

# Only Learn One Sample: Fine-Grained Visual Categorization with One Sample Training

Xiangteng He and Yuxin Peng\*

Institute of Computer Science and Technology, Peking University, Beijing, China  
pengyuxin@pku.edu.cn

## ABSTRACT

The progress of fine-grained visual categorization (FGVC) benefits from the application of deep neural networks, especially convolutional neural networks (CNNs), which heavily rely on large amounts of labeled data for training. However, it is hard to obtain the accurate labels of similar fine-grained subcategories because labeling needs professional knowledge, which is labor-consuming and time-consuming. Therefore, it is appealing and significant to recognize these similar fine-grained subcategories with a few labeled samples or even only one for training, which is a highly challenging task. In this paper, we propose OLOS (Only Learn One Sample), a new data augmentation approach for fine-grained visual categorization with only one sample training, and its main novelties are: (1) A **4-stage data augmentation** approach is proposed to increase both the volume and variety of the one training image, which provides more visual information with multiple views and scales. It consists of a 2-stage data generation and a 2-stage data selection. (2) The **2-stage data generation** approach is proposed to produce image patches relevant to the object and its parts for the one training image, as well as produce new images conditioned on the textual descriptions of the training image. (3) The **2-stage data selection** approach is proposed to conduct screening on the generated images in order that useful information is remained and noisy information is eliminated. Experimental results and analyses on fine-grained visual categorization benchmark demonstrate that our proposed OLOS approach can be applied on top of existing methods, and improves their categorization performance.

## CCS CONCEPTS

• **Computing methodologies** Object recognition;

## KEYWORDS

Fine-grained Visual Categorization; One Sample Training; Data Augmentation; Data Generation; Data Selection

## ACM Reference Format:

Xiangteng He and Yuxin Peng\*. 2018. Only Learn One Sample: Fine-Grained Visual Categorization with One Sample Training. In *MM '18: 2018 ACM*

\*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

MM '18, October 22–26, 2018, Seoul, Republic of Korea

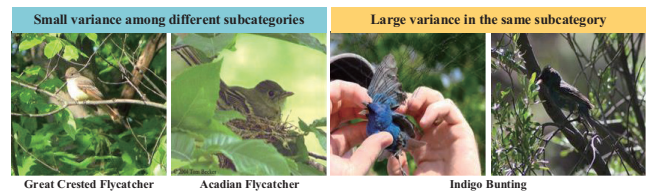
© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-5665-7/18/10...\$15.00

<https://doi.org/10.1145/3240508.3240557>

*Multimedia Conference, Oct. 22–26, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3240508.3240557>

## 1 INTRODUCTION



**Figure 1: Examples of fine-grained subcategories with subtle and local distinctions in the category of birds from CUB-200-2011 dataset [1].**

Fine-grained visual categorization (FGVC) is to distinguish the fine distinctions among similar subcategories, i.e. the fine distinction into species of animals [1] and plants [2], of car [3] and aircraft types [4], etc.. As shown in Figure 1, the variance among different subcategories is small, but that in the same subcategory is large, which make fine-grained visual categorization a highly challenging task. Due to the application of deep learning, especially the convolutional neural networks (CNNs), fine-grained visual categorization has achieved great progress [5–9] in recent years. Deep learning allows computational models that are composed of multiple processing layers to learn the representation of data with multiple levels of abstraction [10]. Its success depends on the advance of hardware accelerator, e.g. fast graphics processing units (GPUs)<sup>1</sup>, which accelerate the training of the networks by 10 ~ 100 times faster, as well as large amounts of labeled data, e.g. the large scale ImageNet dataset [11], which is widely used in computer vision tasks.

Unfortunately, it is difficult and expensive to acquire large amounts of labeled training data [12]. These labeled data is generally acquired by using the service of Amazon Mechanical Turk (AMT)<sup>2</sup> to label the objects in images [13]. AMT is an online platform that asks workers to complete the labeling task, which leads to the fact that workers need to have the specialized knowledge about the data to be labeled. For the datasets of basic-level visual categorization, workers only need to identify categories with large variances, such as birds, dogs, cars and chairs. Such labeling task can be completed by common workers. However, for the datasets of fine-grained visual categorization, which aims to distinguish the fine distinctions among similar subcategories, workers need to

