_{i=1}^{N} \exp(S_{ij})}, \text{ where } s_{ij} = f(x_i)^T g(x_j) \tag{4}$$

where $\beta_{j,i}$ denotes whether the model concerns the ith position when distinguishing the jth region. Besides,all of $f(x)$, $g(x)$ and $h(x)$ are the fundamental 1×1 convolution kernel,and the only difference between them is the size of the channel output.

The output of $f(x)$ is transposed and multiplied by the output of $g(x)$, and then require a self-attention map by the result normalized with *softmax* function processing.

Multiply the self-attention map with $h(x)$ by pixel to obtain the feature value o_j, $o = (o_1, o_2, \ldots, o_j, \ldots, o_N) \in \mathbb{R}^{C \times N}$ of the adaptive attention layer,

$$o_j = \sum_{i=1}^{N} \beta_{j,i} h(x_i) = W_h x_i, \text{ where } \mathrm{h}_i = W_h x_i. \tag{5}$$

In this experiment, we only need the discriminator model in SAGAN to discriminate the destructed image and the original image and then calculate the loss until the discriminator cannot discriminate the corrupted image and the original image. This can effectively prevent RCM noise from entering the feature space. $\mathcal{L}_{\text{sag}}$ is a loss function that satisfies Eq. 5:

$$\begin{aligned} \mathcal{L}_{\text{sag}} = & -\mathbb{E}_{(x,y)\sim p_{\text{data}}}[\min(0, -1 + D(x,y))] \\ & - \mathbb{E}_{z\sim p_z, y\sim p_{\text{data}}}[\min(0, -1 - D(G(z), y))] \end{aligned}. \tag{6}$$

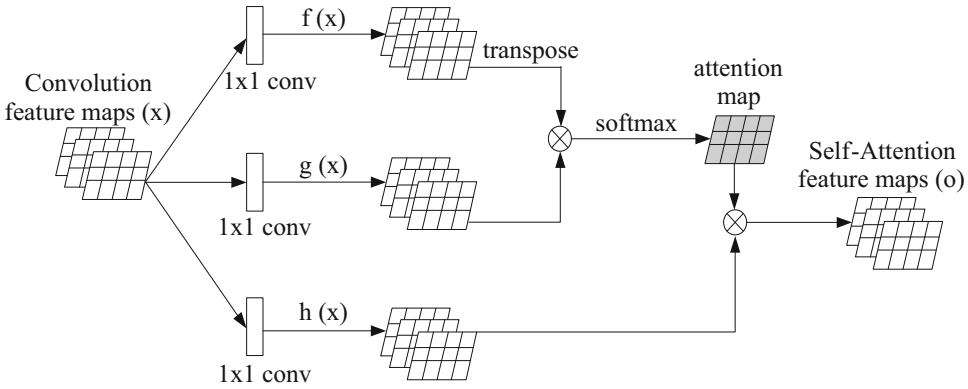

Fig. 3 Self-attention feature maps

3.2 Spectral normalization

In addition to generating self-attention adversarial learning network, we also add a spectral normalization [5, 31–34] to the network to improve training stability and effectively avoid model overfitting. Spectral normalization is a technique used to stabilize the training of discriminators network. It can enhance the generalization ability of the model. It is very simple to implement this strategy by adding Lipschitz constraints to the parameter matrix of the traditional discriminators.

The Lipschitz constraint is

$$\frac{||f(x)-f(x')||_2}{||x-x'||_2} \leq D, \tag{7}$$

where D is a constant. The Lipschitz constraint satisfies that: The function change is not too quick, and the gradient is restricted. The function value is limited to the range of D or less even at the most extreme value.

When the discriminator adopts this distance to train, it can alleviate the convergence problem in the traditional adversarial learning network training in DCL. At the same time working with the Two Time-Scale Update Rule (TTUR) mechanism [7], discriminator can prevent the RCM noise from entering the feature map more effectively.

