# Destruction and Construction Learning for Fine-grained Image Recognition

Yue Chen[1]* Yalong Bai[2]* Wei Zhang[3] Tao Mei[4]
JD AI Research, Beijing, China
[1]chenyue21@jd.com, [2]ylbai@outlook.com, [3]wzhang.cu@gmail.com, [4]tmei@live.com

## Abstract

*Delicate feature representation about object parts plays a critical role in fine-grained recognition. For example, experts can even distinguish fine-grained objects relying only on object parts according to professional knowledge. In this paper, we propose a novel "Destruction and Construction Learning" (DCL) method to enhance the difficulty of fine-grained recognition and exercise the classification model to acquire expert knowledge. Besides the standard classification backbone network, another "destruction and construction" stream is introduced to carefully "destruct" and then "reconstruct" the input image, for learning discriminative regions and features. More specifically, for "destruction", we first partition the input image into local regions and then shuffle them by a Region Confusion Mechanism (RCM). To correctly recognize these destructed images, the classification network has to pay more attention to discriminative regions for spotting the differences. To compensate the noises introduced by RCM, an adversarial loss, which distinguishes original images from destructed ones, is applied to reject noisy patterns introduced by RCM. For "construction", a region alignment network, which tries to restore the original spatial layout of local regions, is followed to model the semantic correlation among local regions. By jointly training with parameter sharing, our proposed DCL injects more discriminative local details to the classification network. Experimental results show that our proposed framework achieves state-of-the-art performance on three standard benchmarks. Moreover, our proposed method does not need any external knowledge during training, and there is no computation overhead at inference time except the standard classification network feed-forwarding. Source code:* `https://github.com/JDAI-CV/DCL`.

## 1. Introduction

In the past decade, generic object recognition has achieved steady progress with efforts from both large-scale annotated dataset and sophisticated model design. However, recognizing fine-grained object categories (e.g., bird species [3], car models [14] and aircraft [18]) is still a challenging task, which attracts extensive research attention. Although fine-grained objects are visually similar by a rough glimpse, they can be correctly recognized by details in discriminative local regions.

Learning discriminative feature representations from discriminative parts plays the key role in fine-grained image recognition. Existing fine-grained recognition methods can be roughly grouped into two categories, as illustrated in Figure 1. One group (a) first locates the discriminative object parts and then classifies based on the discriminative regions. These two-steps methods [21, 11, 1] mostly need additional bounding box annotations on objects or parts, which are expensive to collect. The other group (b) tries to automatically localize discriminative regions by attention mechanism in an unsupervised manner, and thus does not needs extra annotations. However, these methods [7, 42, 41, 22] usually need additional network structure (e.g., attention mechanism), and thus introduce extra computation overhead for both training and inference stages.

In this paper, we propose a novel fine-grained image recognition framework named "Destruction and Construction Learning" (DCL), as shown in Figure 1 (c). Besides the standard classification backbone network, we introduce a DCL stream to learn from discriminative regions automatically. An input image is first carefully *destructed* to emphasize discriminative local details, and then *reconstructed* to model the semantic correlation among local regions. On one hand, DCL automatically localizes discriminative regions, and thus does not need any extra knowledge while training. On the other hand, the DCL structure is only adopted at the training stage, and thus introduces no computational overhead at inference time.

For "Destruction", we propose a Region Confusion Mechanism (RCM) to deliberately "confuse" the global structure, which partitions the input image into local patches and then shuffles them randomly (Figure 3). For fine-grained recognition, local details play a more important role than global structures, since images from different fine-

*Equal contribution.

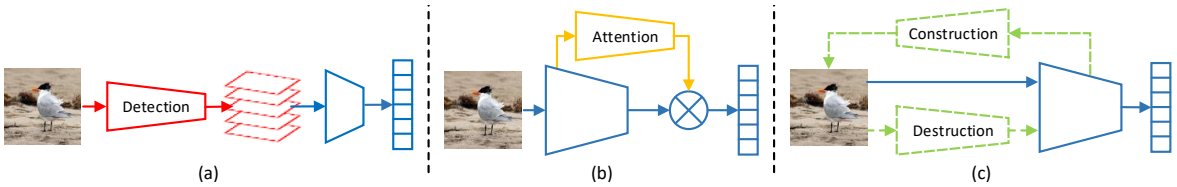


Figure 1. Illustrations of two previous general frameworks (a,b) and our proposed framework (c) for fine-grained classification. (a) Two-stage part detection based framework. (b) Attention based framework. (c) Our proposed destruction and construction learning framework. The network structures in dashed lines are disabled during inference.

grained categories usually share the same global structure or shape, but only differ in local details. Discarding global structure and keeping local details could force the network to identify and focus on the discriminative local regions for recognition. After all, the devil is in the details. Shuffling is also adopted in natural language processing [15] to let the neural network focus on discriminative words. Similarly, if local regions in an image are "shuffled", irrelevant regions that are non-critical to fine-grained recognition will be neglected, and the network will be forced to classify images based on the discriminative local details. With RCM, the visual appearance of the image has been substantially changed. As shown in the bottom row of Figure 3, though it becomes more difficult for recognition, bird experts can still spot the difference easily. Car enthusiasts can distinguish car models by only examining parts of car [34]. Similarly, the neural network also needs to learn expert knowledge to classify the destructed images.

It is worth noting that "destruction" is not always beneficial. As a side effect, RCM introduces several noisy visual patterns as in Figure 3. To offset the negative impact, we apply an adversarial loss to distinguish original images from destructed ones. As a result, the effect of noisy patterns can be minimized, keeping only beneficial local details. Conceptually, the adversarial and classification losses work in an adversarial manner to carefully learn from "destruction".