<sup>1</sup><http://www.nvidia.cn>

<sup>2</sup><https://www.mturk.com>

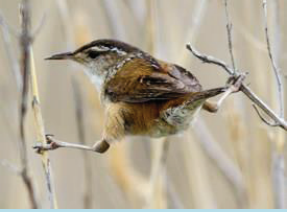



Category	Marsh Wren	Rock Wren	Canyon Wren	Carolina Wren
Image				
Description	<ul style="list-style-type: none"> <li>• Small brown bird with thin bill.</li> <li>• Tail often held upright.</li> <li>• Dark cap.</li> <li>• Whitish eyeline.</li> <li>• Bold black-and-white streaks on back.</li> <li>• Buffy flanks, whitish chest.</li> </ul>	<ul style="list-style-type: none"> <li>• Small songbird.</li> <li>• Pale gray back.</li> <li>• Faintly striped throat.</li> <li>• Long, barred tail.</li> <li>• Long, thin bill.</li> </ul>	<ul style="list-style-type: none"> <li>• Small songbird.</li> <li>• Brown body.</li> <li>• White throat.</li> <li>• Bright rufous, barred tail.</li> <li>• Long, thin, decurved bill.</li> </ul>	<ul style="list-style-type: none"> <li>• Small but chunky bird.</li> <li>• Round body.</li> <li>• Long tail.</li> <li>• The head is large with very little neck.</li> <li>• Distinctive bill.</li> </ul>

Figure 2: Examples of images and their textual descriptions coexisted in the same web page of “All About Birds” website <sup>4</sup>.

have professional knowledge or undergo prior training before the labeling. It is very expensive to employ the experts to label. Gebru et al. [13] report that building a fine-grained car dataset with over 2 million annotations would cost over \$300,000. We can conclude that it is difficult and expensive to acquire a dataset of fine-grained visual categorization.

Fortunately, with the development of Internet, we can easily acquire a few labeled images (may be only one labeled image) from the encyclopedia or e-commerce websites, such as the websites of “Wikipedia” <sup>3</sup>, “All About Birds” <sup>4</sup> and “Cars” <sup>5</sup>. There exists a phenomenon that we should pay attention to: On the above websites, there are some textual descriptions coexisting with the image of the subcategory, which can tell the key characteristics of the subcategory, and provide the complementary information for the visual information. Figure 2 shows some examples of images and their textual descriptions coexisted in the same web page. Inspired by this, an intuitive idea is to learn representations and knowledge of the fine-grained subcategories with a little labeled data, especially, with only one labeled sample per subcategory (e.g. one image and its subcategory label as well as its textual descriptions), which is a significant and challenging problem.

Besides, fine-grained visual categorization has its own challenges: large variance in the same subcategory and small variance among different subcategories, as shown in the first line and second line of Figure 1 respectively. So fine-grained visual categorization with only one sample training is more challenging and significant. To the best of our knowledge, there are few researches on this problem.

Note that it is different from zero-shot learning [14], and has two greater challenges: (i) *Smaller data scale*. As is known to all, good performance of deep learning relies on a large scale of training data. However, we only use one image per subcategory. In zero-shot learning, all training images of seen subcategories are used.

For example, in CUB-200-2011 dataset [1], we only use 200 images, while zero-shot learning uses about 4500 images, 22 times of ours. (ii) *Less extern annotations or prior knowledge*. In zero-shot learning, human-encoded attributes [15], WordNet-hierarchy-derived features [16] and Word2Vec [17] are also used.

Therefore, this paper proposes a new data augmentation approach (OLOS) for fine-grained visual categorization only with one sample training, which consists of a 2-stage data generation and a 2-stage data selection. Its main contributions can be summarized as follows:

- A **4-stage data augmentation** approach is proposed to increase both the volume and variety of the one training image, which provides more visual information with multiple views and scales.
- A **2-stage data generation** approach is proposed to produce image patches relevant to the object and its parts for the one training image, as well as produce new images conditioned on the textual descriptions of the training image.
- A **2-stage data selection** approach is proposed to conduct screening on the generated images of the one training image in order that useful information is remained and noisy information is eliminated.
- Our proposed OLOS approach can be applied on top of existing methods, and improves their categorization accuracies, which has been verified by the experimental results.

The rest of this paper is organized as follows: Section 2 briefly reviews the related works on fine-grained visual categorization, zero-shot learning and data augmentation. Section 3 presents our OLOS approach in detail, and Section 4 introduces the experimental results as well as the experimental analyses. Finally, Section 5 presents the conclusion and future works of this paper.

## 2 RELATED WORK

In this section, we review the related works of fine-grained visual categorization, zero-shot learning and data augmentation.