3.3 Construction learning

Considering that the retail product image recognition has a similar local correlation with bird image recognition, we choose to use the same region construction loss and region align network [4], certainly, to measure the positional accuracy of different regions in the image, and guide the base network to model the semantic correlation between regions through end-to-end training. The predicted area $R_\sigma(i, j)$ is $M_{\sigma(i,\ j)}(\phi(I))$ in I, and the predicted area $R(i, j)$ is $M_{i,\ j}(I,\ i,\ j)$ in I. The true values of these two predictions $M_{\sigma(i,\ j)}(\phi(I))$ and $M_{i,\ j}(I,\ i,\ j)$ are both (i, j). Eventually, the calculated area alignment loss is $\mathcal{L}_{\text{loc}}$, defined as the predicted coordinates and the distance from the original coordinates, the expression is:

$$\mathcal{L}_{\text{loc}} = \sum_{I\in\mathcal{I}}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|M_{\sigma(i,j)}(\phi(I)) - \begin{bmatrix} i \\ j \end{bmatrix}\right|_1 + \left|M_{i,j}(I) - \begin{bmatrix} i \\ j \end{bmatrix}\right|_1. \tag{8}$$

Experiments show that the loss of region construction is helpful to locate the key information in the image, and it is easy to find the correlation between the sub-regions. Moreover, the region alignment loss can allow the backbone classification network to have deeper training and facilitate the modeling of structural information (the shape of the object and the semantic correlation between the various parts of the object).

3.4 Self-attention destruction and construction learning

In the improved DCL framework, classification loss, self-attention adversarial loss and region alignment loss are trained in an end-to-end manner, these three networks can use fine local detail and well-modeled object partial correlation for fine-grained identification. Figure 4 illustrates the structure of the improved SADCL classification model.

The destruction learning is used to learn from discriminative areas, while construction learning is used to rearrange the learned local details according to the semantic relevance between regions. Hence, DCL is able to generate a set of diverse and complex visual representations based on the well-structured detail features extracted from discriminative regions. The loss function $\mathcal{L}$ of the improved SADCL model consists of three parts with weights, the classification network loss $\mathcal{L}_{\text{cls}}$, the self-attention adversarial loss $\mathcal{L}_{\text{sag}}$ and the region alignment loss $\mathcal{L}_{\text{loc}}$

$$\mathcal{L} = p\mathcal{L}_{cls} + q\mathcal{L}_{\text{sag}} + r\mathcal{L}_{\text{loc}}. \tag{9}$$

In this paper, p and q are equal, denoted as k, so the loss function is:

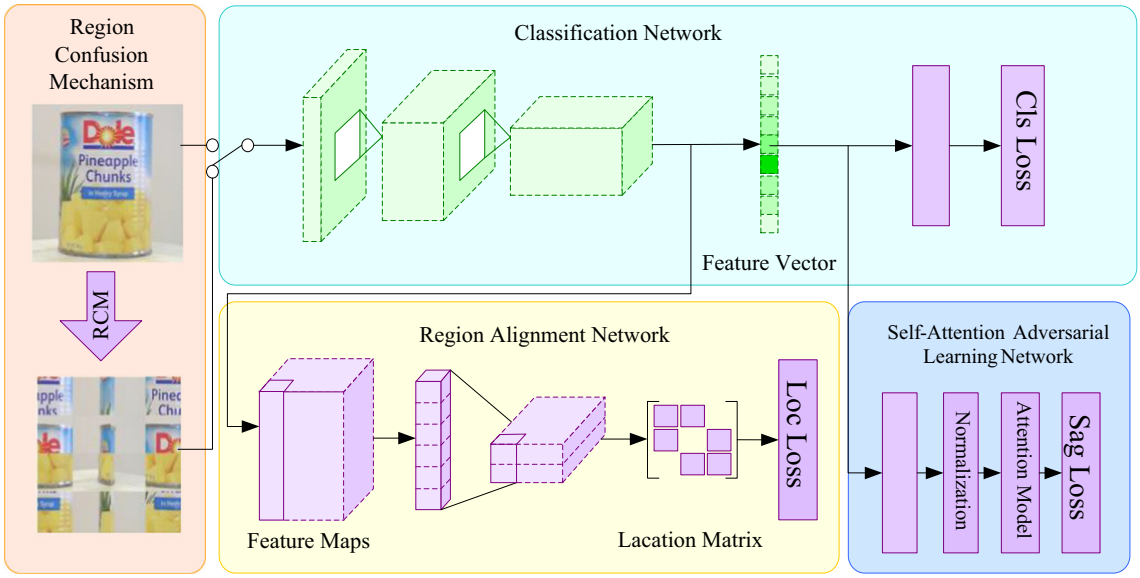

Fig. 4 Improved SADCL classification model structure

$$\mathcal{L} = k\mathcal{L}_D + r\mathcal{L}_C \tag{10}$$

4 Experiments

In this section, we present the datasets in Sect. 4.1 and experimental settings in Sect. 4.2 for our experiments, and then discuss the experimental results in Sects. 4.3 and 4.4.

4.1 Datasets

We comprehensively evaluate our proposed SADCL method on Caltech-UCSD Birds (CUB-200-2011) [1] and MEGVII RPC (Retail Product Checkout, RPC) [19] datasets. These two datasets are extensively used for fine-grained image classification competitions and rankings. Besides, the part annotations and the bounding box are not used during our experiments.

To validate the generalization of SADCL, our experiments have an independent dataset test and sub-sampling (or named k-fold cross-validation). In the k-fold cross-validation, single product images are divided into k non-overlapping and equal-sized subsets by random. Of the k subsets, a subset is used for test and the remaining k-1 subsets are retained as the training set.

There are 53,739 images of retail products in RPC. Figure 5 shows a sample with different photo angles in the RPC dataset. We divide RPC into 10 parts randomly. By setting the value of k as 10, a ten fold cross-validation is performed. After validation, the results of the 10 divided evaluations are obtained, the final result is to average the results of each evaluation.

In our experiments, a single sample image of the RPC dataset is preprocessed to match the size of 448 × 448. This meets the input requirements of the model and ignores the large white background in the image without losing feature information. In addition, we select a part of the training set for subsequent training and cross-validation. Cross-validation greatly reduces the contingency caused by a random division by multiple divisions. At the same time, through multiple divisions and multiple pieces of training, the model can also encounter various commodity maps, thus improving its generalization ability.

4.2 Implementation details

The implementation details of the experiments are as follows:

Lab environment: Python3.6, PyTorch 1.2.0, torchvision 0.3.0, OpenCV-Python 4.1.0.

Backbone: We use Resnet-50 and VGG-16 as basic classification network (backbone) models pre-trained by ImageNet at the same time.

Data enhancement: To facilitate the experiment without losing information, we preprocess the images in the two datasets. The images in the CUB-200-2011 dataset are scaled to 512 × 512 sizes, and then randomly cropped to 448 × 448 sizes [35]. The images in the RPC dataset are scaled to 1024 × 1024 size and then cropped to 448 × 448 sizes. In addition, some data enhancement methods, such as randomly rotation, random horizontal flip, and random brightness will be used on these preprocessed images to increase the diversity of the images in the training set.

Moreover, to identify high-resolution images, we changed the first and second fully connected layers in VGG-16