For "Construction", a region alignment network is introduced to restore the original region arrangement, which acts in the opposite way of RCM. By learning to restore the original layout as in [19, 6], the network needs to understand the semantics of each region, including those discriminative ones. Through "construction", the correlation between different local regions can be modeled.

The main contributions are summarized as follows:

- A novel "Destruction and Construction Learning (DCL)" framework is proposed for fine-grained recognition. For destruction, the region confusion mechanism (RCM) forces the classification network to learn from discriminative regions, and the adversarial loss prevents over-fitting the RCM-induced noisy patterns. For construction, the region alignment network restores the original region layout by modeling the semantic correlation among regions.
- State-of-the-art performances are reported on three standard benchmark datasets, where our DCL consistently outperforms existing methods.
- Compared to existing methods, our proposed DCL does not need extra part/object annotation and introduces no computational overhead at inference time.

## 2. Related works

Researches for fine-grained image recognition task mainly proceed along two dimensions. One is learning better visual representations from the original image directly [26, 25, 28] and the other is using part/attention based methods [41, 42, 7, 13] to obtain discriminative regions in images and learn region-based feature representations.

Due to the success of deep learning, fine-grained recognition methods have shifted from multistage frameworks based on hand-crafted features [39, 36, 23, 10] to multi-stage frameworks with CNN features [13, 31, 29]. Second order bilinear feature interactions were shown to have a significant improvement for visual representations learning [16, 30]. This method was later extended to a series of related works with further improvements [12, 4, 8]. Deep metric learning is also used to capture subtle visual differences. Zhang *et al*. [40] introduced label structures and a generalization of triplet loss to learn fine-grained feature representations. Chen *et al*. [27] investigate simultaneously predicting categories of different levels in the hierarchy and integrating this structured correlation information into the network by an embedding method. However, these pairwise neural network models often bring complex network computing.

There is also a large amount of part localization based methods proposed regarding the theory that the object parts are essential to learning discriminative features for fine-grained classification [32]. Fu *et al*. [7] proposed a reinforced attention proposal network to obtain discriminating attention regions and region-based feature representation of multiple scales. Sun *et al*. [20] proposed a one-squeeze multi-excitation module to learns multiple attention region features of each input image, and then apply a multi-attention multi-class constraint in a metric learning framework. Zheng *et al*. [42] adopted a channel grouping network to generate multiple parts by clustering, then classi-

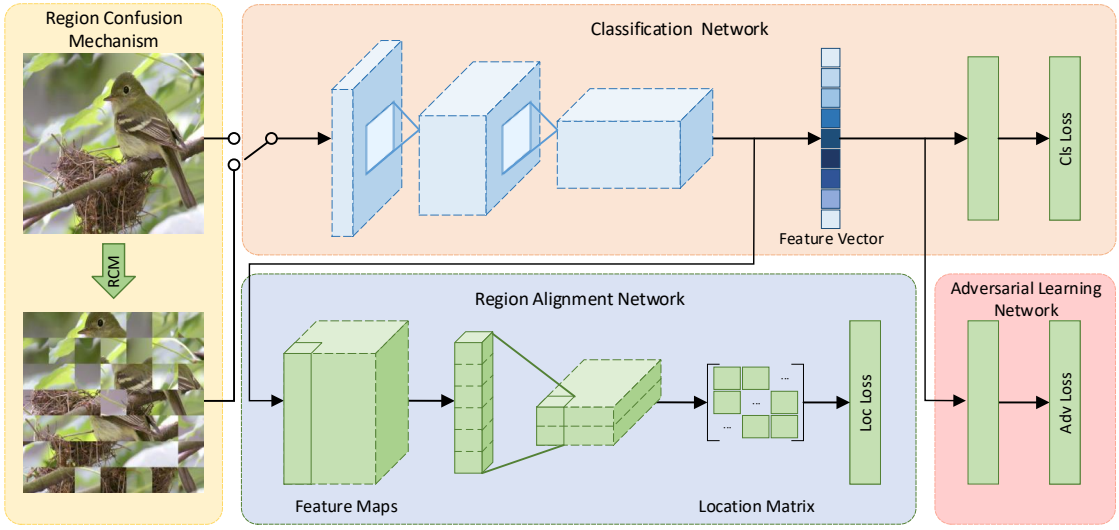


Figure 2. The framework of the proposed DCL method which consists of four parts. (1) Region Confusion Mechanism: a module to shuffle the local regions of the input image. (2) Classification Network: the backbone classification network that classifies images into fine-grained categories. (3) Adversarial Learning Network: an adversarial loss is applied to distinguish original images from destructed ones. (4) Region Alignment Network: appended after the classification network to recover the spatial layout of local regions.

fied these parts features to predict the categories of input images. Compared with earlier part/attention based methods, some of the recent methods tend to be weak supervised and do not require the annotations of parts or key areas [21, 35]. In particular, Peng *et al*. [21] proposed a part spatial constraint to make sure that the model could select discriminative regions, and a specialized clustering algorithm is used to integrate the features of these regions. Yang *et al*. [35] introduced a method to detect informative regions and then scrutinizes them to make final predictions. However, the correlation among regions is helpful to build deep understanding about the objects, it is usually ignored by previous works. The research [19] also shows that utilizing the location information of regions can enhance the visual representation ability of the neural network and result in improving performance on classification and detection tasks.

Our proposed method differs previous works in three aspects: First, by training classifier with our proposed RCM, the discriminative regions can be automatically detected without using any prior knowledge except object labels. Second, our formulation considers not only the fine-grained local region feature representations but also the semantic correlation among different regions in the whole image. Third, our proposed method is highly efficient, that there is no additional overhead except backbone network feed-forward in prediction time.

