

Fine-Grained Visual-Textual Representation Learning

Xiangteng He^{1b} and Yuxin Peng^{2b}

Abstract—Fine-grained visual categorization is to recognize hundreds of subcategories belonging to the same basic-level category, which is a highly challenging task due to the quite subtle and local visual distinctions among similar subcategories. Most existing methods generally learn part detectors to discover discriminative regions for better categorization performance. However, not all parts are beneficial and indispensable for visual categorization, and the setting of part detector number heavily relies on prior knowledge as well as experimental validation. As is known to all, when we describe the object of an image via textual descriptions, we mainly focus on the pivotal characteristics and rarely pay attention to common characteristics as well as the background areas. This is an involuntary transfer from human visual attention to textual attention, which leads to the fact that textual attention tells us how many and which parts are discriminative and significant to categorization. So, textual attention could help us to discover visual attention in the image. Inspired by this, we propose a fine-grained visual-textual representation learning (VTRL) approach, and its main contributions are: 1) fine-grained visual-textual pattern mining devotes to discovering discriminative visual-textual pairwise information for boosting categorization performance through jointly modeling vision and text with generative adversarial networks, which automatically and adaptively discovers discriminative parts and 2) VTRL jointly combines visual and textual information, which preserves the intra-modality and inter-modality information to generate complementary fine-grained representation, as well as further improves categorization performance. Comprehensive experimental results on the widely used CUB-200-2011 and Oxford Flowers-102 datasets demonstrate the effectiveness of our VTRL approach, which achieves the best categorization accuracy compared with the state-of-the-art methods.

Index Terms—Fine-grained visual categorization, fine-grained visual-textual pattern mining, visual-textual representation learning.

I. INTRODUCTION

FINE-GRAINED visual categorization aims to recognize similar subcategories in the same basic-level category. It is one of the most challenging and significant open problems in multimedia and computer vision areas, which has achieved great progress as well as attracted extensive attention of

Manuscript received June 27, 2018; revised October 1, 2018 and December 16, 2018; accepted January 5, 2019. Date of publication January 14, 2019; date of current version February 5, 2020. This work was supported by the National Natural Science Foundation of China under Grant 61771025 and Grant 61532005. This paper was recommended by Associate Editor S. Satoh. (Corresponding author: Yuxin Peng.)

The authors are with the Institute of Computer Science and Technology, Peking University, Beijing 100871, China (e-mail: pengyuxin@pku.edu.cn).

Color versions of one or more of the figures in this article are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TCSVT.2019.2892802

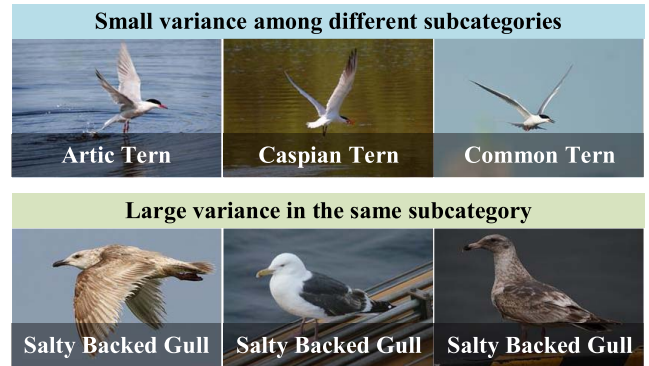


Fig. 1. Examples from CUB-200-2011 dataset [1]. Note that fine-grained visual categorization is a technically challenging task even for humans to recognize these subcategories, due to small variances among different subcategories and large variances in the same subcategory.

academia and industry in recent years. The progress incarnates in three aspects: (1) More fine-grained domains have been covered, such as animal species [1], [2], plant breeds [3], [4], car types [5] and aircraft models [6]. (2) Methodologies of fine-grained visual categorization have achieved promising performance in recent years [7]–[11], due to the application of deep neural networks (DNNs). (3) Some information technology companies, such as Microsoft and Baidu, begin to turn fine-grained visual categorization technologies into their applications.^{1,2}

Fine-grained visual categorization lies in the continuum between basic-level visual categorization (e.g. object recognition) and identification of individuals (e.g. face recognition). Its main challenges can be summarized as the following two aspects: (1) Variances among similar subcategories are subtle and local, because they belong to the same genus. (2) Variances in the same subcategory are large and diverse, due to different poses and views, as well as for animals or plants also because of different living environments and growth periods. For example, as shown in Fig. 1, the images of “Arctic Tern” and “Caspian Tern” look similar in global appearance, but the images of “Salty Backed Gull” look different in the pose, view and feather color. So it is hard for a person without professional knowledge to recognize them.