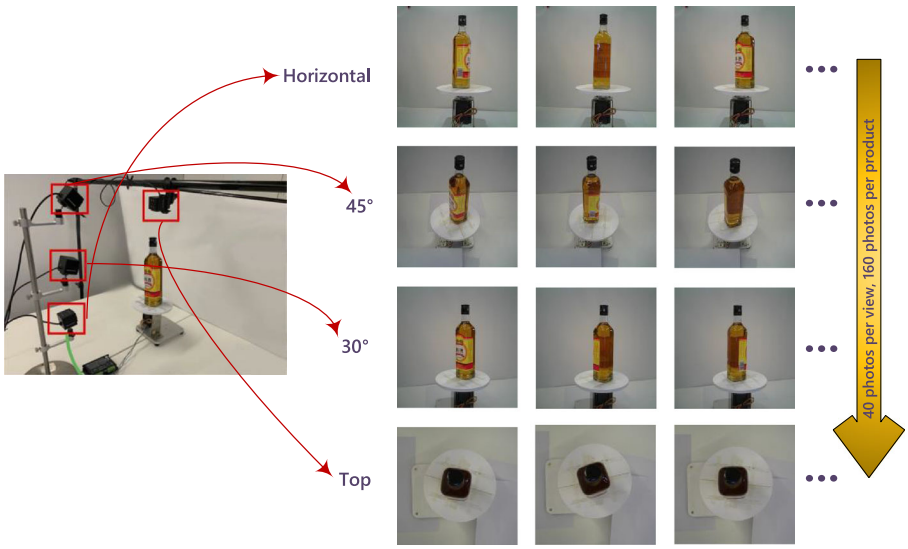

Fig. 5 A sample in the RPC dataset

for full convolution [20, 36, 37]. At the same time, the feature map output of the last convolutional layer of the base network will be sent to a region alignment network as an input. The eigenvector obtained by average pooling from the last convolutional layer is sent to the self-attention adversarial network.

Key parameter setting: All methods use 200 epochs during training; the learning rate decay by a factor of 10 for every 60 epochs; 32 batches are set for the training and test sets. Since the RPC dataset is huge, and each of the images in the dataset has a high resolution, we cannot use Full Batch Learning. In general, within a certain range acceptable to the hardware, the larger the batch_size, the more accurate the direction of its determined decline, and causing the smaller training shock. After several tests, it shows that setting batch_size to 32 can achieve the highest accuracy under the premise of iterating all epochs.

The values of train_num_workers and val_num_workers in PyTorch are both set to 8. Setting save_points every 30 epochs. The size of an image must be divisible by the region number N in RCM; the height and width of the region are set to 32; N is set to 7 as initialization. In addition, we set $k = p{=}q = 1$. As verified in [4], the correlations between different regions are very important for establishing learning of objects.

The setting of r is related to the appearance characteristics of the recognized object. Specifically, changes in the appearance characteristics can lead to reduced effectiveness of deconstruction learning. In this case, increasing the weight of reconstruction learning r is a good solution. Further, the retail product recognition task is similar to the rigid object recognition task. The appearance characteristics of retail products such as packaging cannot change easily or frequently. Thus, the appearance characteristics of retail products can play an important role in retail product recognition.

Based on the above analysis, in our experiments, we set $r = 1$, 0.5 and 0.1 to the proposed method and conducted multiple experiments on the RPC dataset. Experimental results show that when $r = 0.1$, the proposed method has the highest recognition accuracy for retail products. On the contrary, since the bird information in the CUB-200-2011 dataset is not fixed, for the experiments on the CUB-200-2011 dataset, we set r to 1 to the proposed method.

4.3 Performance comparison

As indicated in Table 3, the proposed SADCL method is more suitable for retail product recognition than other fine-grained image classification methods. When combining with ResNet-50 for feature extraction, the accuracy of SADCL on the RPC dataset is as high as 81.4%. Although the accuracy of SADCL on the CUB-200-2011 dataset is lower than that of DCL due to some factors such as the experimental environment, SADCL has achieved a considerable accuracy on the RPC dataset.

4.4 Ablation experiments

To test the exact effect of the proposed self-attention mechanism, we performed an ablation experiment at the same time, and compared the recognition accuracy on the CUB-200-2011 dataset and the RPC dataset in each module by controlling variables. The ablation experimental results are shown in Table 4.

As shown in Table 4, the accuracy of the destruction learning-based classification network with self-attention mechanism (i.e., SAD in Table 4) is much higher than that

Table 3 Experimental results data on CUB-200-2011 and RPC datasets

Method	Base Model	Acc on CUB-200-2011	Acc on RPC
B-CNN [26]	VGGnet	84.1%	47.2%
NTS-NET [2] (k = 2)	ResNet-50	87.3%	52.5%
NTS-NET [2] (k = 4)	ResNet-50	87.5%	54.7%
TASN [3]	VGG-19	87.1%	61.4%
	ResNet-50	87.9%	63.9%
DCL [4]	VGG-16	86.9%	69.5%
	ResNet-50	87.8%	71.8%
Ours SADCL	VGG-16	85.2%	78.7%
	ResNet-50	85.9%	81.4%