# 3. Proposed Method

In this section, we present our proposed Destruction and Construction Learning (DCL) method. As shown in Figure 2, the whole framework is composed of four parts. Please note that only the "classification network" is needed during inference time.

## 3.1. Destruction Learning

The devil is in the details. For fine-grained image recognition, local details are much more important than the global structure. In most cases, different fine-grained categories usually share a similar global structure and only differ in certain local details. In this work, we propose to carefully destruct the global structure by shuffling the local regions for better identifying discriminative regions and learning discriminative features (Section 3.1.1). To prevent the network learning from noisy patterns introduced by destruction, an adversarial counterpart (Section 3.1.2) is proposed to reject RCM-induced patterns that are irrelevant to fine-grained classification.

### 3.1.1 Region Confusion Mechanism

As an analogy [15] to natural language processing, shuffling words in a sentence would force the neural network to focus on discriminative words and neglect irrelevant ones. Similarly, if local regions in an image are "shuffled", the neural network would be forced to learn from discriminative region details for classification.

As shown in Figure 3, our proposed Region Confusion Mechanism (RCM) is designed to disrupt the spatial layout of local image regions. Given an input image $I$, we first uniformly partition the image into $N \times N$ sub-regions denoted by $R_{i,j}$, where $i$ and $j$ are the horizontal and vertical indices respectively and $1 \leq i, j \leq N$. Inspired by [15], our proposed RCM shuffles these partitioned local regions

in their 2D neighbourhood. For the $j^{th}$ row of $R$, a random vector $q_j$ of size $N$ is generated, where the $i^{th}$ element $q_{j,i} = i + r$, where $r \sim U(-k, k)$ is a random variable following a uniform distribution in the range of $[-k, k]$. Here, $k$ is a tunable parameter ($1 \leq k < N$) defining the neighbourhood range. Then we can get a new permutation $\sigma_j^{row}$ of regions in $j^{th}$ row by sorting the array $q_j$, verifying the condition:

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

Similarly, we apply the permutation $\sigma_i^{col}$ to the regions column-wisely, verifying the condition:

$$\forall j \in \{1, ..., N\}, \left|\sigma_i^{col}(j) - j\right| < 2k. \tag{2}$$

Therefore, the region at $(i, j)$ in original region location is placed to a new coordinate:

$$\sigma(i, j) = (\sigma_j^{row}(i), \sigma_i^{col}(j)). \tag{3}$$

This shuffling method destructs the global structure and ensures that the local region jitters inside its neighbourhood with a tunable size.

The original image $I$, its destructed version $\phi(I)$ and its ground truth one-vs-all label $\boldsymbol{l}$ indicating the fine-grained categories are coupled as $\langle I, \phi(I), \boldsymbol{l}\rangle$ for training. The classification network maps input image into a probability distribution vector $C(I, \theta_{cls})$, where $\theta_{cls}$ is all learnable parameters in the classification network. The loss function of the classification network $\mathcal{L}_{cls}$ can be written as:

$$\mathcal{L}_{cls} = -\sum_{I \in \mathcal{I}} \boldsymbol{l} \cdot \log\left[C\left(I\right) C\left(\phi(I)\right)\right], \tag{4}$$

where $\mathcal{I}$ is the image set for training.

Since the global structure has been destructed, to recognize these randomly shuffled images, the classification network has to find the discriminative regions and learn the delicate differences among categories.

#### 3.1.2 Adversarial Learning

Destructing images with RCM does not always bring beneficial information for fine-grained classification. For example in Figure 3, RCM also introduces noisy visual patterns as we shuffle the local regions. Features learned from these noise visual patterns are harmful to the classification task. To this end, we propose another adversarial loss $\mathcal{L}_{Adv}$ to prevent overfitting the RCM-induced noise patterns from creeping into the feature space.

Considering the original images and the destructed ones as two domains, the adversarial loss and classification loss work in an adversarial manner to 1) keep domain-invariant patterns, and 2) reject domain-specific patterns between $I$ and $\phi(I)$.


Figure 3. Example images for fine-grained recognition (top) and the corresponding "destructed" images by our proposed RCM (bottom).

We label each image as a one-hot vector $\boldsymbol{d} \in \{0, 1\}^2$ indicating whether the image is destructed or not. A discriminator can be added as a new branch in the framework to judge whether an image $I$ is destructed or not by:

$$D(I, \theta_{adv}) = \text{softmax}(\theta_{adv} C(I, \theta_{cls}^{[1,m]})), \tag{5}$$

where $C(I, \theta_{cls}^{[1,m]})$ is the feature vector extract from the outputs of the $m^{th}$ layer in backbone classification network, $\theta_{cls}^{[1,m]}$ is the learnable parameters from the $1^{st}$ layer to $m^{th}$ layer in the classification network, and $\theta_{adv} \in \mathbb{R}^{d \times 2}$ is a linear mapping. The loss of the discriminator network $\mathcal{L}_{adv}$ can be computed as:

$$\mathcal{L}_{adv} = -\sum_{I \in \mathcal{I}} \boldsymbol{d} \cdot \log\left[D(I)\right] + (\boldsymbol{1} - \boldsymbol{d}) \cdot \log\left[D(\phi(I))\right]. \tag{6}$$

**Justification.** To better understand how the adversarial loss tunes feature learning, we further visualize the features of backbone network ResNet-50 with and without the adversarial loss. Given an input image $I$, we denote the $k^{th}$ feature map in $m^{th}$ layer by $F_m^k(I)$. For ResNet-50, we extract feature from the outputs of the convolutional layer with average pooling next to the last fully-connect layer for adversarial learning. Thus, the response of $k^{th}$ filter in the last convolutional layer for ground truth label $c$ can be measured by $r^k(I, c) = \bar{F}_m^k(I) \times \theta_{cls}^{[m+1]}[k, c]$, where $\theta_{cls}^{[m+1]}[k, c]$ is the weight between the $k^{th}$ feature map and the $c^{th}$ output label.