<sup>3</sup><https://en.wikipedia.org>

<sup>4</sup><https://www.allaboutbirds.org>

<sup>5</sup><https://www.cars.com/>

## 2.1 Fine-grained Visual Categorization

Fine-grained visual categorization is one of the most fundamental and challenging open problems in computer vision, and has drawn extensive attention in both academia and industry. Early works [18, 19] focus on the design of feature representations and classifiers based on the basic low-level descriptors, such as SIFT [20]. However, the performance of these methods is limited due to the low representation ability of the handcrafted features. Recently, deep learning has achieved great success in the domains of computer vision, speech recognition, natural language processing and so on. Inspired by this, many researchers begin to study on the problem of fine-grained visual categorization based on deep learning, and have achieved great progress [5, 6, 21–23].

Since the discriminative characteristics generally localize in the regions of the object and its parts, most existing works generally follow the 2-stage pipeline: first localize the object and its parts, and then extract their features to train classifiers. For the first stage, some works [24, 25] directly utilize the human annotations (i.e. the bounding box of the object and part locations) to localize the object and parts. Since the human annotations are labor-consuming, some researchers begin to only utilize them in the training phase. Zhang et al. propose the Part-based R-CNN [6] to directly utilize the object and part annotations to learn the whole-object and part detectors with geometric constraints between them. This framework is widely used in fine-grained visual categorization.

Recently, fine-grained visual categorization methods begin to focus on how to achieve promising performance without using any object or part annotations. The first work under such weakly supervised setting is the two-level attention model [5], which utilizes the attention mechanism of the convolutional neural networks (CNNs) to select region proposals related to the object and its parts, and achieves promising results even compared with those methods relying on the object and part annotations. Inspired by this work, Zhang et al. [23] incorporate deep convolutional filters for both parts selection and description. He and Peng [21] integrate two spatial constraints for improving the performance of parts selection.

## 2.2 Zero-shot Learning

Zero-shot learning aims to recognize new categories that are not seen in the training phase, so it is a challenging task. Most of the existing methods take the advantage of external knowledge, such as attribute, Wikipedia. Lampert et al. [15] apply the attributes, such as object's color or shape, to recognize the new categories based on their attributes. Elhoseiny et al. [26] utilize the knowledge extracted from Wikipedia to represent the new categories.

It is noted that zero-shot learning is different from the one training sample problem explored in this paper. One training sample problem has smaller data scale and less external annotations or knowledge, which make it more challenging.

## 2.3 Data Augmentation

Data augmentation is widely used to reduce the model overfitting on the training data [27]. Traditional data augmentation in CNNs usually contain generating image translations and horizontal reflections, as well as randomly cropping some patches from the original images, which effectively improve the performance of CNNs. Scale

augmentation [28] and color augmentation [27] are also used to improve the performance and generalization of CNNs [29]

## 3 OUR OLOS APPROACH

In this section, we present the OLOS approach, which is a new approach to augment training data for fine-grained visual categorization only with one sample training, as shown in Figure 3. It applies a 4-stage process to augment the training data, which consists of a 2-stage data generation (i.e. data proposal and data synthesis) and a 2-stage data selection (i.e. data filtering and data re-selection). The 4 stages are as follows: (1) Stage 1: Data proposal is to generate image patches for the one training image. (2) Stage 2: Data filtering is to remove the image patches that are full of background. (3) Stage 3: Data re-selection is to further select the truly useful image patches for the learning of CNNs. (4) Stage 4: Data synthesis is to further generate images to enrich the variety of training data based on the textual descriptions of the training image. In the following subsections, we present each stage in detail.

### 3.1 Stage 1: Data Proposal

The performance of CNN relies on a large amount of labeled training data. So we propose a 2-stage data generation to produce some more images for the one training image, which provides more visual information with multiple views and scales. The two stages are data proposal and data synthesis, which are presented in Section 3.1 and Section 3.4 respectively. (1) Data proposal is to generate thousands of image patches with high objectness by grouping related pixels into regions, some of which may contain the object and its parts. (2) Data synthesis is to generate images corresponding to the textual descriptions of the fine-grained subcategory by generative adversarial network (GAN) [30], which can relate the visual information and textual descriptions of the same fine-grained subcategory.

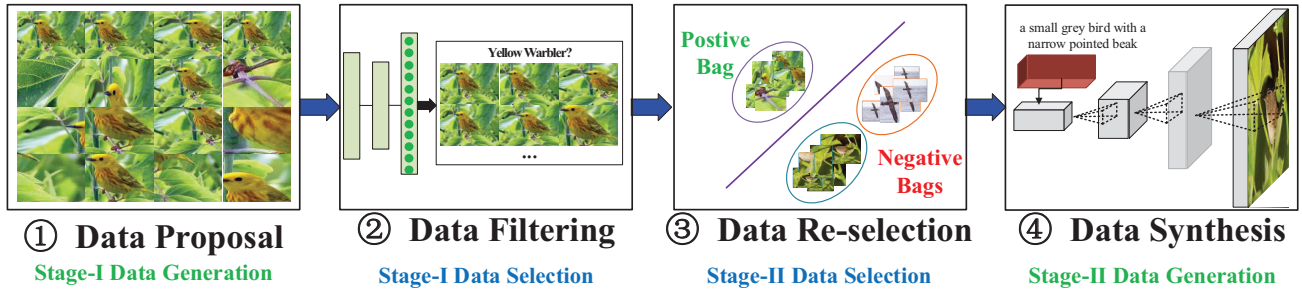
Data proposal can be implemented by bottom-up process, such as selective search [31], an unsupervised and widely-used method to generate such image patches. Some of these image patches are relevant to the object and its parts, which provide multiple views and scales of original image to boost the learning of CNNs.