Table 4 The ablation experimental results

Method	Acc on CUB-200-2011	Acc on RPC
ResNet-50	83.7%	63.2%
+RCM	84.2%	67.5%
Destruction (D)	84.7%	68.9%
Construction (C)	84.4%	67.4%
Destruction & Construction (DC)	86.5%	71.8%
Ours SAD	85.3%	78.7%
Ours SADC	85.9%	81.4%

of the network only using general adversarial learning (i.e., Destruction in Table 4). The ablation experiment also demonstrates that the self-attention-based fine-grained image classification is capable of dealing with retail product recognition tasks.

5 Conclusion

In this paper, we propose an improved fine-grained classification method named SADCL for retail product recognition. A self-attention mechanism is utilized by SADCL, which can effectively prevent noise overfitting after image destruction. SADCL highlights the characteristics of retail merchandise and enables end-to-end training. Compared to the existing methods, the proposed method can better learn the dependencies between global features, making SADCL is more suitable for fine-grained classification on retail product recognition. Experimental results demonstrate that the proposed method is more robust than the reference methods.

Experimental results show that the accuracy of the proposed method on the RPC data set exceeds 80%, much higher than other fine-grained classification methods. Further, SADCL has a low computational cost. It only has the basic network computational cost for reasoning and does not require other external computational costs.

To summarize, SADCL is an effective fine-grained classification method, which has further research value in both practical applications and scientific research and can be used in a wider field than retail product recognition. For future work, we intend to use SADCL to detect and recognize other rigid objects such as indoor furniture. We will also continue the research on fine-grained classification for retail product recognition/classification, and test better fine-grained classification methods on the RPC dataset.

Acknowledgements This work is supported by the National Natural Science Foundation of China (61672321, 61771289, 61832012, 61802062, 51977113, and 51507084), Major Basic Research of Natural Science Foundation of Shandong Province (ZR2019ZD10), Shandong Province Key Research and Development Plan (2019GGX101050), and the Project of Department of Education of Guangdong Province (2017KQNCX209).

References

1. Krause J, Stark M, Deng J, Fei-Fei L (2013) 3d object representations for fine-grained categorization. In: Proceedings of the IEEE international conference on computer vision workshops, pp 554–561
2. Yang Z, Luo T, Wang D, Hu Z, Gao J, Wang L (2018) Learning to navigate for fine-grained classification. In: ECCV 2018, arXiv:1809.00287
3. Zheng H, Fu J , Zha Z, Luo J (2019) Looking for the devil in the details: learning trilinear attention sampling network for fine-grained image recognition. arXiv:1903.06150
4. Chen Y, Bai Y, Zhang W, Mei T (2019) Destruction and construction learning for fine-grained image recognition. In: The IEEE conference on computer vision and pattern recognition (CVPR), pp 5157–5166