We compare the responses of different filters for original image and its destructed version in scatter plot shown as Figure 4, where every filter with positive response is mapped to the data point $(r(I, c), r(\phi(I), c))$ in the scatter plot. We can find that the distributions of feature maps trained by $\mathcal{L}_{cls}$ is more compact than those trained by $\mathcal{L}_{cls} + \mathcal{L}_{adv}$. It means that the filters have large responses on the noise patterns introduced by RCM may also have large responses on the original image (as the visual patterns visualized in $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$, there are lots of filters responding

to edge-style visual patterns or irrelevant patterns that are introduced by RCM). These filters may mislead the predictions on the original image.

We also colored the points in scatter plot about backbone network trained by $\mathcal{L}_{cls} + \mathcal{L}_{adv}$, according to the value of

$$\delta_k = \bar{F}_m^k(I) \times \theta_{adv}[k,1] - \bar{F}_m^k(\phi(I)) \times \theta_{adv}[k,2], \quad (7)$$

where $\theta_{adv}[k,1]$ is the weight connecting feature map $F_m^k(\cdot)$ and the label representing orignal image, and $\theta_{adv}[k,2]$ is the weight connecting $F_m^k(\cdot)$ and the label representing destructed image. $\delta_k$ evaluates whether the $k_{th}$ filter tends to be visual patterns in original image or not. It can be observed that the filter respond to noisy visual pattern can be distinguished ($\boldsymbol{D}$ VS. $\boldsymbol{F}$) by using adversarial loss. The points in figures can be divides into three parts. $\boldsymbol{D}$: filters that tend to respond to noisy patters (RCM-induced image features); $\boldsymbol{F}$: filters that tend to respond to global context description (orignal image specific image features); $\boldsymbol{E}$: the vast majority of filters are related to the detailed local region descriptions that enhanced by $\mathcal{L}_{cls}$ (common image feature maps between orignal image and destructed image).

$\mathcal{L}_{cls}$ and $\mathcal{L}_{adv}$ together contribute to the "destruction" learning, where only discriminative local details are enhanced and irrelevant features are filtered out.

### 3.2. Construction Learning

Considering it is the combination of correlative regions in images constitute the complex and diverse visual patterns, we propose another learning method to model the correlation among local regions. Specifically, we propose a region alignment network with region construction loss $\mathcal{L}_{loc}$, that measures the location precision of different regions in images, to induce backbone network to model the semantic correlative among regions by end-to-end training.

Given an image $I$ and its corresponding destructed version $\phi(I)$, the region $R_{i,j}$ located at $(i,j)$ in $I$ is consistent with the region $R_{\sigma(i,j)}$ in $\phi(I)$. Region alignment network works on the output features of one convolution layer of the classification network $C(\cdot, \theta_{cls}^{[1,n]})$, where the $n^{th}$ layer is a convolutional layer. The features are processed by a $1 \times 1$ convolution to obtain outputs with two channels. Then the outputs are handled by an ReLU and an average pooling to get a map with the size of $2 \times N \times N$. The outputs of region alignment network can be written as:

$$M(I) = h\left(C(I, \theta_{cls}^{[1,n]}), \theta_{loc}\right), \quad (8)$$

where the two channels in $M(I)$ correspond to the location coordinates of rows and columns, respectively, $h$ is our proposed region alignment network, and $\theta_{loc}$ is the parameters in region alignment network. We denote the predicted location of $R_{\sigma(i,j)}$ in $I$ as $M_{\sigma(i,j)}(\phi(I))$, predicted location of $R_{i,j}$ in $I$ as $M_{i,j}(I,i,j)$. Both ground truth of