However, we cannot directly utilize these image patches to augment the training data, which will cause reduction of CNNs' performance, as analyzed in Section 4.5. This is because the bottom-up process has high recall but low precision, which means that some image patches are full of background, and not helpful or even have side effects to the learning of CNNs. So it is significant to filter the bad image patches. We propose a 2-stage data selection to address this problem, which consists of data filtering and data re-selection, as presented in Section 3.2 and Section 3.3.

### 3.2 Stage 2: Data Filtering

In this stage, we aim to eliminate the image patches which are full of background, and remain the image patches that are relevant to the object or its parts. Besides, we train the CNNs progressively with the filtered data. Data filtering and progressive training supplement each other, and further improve the performance of fine-grained visual categorization with one sample training.

First, we train the CNN model with one image per fine-grained subcategory based on a pre-trained model, which is trained on



**Figure 3: An overview of our OLOS approach, which consists of 4 stages: (1) Data proposal. (2) Data filtering. (3) Data re-selection. (4) Data synthesis.**

the ImageNet dataset [11]. Then, we obtain the first CNN model, denoted as  $M_1$ , which has the ability to filter the image patches that are full of background. We feed the image patches generated by data proposal to  $M_1$ , and output the activations of neurons in softmax layer. If the activation of the neuron corresponding the labeled subcategory is the maximum value among all the activations, the image patch is remained, otherwise eliminated. The remained image patches contain the decisive regions of the object for categorization, which are mostly contain the object and its parts, and provide more visual information of the original image with multiple views and scales. Then, we use the remained image patches to augment the training data and fine-tune CNN model as  $M_2$ , which has the more powerful ability of categorization than  $M_1$ .

### 3.3 Stage 3: Data Re-selection

Even the image patches obtained through data proposal and data filtering have played a positive role in the learning of CNNs, a few of them still have side effects. In this stage, we aim to further remove the bad image patches via data re-selection.

A problem we face to is that we do not know which patch is good and which one is bad. Fortunately, we clearly and certainly know there are good image patches in the remained image patches. Therefore, we address this problem via multi-instance learning (MIL). We define the problem as follows:

For the one image  $I$  used for training, which belongs to the fine-grained subcategory  $c_i$ , its remained image patches are denoted as  $P_{c_i} = \{p_1, p_2, \dots, p_n\}$ , where  $n$  is different for different images. They are grouped into a bag, and denoted as  $B_{c_i}$ . Therefore, we can obtain a set of bags for the dataset, denoted as  $B = \{B_{c_1}, B_{c_2}, \dots, B_{c_S}\}$ , where  $S$  denotes the number of fine-grained subcategories in the dataset. We regard the fine-grained visual categorization as multiple binary categorization problems, which means that we need to train  $S$  binary classifiers. For clarity, we describe the process of training  $i$ -th binary classifier in detail.

Since there must be at least one image patch  $p_j \in P_{c_i}$  is a positive example belonging to fine-grained subcategory  $c_i$ , and its label  $y_{p_j} = 1$ , then the bag  $B_{c_i}$  is associated a label  $Y_{c_i}$ , and  $Y_{c_i} = 1$ . For the bags of the other subcategories, their labels are  $-1$ . In general, the relation between the patch label  $y_{p_j}$  and bag label  $Y_{c_i}$  can be

expressed as a set of linear constraints

$$\sum_{p_j \in P_{c_i}} \frac{y_{p_j} + 1}{2} \geq 1, \forall P_{c_i} \text{ s.t. } Y_{P_{c_i}} = 1,$$

$$\text{and } y_{p_j} = -1, \forall P_{c_i} \text{ s.t. } Y_{P_{c_i}} = -1 \quad (1)$$

Then we apply a generalized soft-margin SVM to formulate the multi-instance learning, and it can be expressed as follows:

$$\min_{y_{p_j}} \min_{w, b, \xi} \frac{1}{2} \|w\|^2 + C \sum_{p_j} \xi_{p_j}$$

$$\text{s.t. } \forall p_j : y_{p_j} (\langle w, p_j \rangle + b) \geq 1 - \xi_{p_j}, \xi_{p_j} \geq 0,$$

$$y_{p_j} \in \{-1, 1\}, \text{ and (1) hold.} \quad (2)$$

Finally, we follow Andrews et al. [32] to optimize the SVM, and adopt the implementation by Yang [33]. We can re-select the truly useful image patches by multi-instance learning.