5. Miyato T (2018) Spectral normalization for generative adversarial networks, arXiv:1802.05957v1
6. Zhu JY, Parck T, Isola P, Efros A (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: CVPR
7. Heusel M, Ramsauer H, Unterthiner T, Nesslerm B (2018) GANs trained by a two time-scale update rule converge to a local nash equilibrium. arXiv: 1706.08500v6
8. Catherine w, Steve B, Peter W, Pietro P, Serge B (2011) The caltech-ucsd birds-200-2011 dataset. (CNS-TR-2011-001)
9. Zheng H, Wang R, Ji W, Zong M, Wong WK, Lai Z, Lv H (2020) Discriminative deep multi-task learning for facial expression recognition. Inf Sci. https://doi.org/10.1016/j.ins.2020.04.041
10. Zou F, Xiao W, Ji W, He K, Yang Z, Song J, Zhou H, Li K (2020) Arbitrary-oriented object detection via dense feature fusion and attention model for remote sensing super-resolution image. Neural Comput Appl. https://doi.org/10.1007/s00521-020-04893-9
11. M.Mirza, S.Osindero. Conditional Generative Adversarial Nets. arXiv:1411.1784v1, 2014
12. Maji S, Rahtu E, Kannala J, Blaschko MB, Vedaldi A (2013) Fine-grained visual classification of aircraft. CoRR, abs/1306.5151
13. Liu X, Xia T, Wang J, Lin Y (2016) Fully convolutional attention localization networks: efficient attention localization for fine-grained recognition. CoRR, abs/1603.06765
14. Zheng H, Fu J, Mei T, Luo J (2017) Learning multi-attention convolutional neural network for fine-grained image recognition. 10.1109/ICCV.2017.557
15. Cui Y, Song Y, Sun C, Howard A, Belongie S (2018) Large scale fine-grained categorization and domain-specific transfer learning. In: CVPR, pp 4109– 4118, 2018. 2
16. Huang C, He Z, Cao G, Cao W (2016) Task-driven progressive part localization for fine-grained object recognition. IEEE Trans Multimed 18(12):2372–2383
17. Fu J, Zheng H, Mei T (2017) Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In: CVPR, pp 4438–4446
18. Rodŕıguez P, Gonfaus JM, Cucurull G, XavierRoca F, Gonzalez J (2018) Attend and rectify: a gated attention mechanism for fine-grained recovery. In: Proceedings of the European conference on computer vision (ECCV), pp 349–364
19. Wei XS, Cui Q, Yang L, Wang P, Liu L (2019) RPC: a large-scale retail product checkout dataset. arXiv:1901.07249
20. Mehdi N, Paolo F (2016) Unsupervised learning of visual representations by solving jigsaw puzzles. In: Computer vision–ECCV 2016, pp 69–84, Cham, 2016. Springer International Publishing
21. Ming S, Yuchen Y, Feng Z, Errui D (2018) Multi-attention multi-class constraint for fine-grained image recognition. pp 834–850
22. Peng Y, He X, Zhao J (2018) Object-part attention model for fine-grained image classification. IEEE Trans Image Process 27(3):1487–1500
23. Cai S, Zuo W, Zhang L (2017) Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization. In: 2017 IEEE international conference on computer vision, pp 511–520
24. Doersch C, Gupta A, Efros AA (2015) Unsupervised visual representation learning by context prediction. In: 2015 IEEE international conference on computer vision, pp 1422–1430
25. Lample G, Conneau A, Denoyer L, Ranzato M (2018) Unsupervised machine translation using monolingual corpora only
26. Lin T, RoyChowdhury A, Maji S (2015) Bilinear cnn models for fine-grained visual recognition. In: 2015 IEEE international conference on computer vision, pp 1449–1457
27. Berg T, Liu J, Lee SW, Alexander ML, Jacobs DW, Belhumeur PN (2014) Birdsnap: large-scale fine-grained visual categorization of birds. In 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp 2019–2026
28. Branson S, Horn GV, Belongie SJ, Perona P (2014) Bird species categorization using pose normalized deep convolutional nets. In: BMVC, 2014. 1
29. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair D, Courville AC, Bengio Y (2014) Generative adversarial nets. In: NIPS
30. Donahue J, Kr̈ahenb̈uhl P, Darrell T (2017) Adversarial feature learning. In: ICLR
31. Lin H, Goodfellow I, Metaxas D, Odena A (2018) Self-attention generative adversarial networks. arXiv:1805.08318
32. Yoshida Y (2017) Spectral norm regularization for improving the generalizability of deep learning, National Institute of Informatics. arXiv: 1705.10941v1, 2017
33. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv:1409.0473,
34. Che T, Li Y, Jacob AP, Bengio Y, Li W (2017) Mode regularized generative adversarial networks. In: ICLR
35. Dziugaite GK, Ghahramani Z, Roy DM (2016) A study of the effect of jpg compression on adversarial images. arXiv preprint arXiv:1608.00853
36. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2018) Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution and fully connected crfs. TPAMI 40(4):834–848
37. Cheng J, Dong L, Lapata M (2016) Long short-term memory-networks for machine reading. In: EMNLP