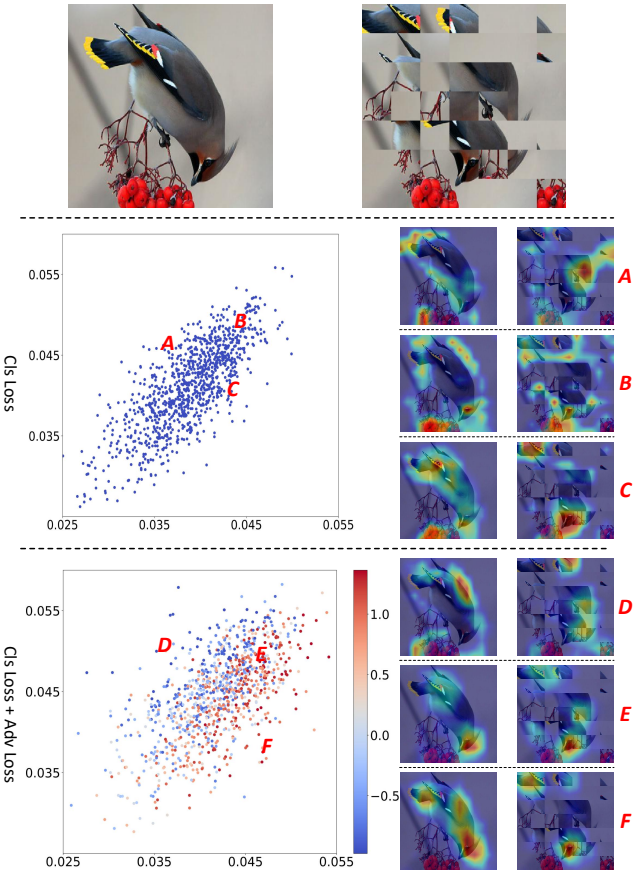


Figure 4. Visualization of filters learned by using $\mathcal{L}_{cls}$ and $\mathcal{L}_{cls}$ + $\mathcal{L}_{adv}$ respectively. The $1^{st}$ row shows the original image $I$ and its destructed version $\phi(I)$. The left side of $2^{nd}$ and $3^{rd}$ rows show the scatter plots about the filters' responses to $I$ and $\phi(I)$. The right side of $2^{nd}$ and $3^{rd}$ rows show the visualization of feature maps belongs to filters that have various responses on $I$ and $\phi(I)$. $\boldsymbol{A}$,$\boldsymbol{D}$: filters with larger response to $\phi(I)$. $\boldsymbol{C}$,$\boldsymbol{F}$: filters with larger response to $I$. $\boldsymbol{B}$,$\boldsymbol{E}$: filters with large responses on both of $I$ and $\phi(I)$. (The figure is best viewed in color.)

$M_{\sigma(i,j)}(\phi(I))$ and $M_{i,j}(I)$ should be $(i,j)$. The region alignment loss $\mathcal{L}_{loc}$ is defined as the $L1$ distance between the predicted coordinates and original coordinates, which can be expressed as:

$$\mathcal{L}_{loc} = \sum_{I \in \mathcal{I}} \sum_{i=1}^{N} \sum_{j=1}^{N} \left| M_{\sigma(i,j)}(\phi(I)) - \begin{bmatrix} i \\ j \end{bmatrix} \right|_1 + \left| M_{i,j}(I) - \begin{bmatrix} i \\ j \end{bmatrix} \right|_1 \quad (9)$$

The region construction loss is helpful to locate the main objects in images and tends to find the correlation among sub-regions. By end-to-end training, the region construction loss can help the classification backbone network to build deep understanding about objects and model the structure information, such as the shape of objects and semantic correlation among parts of object.

### 3.3. Destruction and Construction Learning

In our framework, the classification, adversarial and region alignment losses are trained in an end-to-end manner, in which the network can leverage both enhanced local details and well-modeled object parts correlation for fine-grained recognition. Specifically, we want to minimize the following objective:

$$\mathcal{L} = \alpha\mathcal{L}_{cls} + \beta\mathcal{L}_{adv} + \gamma\mathcal{L}_{loc}. \tag{10}$$

Figure 2 shows the architecture of the DCL framework. The destruction learning mainly helps to learn from discriminative regions, while the construction learning helps to re-arrange the learned local details according to semantic correlation among regions. Hence, DCL yields to a set of complex and diverse visual representations based on the well-structured detail features from discriminative regions.

Note that only $f(\cdot, \theta_{cls}^{[1,l]})$ is used for predicting the category label of given images. Thus, there is no external computational overhead except the backbone classification network for inference.

## 4. Experiments

We evaluate the performance of our proposed DCL on three standard fine-grained object recognition datasets: CUB-200-2011 (CUB) [3], Stanford Cars (CAR) [14] and FGVC-Aircraft (AIR) [18]. We do not use any bounding box/part annotations in all our experiments.

### 4.1. Implementation Details

**Backbone network:** We evaluate our proposed method on two classification widely used backbone networks: ResNet-50 [9] and VGG-16 [25]. These two networks are pre-trained on ImageNet dataset. The category label of the image is the only annotation used for training. The input images are resized to a fixed size of $512 \times 512$ and randomly cropped into $448 \times 448$. Random rotation and random horizontal flip are applied for data augmentation. All above settings are standard in the literature. To recognize high-resolution images on VGG-16 without sub-sampling, the first two fully connected layers in VGG-16 are transformed into two convolution layers respectively. For all the experiments in this paper, the feature maps of the last convolutional layer of backbone network are feed into the region alignment network, and the feature vector formed by the output of average pooling following the last convolutional layer is feed into the adversarial learning network.

The number of regions $N$ in RCM is based on the backbone network and the size of the input image. The width $w$ and length $h$ of the region should be divisible by the stride of the last convolutional layer, which is 32 for VGG-16 and ResNet-50. Meanwhile, to ensure the feasibility of region alignment, the width and height of the input image should also be divisible by $N$. Without special mention, the default value of division number $N$ for RCM is set to 7 in this paper. The influence of choice of $N$ is discussed in Section 4.4.

All models in experiments were trained for 180 epochs, and learning rates decay by a factor of 10 for every 60 epochs. At test time, RCM is disabled, and the networks structures for adversarial loss and region construction are removed. The input images are center cropped and then feed into the backbone classification network for final predictions.

### 4.2. Performance Comparison

The results on CUB-200-2011, Stanford Cars, and FGVC-Aircraft are presented in Table 1. Considering that some of the compared methods use image-level labels or bounding box annotations, the information of extra annotations is also presented in parentheses for direct comparisons. The single model and single crop performance of our proposed DCL achieved state-of-the-art with no extra annotation on all of the three datasets.

We set $\alpha = \beta = 1$ for all experiments reported in this paper. For non-rigid objects recognition tasks like CUB-200-2011, the correlation among different regions is important for building deep understanding about objects. Thus we set $\gamma = 1$. While for rigid objects recognition tasks like Stanford Cars and FGVC-Aircraft, parts of objects are discriminative and complementary. Thus object and part location may play a significant role [34]. We set $\gamma = 0.01$ for rigid objects recognition tasks to highlight the role of destruction learning in learning detail visual representations from discriminative regions. Different from other fine-grained categories like bird and car, the structure of aircraft can change with their design significantly [18]. For example, the number of wings, undercarriages, wheels per undercarriage, engines, etc. varies. Thus we set $N$ as 2 for DCL on FGVC-Aircraft in Table 1 to retain the structure information to a certain extent.