Here we conclude the 2-stage data selection, which actually aims for background-foreground classification via data filtering and multi-instance learning. First, data filtering is used to remove the image patches full of background, which makes MIL learn more valuable knowledge and speed up the learning process. Then, MIL further eliminates the image patches which have large area of background, and reserves the truly useful patches. With the selected image patches by the 2-stage data selection, we obtain the CNN model  $M_3$ .

### 3.4 Stage 4: Data Synthesis

Through stages 1 to 3, we have augmented the training data via generating and selecting the image patches. The augmented images are the local parts of the original image, which provides more visual information with multiple views and scales. Besides, we hope to generate new images to enrich the variety of the training image to boost the learning of CNNs. With the development of Internet, the image and its textual descriptions are easily to be obtained in the same website, as shown in Figure 2. Textual descriptions point out the characteristics of the object in the image, which are complementary to visual information and boost the fine-grained categorization performance [8]. Therefore, in this stage, we generate new images by GANs based on the textual descriptions. GANs generate the new images through learning the data distribution, which requires a large amount of training data. It is hard to learn a good GAN model

with only one training sample. So we augment the training data by data proposal and data filtering, which are the first 2 stages in our OLOS approach.

Specifically, we follow Reed et al. to train a deep convolutional generative adversarial network (DC-GAN) [30] conditioned on text features encoded by a hybrid character-level convolutional-recurrent neural network [34]. In the training phase, we use the original training image, and its augmented data through stages 1 to 3 as well as its textual descriptions. Then we feed the textual description to the generative model of DC-GAN, whose outputs are the generated images corresponding to the textual descriptions. Considering the case that some generated images may have side effects for the learning of CNNs, we also conduct data filtering on these generated images.

### 3.5 Summarization of Our OLOS Approach

Here we give a summarization of our proposed OLOS approach. We generate image patches for one training image respectively in data proposal, and then conduct data filtering and data re-selection to choose the truly useful patches and images. Furthermore, we generate new images via data synthesis, and conduct data filtering on the new generated images. Finally, we obtain the augmented data, including the remained useful image patches and new generated images. After all the 4 stages, we complete the data augmentation, the variety of training data is enriched with multiple views and scales, which plays an important role in the learning of CNNs. Using all the augmented images, we train the CNN model  $M_4$ , which is used to categorize testing images. In our experiment, we use 16-layer VGGNet [28] with compact bilinear pooling [35] as the basic CNN model. It is noted that the basic CNN model can be replaced with others, such as AlexNet [27] and GoogleNet [36]. Besides, our OLOS approach can be on top of any existing methods, and improve their categorization performance.




Category	Image	Description
Heermann Gull		(1) a larger bird with a white head, red beak, grey neck and stomach with black back and wings. (2) this bird has a red bill with a head in white color, it's body and covert though in grey color. (3) medium grey white and black bird with long grey tarsus and long orange beak. ...
Red Legged Kittiwake		(1) birds head is white beak yellow and wings are grey feet are orange and short. (2) this is a white bird with grey wings and orange feet. (3) this large white bird has gray wings, yellow bill and red tarsus and feet. ...
Bohemian Waxwing		(1) the bird has small beak when compared to its body, with black throat, reddish brown crown and gray belly. (2) the bird is grey with an orange crown and a black throat with a black eyebrow. (3) this fierce-looking bird has large eyes, a tufted head and a rounded white belly ...

Figure 4: Examples of images and their textual descriptions in the CUB-200-2011 dataset [1].

## 4 EXPERIMENT

In this section, we present comprehensive experimental results and analyses of our OLOS approach on CUB-200-2011 dataset [1] to verify its effectiveness.

### 4.1 Dataset and Evaluation Metric

(I) **CUB-200-2011** [1] is the most widely-used dataset for fine-grained visual categorization task. It contains 11,788 images of 200 subcategories belonging to birds, 5,994 for training and 5,794 for testing. Each image has detailed annotations: 1 subcategory label, 15 part locations, 312 binary attributes and 1 bounding box. Reed et al. [34] expand the CUB-200-2011 dataset by collecting fine-grained natural language descriptions. Ten single-sentence descriptions are collected for each image, as shown in Figure 4. The textual descriptions are collected through the Amazon Mechanical Turk (AMT) platform, and required at least ten words without any information about the fine-grained subcategories and actions.

In our experiments, we just randomly select one image and its ten single-sentence descriptions as training data for each fine-grained subcategory, which means that only 200 images for training. Figure 5 show some images in the training set, where one image denotes the all training data of each subcategory in our experiment. We have released the training data <sup>6</sup> used in our approach for researchers to follow this work easily and fairly. For the testing set, it still contains 5,794 images.

(II) **Accuracy** is adopted to comprehensively evaluate the categorization performance of our OLOS approach, which is widely used in fine-grained visual categorization [6, 21, 23], and its definition is as follows:

$$Accuracy = \frac{R_a}{R} \tag{3}$$

where  $R$  denotes the number of images in testing set, and  $R_a$  denotes the number of images that are correctly recognized.

### 4.2 Implementation Details

The implementation and the parameters of training the CNN model with one sample training are as follows: We follow Gao et al. [35] to train the compact bilinear model. First, we initialize the weights with the network pre-trained on the ImageNet dataset, and then conduct SGD with a minibatch size of 4. We use a weight decay of  $5e^{-6}$  with a momentum of 0.9 and set the initial learning rate to 1. The learning rate is divided by 0.25 at 600 iterations. We terminate the training at 700 iterations.

### 4.3 Comparisons with State-of-the-art Methods

In this subsection, we present the experimental results on our OLOS approach as well as all the compared methods. For fair comparison, we conduct all the compared methods only with one sample training. Table 1 shows that we achieve the best categorization accuracy and improve the accuracy from 23.28% to 25.33% on CUB-200-2011 dataset [1]. From the result table, we can observe that:

- Only using one sample per subcategory for training, state-of-the-art methods cannot achieve promising results. When using

<sup>6</sup><https://github.com/mumuhe/ACM-MM-2018-OLOS>



Figure 5: Examples of some training images, each of which is the only one training image of its subcategory.

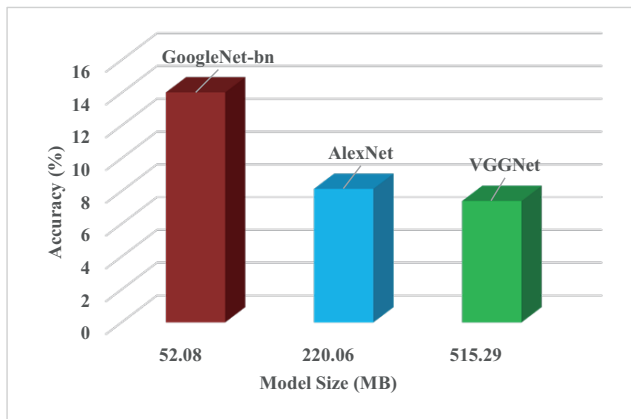


Figure 6: Relation of categorization accuracy and model size.

all the training data, i.e. about 30 images per subcategory, Compact Bilinear [35] can achieve the categorization accuracy of 84.00%, but only 23.28% with one sample training, which demonstrates that fine-grained visual categorization only with one sample training is challenging. But when applying our proposed OLOS approach, the categorization accuracy improves by 2.05%. To further verify the effectiveness of our OLOS approach, we also compare with CVL [8], which jointly models visual and textual information. CVL achieves the accuracy of 18.03%, which is 7.30% lower than ours. The application of our approach brings CVL a 1.65% improvement. It is because that the current state-of-the-art methods are mainly based on deep learning, which relies on large scale of training data. Our OLOS approach can be used on top of any existing state-of-the-arts methods, and improve their categorization accuracy.

- CNNs need a large amount of data to learn representations and knowledge of categorization. Data reduction will cause the sharp decline of categorization accuracy. From Table 2, we can see that the performances of state-of-the-art CNNs decline by at least 51% when reducing the training data to one sample.

Method	Accuracy (%)
<b>Our OLOS Approach</b>	<b>25.33</b>
Compact Bilinear [35]	23.28
CVL [8] + OLOS	19.68
CVL [8]	18.03
GoogleNet-bn [36]	14.03
VGGNet [28]	7.40
AlexNet [27]	8.14

Table 1: Comparisons with state-of-the-art methods on CUB-200-2011 dataset [1].

CNNs	One Sample	All Training Data
GoogleNet-bn [36]	14.03%	82.30%
VGGNet [28]	7.40%	72.20%
AlexNet [27]	8.14%	59.00%

Table 2: Influence of data reduction on state-of-the-art CNNs.