Tables 1, 2 show that our ResNet-50 baseline is already very competitive. Luckily, our proposed DCL can still outperform the strong baseline with a large margin (e.g., 2.3% absolute improvement on average) on all of the three tasks.

### 4.3. Ablation Studies

We conduct ablation studies to understand different components in our proposed DCL. We design different runs in three datasets using ResNet-50 as the backbone network and report the results in Table 2. The results show that the proposed DCL boosts the performance significantly. The performance improvement caused by destruction learning (DL) proves that a well-structured visual feature space that distinguishing the noisy visual pattern, detail visual pattern and the global visual pattern are beneficial to fine-grained recognition task. Likewise, the shape and constitution in-

| Method | Base Model | Accuracy (%) | | |
|---|---|---|---|---|
| | | CUB-200-2011 | Stanford Cars | FGVC-Aircraft |
| CoSeq(+BBox) [13] | VGG-19 | 82.8 | 92.8 | - |
| FCAN(+BBox) [17] | ResNet-50 | 84.7 | 93.1 | - |
| B-CNN [16] | VGGnet | 84.1 | 91.3 | 84.1 |
| HIHCA [2] | VGG-16 | 85.3 | 91.7 | 88.3 |
| RA-CNN [7] | VGG-19 | 85.3 | 92.5 | 88.2 |
| OPAM [21] | VGG-16 | 85.8 | 92.2 | - |
| Kernel-Pooling [5] | VGG-16 | 86.2 | 92.4 | 86.9 |
| Kernel-Pooling [5] | ResNet-50 | 84.7 | 91.1 | 85.7 |
| MA-CNN [42] | VGG-19 | 86.5 | 92.8 | 89.9 |
| DFL-CNN [33] | ResNet-50 | 87.4 | 93.1 | 91.7 |
| DCL | VGG-16 | 86.9 | 94.1 | 91.2 |
| DCL | ResNet-50 | **87.8** | **94.5** | **93.0** |

Table 1. Comparison results on three different standard datasets. Base Model means the backbone network used in the method.

| Method | Accuracy (%) | | |
|---|---|---|---|
| | CUB | CAR | AIR |
| ResNet-50 | 85.5 | 92.7 | 90.3 |
| + RCM | 86.2 | 93.4 | 89.9 |
| DL | 87.2 | 94.4 | 91.6 |
| CL | 86.7 | 94.1 | 90.7 |
| DCL | **87.8** | **94.5** | **92.2** |

Table 2. Ablation studies of the proposed method regarding recognition accuracy on three different datasets. ResNet-50: ResNet-50 finetuned on fine-grained tasks. + RCM: Model trained by $\mathcal{L}_{cls}$. DL: Model trained by $\mathcal{L}_{cls} + \mathcal{L}_{adv}$. CL: Model trained by $\mathcal{L}_{cls} + \mathcal{L}_{loc}$. DCL: Model trained by $\mathcal{L}$.

| N=Divisor($\frac{448}{32}$) | CUB | CAR | AIR |
|---|---|---|---|
| 1 | 85.5% | 92.7% | 90.3% |
| 2 | 86.5% | 93.5% | **93.0%** |
| 7 | **87.8%** | **94.5%** | 92.2% |
| 14 | 85.7% | 93.0% | 92.1% |

Table 3. The recognition accuracy on three datasets of the proposed method by using different $N$.

formation of objects modeled by construction learning (CL) can further improve the performance of fine-grained classification model. Moreover, the adversarial learning and region construction are highly complementary.

### 4.4. Discussions

**Partition Granularity ($N$):** The number of partitions $N$ for RCM is an important parameter for the proposed method. Table 3 shows the recognition accuracy on three datasets with all feasible $N$ with the size of input images $448 \times 448$.

It can be observed that the recognition accuracy increases first and then decreases while $N$ increases. The best performance is achieved at $N = 7$ on CUB-200-2011 and Stanford Cars. For experiments on FGVA-Aircraft, our proposed method can still get better performance than the state-of-the-art method with 0.5% absolute improvement even if we set $N = 7$. In general, if we set $N$ as a small number, the advantage of our proposed method would likely to be restricted. On the other hand, if we set $N$ bigger, the visual patterns can be learned from regions would be more limited, and the region construction network would be more difficult to converge. In particular, the performance of our proposed DCL is equivalent to ResNet-50 baseline when setting $N = 1$.

**Ratio of Destructed Images in a Min-batch:** The default ratio of original images and destructed images in a min-batch is set as $1 : 1$. Table 4 shows the recognition accuracy on CUB-200-2011 with this ratio ranging from $1 : 0$ to $0 : 1$. As shown, the performance decreases by a large margin when we set the ratio as $0 : 1$, since there is no global context information in the training data.

| Ratio | 1:0 | 1:1 | 1:2 | 1:3 | 0:1 |
|---|---|---|---|---|---|
| Accuracy(%) | 85.5 | **87.8** | 86.8 | 86.5 | 84.1 |

Table 4. The recognition accuracy on CUB-200-2011 of the model trained with different composition of training samples. The ratio represents the proportion of original images and images with RCM in one batch.

**Feature Visualization:** We visualize the feature maps of the last convolution layer in Figure 5. Comparing the feature maps from baseline model and proposed method, we can find that the feature map responses of DCL are more concentrated in discriminative regions. With different shuffling, the discriminative regions can be consistently highlighted by DCL based model, which demonstrating the robustness of our DCL method.