Besides, CNN with more parameters is more dependent on large scale of data, as shown in Figure 6. Model size is directly related to the number of parameters. Valiant [37] prove that if the model has  $N$  parameters, training error will be close to test error once you have more than  $\log N$  training samples. This states: (i) Larger model requires more data. (ii) More data achieves better performance. Therefore, among state-of-the-art CNNs, GoogleNet-bn [36], which has the fewest parameters, achieves the best categorization accuracy in one sample training case.

- From above all, we can conclude that data volume is significant to the performance of CNNs. Our OLOS approach focuses on data augmentation to support the training of CNNs. From 4 stages of data proposal, filtering, re-selection and synthesis, we effectively improve the categorization performance.



Figure 7: Examples of image patches generated by the first 3 stages in our OLOS approach: (a) data proposal, (b) data filtering, and (c) data re-selection. The image patches in red rectangles denote the image patches deleted in corresponding stage.

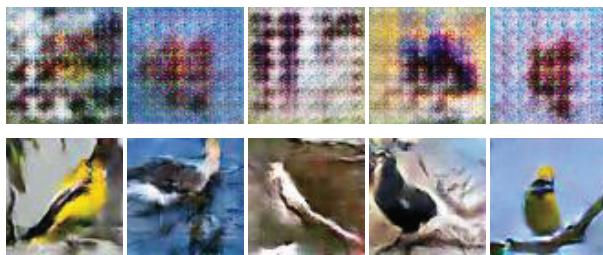


Figure 8: Examples of images generated by the stage 4 in our OLOS approach: data synthesis. The images in the first line are generated by GAN trained with one sample, and those in the second line are generated by GAN trained with augmented data. The images in the same column are corresponding the same subcategory.

#### 4.4 Comparisons with Other Augmentation Methods

Traditional augmentation methods are widely used to enhance the robustness and representation ability of the CNN model, such as rotation, crop, and partition. It’s necessary to compare with them to verify the effectiveness of our approach. In this subsection, we present the categorization results of our OLOS approach compared with the other augmentation methods. The results of all the compared methods in Table 3 are obtained at the same setting that only one training sample per subcategory is used. The compared traditional augmentation methods are as follows:

- Rotation: The images are likely to appear in rotated state due to the shooting angle and image post processing. We rotate the original image by 45°, 90°, 135° and 180° respectively to generate 4 new transformed images.
- Crop: We randomly crop the original image by 75%, 80%, 85% and 90% of the original size respectively to generate 4 new transformed images.
- Random selection: In the first stage of our OLOS approach, i.e. data proposal, we generate hundreds of image patches, which may contain the object, its parts or the background. We randomly select 20 image patches to augment the training data.

- Partition: We directly divide the original image into  $3 \times 3 = 9$  patches.

For the baseline method, we adopt Compact Bilinear [35] using Tensor Sketch, and use its Caffe implementation <sup>7</sup>. We only use one sample per subcategory for baseline method, and then conduct these augmentation methods on the baseline method respectively to verify their effectiveness. These augmentation methods generate new images or image patches to increase the data volume. However, they generate some bad images which only contain a small area of the object or are full of background. These bad images have side effects on the learning of CNNs, so they decline the categorization accuracy compared with baseline method, as shown in Table 3. However, our OLOS approach conducts data proposal, filtering, re-selection and synthesis to generate useful image patches to boost the learning of CNNs.

Method	Accuracy (%)
<b>Our OLOS Approach</b>	<b>25.33</b>
Baseline	23.28
Rotation	22.43
Crop	22.21
Random Selection	12.08
Partition	4.92

Table 3: Comparisons with other augmentation methods.

#### 4.5 Effectiveness of Each Components in Our OLOS Approach

Our OLOS approach focuses on data augmentation, which consists of 4 stages: (1) data proposal, (2) data filtering, (3) data re-selection, and (4) data synthesis. In this subsection, we present the effectiveness of each stage. From Table 4, we can observe that:

- Stage 1: data proposal generates hundreds of image patches. But some of them are bad for categorization, which has discussed in the above subsection. So we need to select the useful image patches from them for the learning of CNNs.
- Stage 2 and stage 3 improve the categorization accuracies by 0.63% and 1.67% respectively. They effectively select useful image patches from stage 1. Figure 7 shows some results of the

<sup>7</sup>[https://github.com/gy20073/compact\\_bilinear\\_pooling](https://github.com/gy20073/compact_bilinear_pooling)

two selection processes. The image patches in the first line are generated by data proposal. The second and third lines show that the bad image patches are filtered at the stage 2 and stage 3 respectively. Stage 3 can further filter worse image patches on the basis of stage 2.