**Object Localization:** We also tested DCL on weakly supervised object localization task on VOC2007 dataset using SPN [43]. We choose Pointing Localization Accuracy ($PL_{Acc}$) as the evaluation criterion, which measures

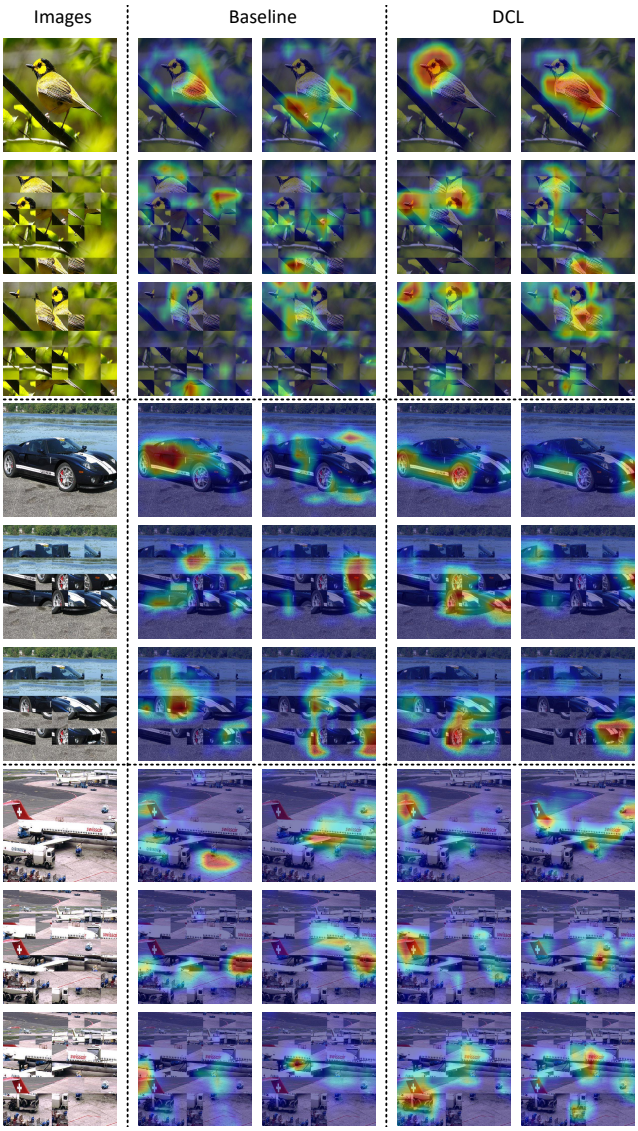


Figure 5. Visualization of the feature maps from the last convolution layer of ResNet-50. For each dataset, the first column shows the orignal image and two destructed versions; the $2^{nd}$ and $3^{rd}$ columns show the feature maps of two filters from baseline ResNet-50; the $4^{th}$ and $5^{th}$ columns show the feature maps of two different filters from the ResNet-50 guided by DCL. This figure is best viewed in color.

whether the network can locate the correct regions of the target. The experimental results is shown in Table 5. We can find that, after applying DCL, $PL_{Acc}$ was improved from $87.5\%$ to $88.7\%$, which serves another numerical evidence that DCL is helpful to learn correct regions.

**Destruction Hyperparameter ($k$):** Since RCM in our proposed method requires the selection of a hyperparameter $k$, we conduct experiments to study the sensitivity of classification performance to the choice of $k$ in Table 6. The recognition accuracy improved, and then decreased as $k$ increases. The best performance is obtained at $k = 2$. In particular, the accuracy decreased slowly when $k$ increased from 2 to 6, which indicates that our method is not particularly sensitive to $k$.

| Method | VGG-16 |
|---|---|
| Center [24] | 69.5 |
| Deconv [37] | 75.5 |
| Grad [24] | 76.0 |
| c-MWP [38] | 80.0 |
| SPN [43] | 87.5 |
| DCL | **88.1** |

Table 5. Pointing localization accuracy (%) on VOC2007 test set. *Center* is a baseline method which uses the image centers as estimation of object centers.

| $k$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| Acc.(%) | 85.5 | 86.7 | **87.8** | 87.6 | 87.4 | 87.3 | 87.2 |

Table 6. The recognition accuracy on CUB-200-2011 of the model trained with different value of $k$.

**Model Complexity:** During training, DCL only requires a simple operation (RCM) and two lightweight network structures (Adversarial Learning Network and Region Alignment Network). For ResNet-50 + DCL, there are 8,192 new parameters introduced by DCL, which is only **0.034%** more parameters than the baseline ResNet-50. Since there are only negligible additional parameters in DCL, the network is efficient to train. Moreover, it takes the same number of iterations as the baseline for finetuning the network upon convergence.

During testing, only the backbone classification network is activated. Compared with the ResNet-50 baseline, our method yields a significantly better result (**+2.3%**) with the same time cost at inference, which adds extra practical value to our proposed method.

## 5. Conclusion

In this paper, we propose a novel DCL framework for fine-grained image recognition. The destruction learning in DCL enhances the difficulty of recognition to guide the network learn expert knowledge for fine-grained recognition. While the construction learning can model the semantic correlation among parts of the object. Our method does not require extra supervision information and can be trained end-to-end in one stage. Extensive experiments against state-of-the-art methods exhibit the superior performances of our method on various fine-grained recognition tasks. Also, our proposed method is lightweight, easy to train, agile for inference and has a good practical value.

**Acknowledgement.** This work is partly supported by National Natural Science Foundation of China No. 61602463.

## References


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

[2] S. Cai, W. Zuo, and L. Zhang. Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization. In *2017 IEEE International Conference on Computer Vision*, pages 511–520, Oct 2017.