- Stage 4: data synthesis further improves the categorization performance. Figure 8 show some examples of generated new images. The images in the first line and second line are generated by GANs with different training data: one sample per subcategory, augmented data through stages 1, 2 and 3. We can observe that images in second line are much better than those in the first line, which also demonstrates the significance of data augmentation. Even the generated images are not better enough, they boost the categorization accuracy due to the fact that they enrich the variety of the training data.

Method	Accuracy (%)
<b>OLOS-stage1+2+3+4</b>	<b>25.33</b>
OLOS-stage1+2+3	24.95
OLOS-stage1+2	23.91
OLOS-stage1	12.08
Baseline	23.28

**Table 4: Effectiveness of each component in our OLOS approach.**

## 5 CONCLUSION

In this paper, the OLOS approach is proposed for fine-grained visual categorization only with one sample training, which consists of 4 stages: data proposal, data filtering, data re-selection and data synthesis. (1) Data proposal is to generate image patches for one image. (2) Data filtering is to remove the image patches full of background. (3) Data re-selection is to further select the truly useful image patches for the learning of CNNs. (4) Data synthesis is to further generate images to enrich the variety of training image based on its textual descriptions. Experimental results on fine-grained visual categorization benchmark demonstrate that the application of our OLOS approach in state-of-the-art methods achieve the best categorization accuracy.

The future work lies in two aspects: First, we will focus on how to learn better image generation with one image-text pair to enrich the variety of the training data. Second, we will explore the k-sample training problem, which is also significant for the fine-grained visual categorization task.

## ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China under Grant 61771025 and Grant 61532005.

## REFERENCES

- [1] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.
- [2] Maria Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, 2008.
- [3] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *International Conference of Computer Vision Workshop (ICCV)*, pages 554–561, 2013.
- [4] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arxiv:1306.5151*, 2013.
- [5] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 842–850, 2015.
- [6] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *International Conference on Machine Learning (ICML)*, pages 834–849, 2014.
- [7] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *International Conference of Computer Vision (ICCV)*, pages 1449–1457, 2015.
- [8] Xiangteng He and Yuxin Peng. Fine-grained image classification via combining vision and language. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [9] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [10] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
- [11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. pages 248–255, 2009.
- [12] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 28(4):594–611, 2006.
- [13] Timnit Gebru, Jonathan Krause, Jia Deng, and Li Fei-Fei. Scalable annotation of fine-grained categories without experts. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 1877–1881. ACM, 2017.
- [14] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2927–2936, 2015.
- [15] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(3):453–465, 2014.
- [16] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
- [17] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Neural Information Processing Systems (NIPS)*, pages 3111–3119, 2013.
- [18] Lingxi Xie, Qi Tian, Meng Wang, and Bo Zhang. Spatial pooling of heterogeneous features for image classification. *IEEE Transactions on Image Processing (TIP)*, 23(5):1994–2008, 2014.
- [19] Shenghua Gao, Ivor Wai-Hung Tsang, and Yi Ma. Learning category-specific dictionary and shared dictionary for fine-grained image categorization. *IEEE Transactions on Image Processing (TIP)*, 23(2):623–634, 2014.
- [20] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, 2004.
- [21] Xiangteng He and Yuxin Peng. Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 4075–4081, 2017.
- [22] Yu Zhang, Xiu-Shen Wei, Jianxin Wu, Jianfei Cai, Jiangbo Lu, Viet-Anh Nguyen, and Minh N Do. Weakly supervised fine-grained categorization with part-based image representation. *IEEE Transactions on Image Processing (TIP)*, 25(4):1713–1725, 2016.
- [23] Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou, Weiyao Lin, and Qi Tian. Picking deep filter responses for fine-grained image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1134–1142, 2016.
- [24] Lingxi Xie, Qi Tian, Richang Hong, Shuicheng Yan, and Bo Zhang. Hierarchical part matching for fine-grained visual categorization. In *International Conference of Computer Vision (ICCV)*, pages 1641–1648, 2013.
- [25] Thomas Berg and Peter Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 955–962, 2013.
- [26] Mohamed Elhoseiny, Babak Saleh, and Ahmed Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2584–2591, 2013.
- [27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012.
- [28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arxiv:1409.1556*, 2014.
- [29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.



- [30] Scott Reed, Zeynep Akata, Xichen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning (ICML)*, pages 1060–1069, 2016.
- [31] Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International Journal of Computer Vision (IJCV)*, 104(2):154–171, 2013.
- [32] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 577–584, 2003.
- [33] Jun Yang. Mill: A multiple instance learning library. URL <http://www.cs.cmu.edu/juny/MILL>, 2008.
- [34] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 49–58, 2016.
- [35] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 317–326, 2016.
- [36] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.
- [37] Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.