[3] W. Catherine, B. Steve, W. Peter, P. Pietro, and B. Serge. The caltech-ucsd birds-200-2011 dataset. (CNS-TR-2011-001), 2011.

[4] Y. Chaojian, Z. Xinyi, Z. Qi, Z. Peng, and Y. Xinge. Hierarchical bilinear pooling for fine-grained visual recognition. pages 595–610, 2018.

[5] Y. Cui, F. Zhou, J. Wang, X. Liu, Y. Lin, and S. Belongie. Kernel pooling for convolutional neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3049–3058, July 2017.

[6] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *2015 IEEE International Conference on Computer Vision*, pages 1422–1430, Dec 2015.

[7] J. Fu, H. Zheng, and T. Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4476–4484, July 2017.

[8] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell. Compact bilinear pooling. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 317–326, June 2016.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, June 2016.

[10] C. Huang, Z. He, G. Cao, and W. Cao. Task-driven progressive part localization for fine-grained object recognition. *IEEE Transactions on Multimedia*, 18(12):2372–2383, Dec 2016.

[11] S. Huang, Z. Xu, D. Tao, and Y. Zhang. Part-stacked cnn for fine-grained visual categorization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1173–1182, June 2016.

[12] S. Kong and C. Fowlkes. Low-rank bilinear pooling for fine-grained classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 7025–7034, July 2017.

[13] J. Krause, H. Jin, J. Yang, and L. Fei-Fei. Fine-grained recognition without part annotations. In *2015 IEEE Conference on Computer Vision and Pattern Recognition*, pages 5546–5555, June 2015.

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

[15] G. Lample, A. Conneau, L. Denoyer, and M. Ranzato. Unsupervised machine translation using monolingual corpora only. 2018.

[16] T. Lin, A. RoyChowdhury, and S. Maji. Bilinear cnn models for fine-grained visual recognition. In *2015 IEEE International Conference on Computer Vision*, pages 1449–1457, Dec 2015.

[17] X. Liu, T. Xia, J. Wang, and Y. Lin. Fully convolutional attention localization networks:efficient attention localization for fine-grained recognition. *CoRR*, abs/1603.06765, 2016.

[18] S. Maji, E. Rahtu, J. Kannala, M.B. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. *CoRR*, abs/1306.5151, 2013.

[19] N. Mehdi and F. Paolo. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Computer Vision – ECCV 2016*, pages 69–84, Cham, 2016. Springer International Publishing.

[20] S. Ming, Y. Yuchen, Z. Feng, and D. Errui. Multi-attention multi-class constraint for fine-grained image recognition. pages 834–850, 2018.

[21] Y. Peng, X. He, and J. Zhao. Object-part attention model for fine-grained image classification. *IEEE Transactions on Image Processing*, 27(3):1487–1500, March 2018.

[22] Pau Rodríguez, Josep M Gonfaus, Guillem Cucurull, F XavierRoca, and Jordi Gonzalez. Attend and rectify: a gated attention mechanism for fine-grained recovery. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 349–364, 2018.

[23] X. Shu, J. Tang, G. Qi, Z. Li, Y. Jiang, and S. Yan. Image classification with tailored fine-grained dictionaries. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(2):454–467, Feb 2018.

[24] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. 2013.

[25] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[26] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, June 2015.

[27] C. Tianshui, W. Wenxi, G. Yuefang, D. Le, L. Xiaonan, and L. Liang. Fine-grained representation learning and recognition by exploiting hierarchical semantic embedding. In *The 26th ACM International Conference on Multimedia*, MM '18, pages 2023–2031. ACM, 2018.

[28] J. Wang, J. Fu, Y. Xu, and T. Mei. Beyond object recognition: Visual sentiment analysis with deep coupled adjective and noun neural networks. August 2016.

[29] Y. Wang, J. Choi, V. I. Morariu, and L. S. Davis. Mining discriminative triplets of patches for fine-grained classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1163–1172, June 2016.

[30] Xing Wei, Yue Zhang, Yihong Gong, Jiawei Zhang, and Nanning Zheng. Grassmann pooling as compact homogeneous bilinear pooling for fine-grained visual classification.

In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 355–370, 2018.

[31] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition*, pages 842–850, June 2015.

[32] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition*, pages 842–850, June 2015.

[33] W. Yaming, M. Vlad I, and D. Larry S. Learning a discriminative filter bank within a cnn for fine-grained recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4148–4157, 2018.

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

[35] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. Learning to navigate for fine-grained classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 420–435, 2018.

[36] B. Yao, G. Bradski, and L. Fei-Fei. A codebook-free and annotation-free approach for fine-grained image categorization. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3466–3473, June 2012.

[37] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[38] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018.

[39] N. Zhang, R. Farrell, F. Iandola, and T. Darrell. Deformable part descriptors for fine-grained recognition and attribute prediction. In *2013 IEEE International Conference on Computer Vision*, pages 729–736, Dec 2013.

[40] X. Zhang, F. Zhou, Y. Lin, and S. Zhang. Embedding label structures for fine-grained feature representation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1114–1123, June 2016.

[41] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan. Diversified visual attention networks for fine-grained object classification. *IEEE Transactions on Multimedia*, 19(6):1245–1256, June 2017.

[42] H. Zheng, J. Fu, T. Mei, and J. Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *2017 IEEE International Conference on Computer Vision*, pages 5219–5227, Oct 2017.

[43] Yi Zhu, Yanzhao Zhou, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Soft proposal networks for weakly supervised object localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1841–1850, 2017.