These subcategories can be distinguished by the subtle and local variances of the discriminative parts. It is crucial for fine-grained visual categorization to localize the object and

¹<https://www.microsoft.com/en-us/research/project/flowerreco-cn/>

²<http://image.baidu.com/?fr=shitu/>

its discriminative parts. Researchers generally adopt a two-stage categorization pipeline: the first stage is to localize the object or its discriminative parts, and the second is to extract their features to categorize the subcategory. For example, Zhang *et al.* [12] utilize R-CNN [13] with geometric constraints to detect object and its parts first, and then extract the features of the object and its parts, finally train one-versus-all linear SVMs for categorization. However, not all the parts are beneficial and indispensable for fine-grained categorization. The conclusive distinctions among subcategories generally locate at a few specific parts, such as the red beak or the black tail. So the categorization performance depends on the number of part detectors and whether the detected parts are discriminative or not. However, mainstream methods generally set the detector number due to their prior knowledge or the experimental validation, which is highly empirical and limited. For example, when the number of part detectors applied in the experiments increase from eight to fifteen, the performance of fine-grained categorization declines, as reported in [14]. Six part detectors are applied by Zhang *et al.* [15] to achieve the best categorization accuracy. He and Peng [16] apply two discriminative parts for fine-grained categorization. They are limited in flexibility, and hard to generalize.

Therefore, it is significant to automatically learn how many and which parts really make sense to fine-grained visual categorization. When human beings see two images of two different subcategories, human visual attention mechanism plays an important role in focusing on the pivotal distinctions between them. Inspired by this, researchers begin to apply human visual attention mechanism in their works, aiming to find the most discriminative characteristics for categorization. Xiao *et al.* [17] propose a two-level attention model (TL Atten), in which object-level attention selects relevant image proposals to a certain object, and part-level attention selects relevant image proposals to the discriminative parts of the object. Fu *et al.* [18] propose a recurrent attention convolutional neural network (RA-CNN) to recursively learn discriminative region attention and region-based feature representation. These works simulate human visual attention mechanism to find discriminative parts for categorization from visual information.

Attention is the behavioral and cognitive process of selectively concentrating on a discrete aspect of information, whether deemed subjective or objective, while ignoring other perceivable information [19]. As is known to all, when human beings give the interpretation of the visual data by textual descriptions, they tend to indicate how many and which parts are distinguishing from other subcategories. These words describing the part attributes are regarded as textual attention, which generally appears frequently in the textual descriptions. This is an involuntary transfer from human visual attention to textual attention. In this transfer process, common characteristics of object and background areas are ignored naturally. Textual attention can be obtained by discovering the frequent item sets in the textual descriptions, which point out the discriminative parts of the subcategory. From Fig. 2, we can see that the frequent item sets contain “red break”, which is

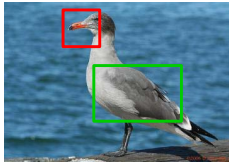
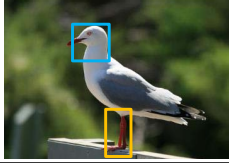
Subcategory	Vision	Text
Heermann Gull		<p>(1) A large bird with different shades of grey all over its body, white and black tail feathers, and a long sharp orange beak.</p> <p>(2) This bird is grey and black in color, with a orange beak.</p> <p>(3) This bird has black outer retires and white inner retires and an orange beak.</p> <p>...</p>
Red Legged Kittiwake		<p>(1) This bird has a white head, breast and belly with gray wings, red feet and thighs, and a red beak.</p> <p>(2) This is a white bird with gray wings, red webbed feet and a red beak.</p> <p>(3) This bird has a white head, nape, breast and belly, with gray wings, the underside of which are black-tipped.</p> <p>...</p>

Fig. 2. Examples of visual and textual attentions. The images come from CUB-200-2011 dataset [1], and text are collected by Reed *et al.* [20] through Amazon Mechanical Turk (AMT) platform.

a discriminative characteristic that distinguishes “Heermann Gull” from “Red Legged Kittiwake”.

Therefore, how to exactly relate textual attention to visual attention and mine the discriminative parts are pivotal to fine-grained visual categorization. This paper proposes a fine-grained visual-textual representation learning (VTRL) approach, and its main contributions are:

- **Fine-grained visual-textual pattern mining** devotes to discovering discriminative visual-textual parts for categorization by jointly modeling vision and text with generative adversarial networks (GANs). Different from existing methods, the localized discriminative parts in this paper could not only tell us how many and which parts are significant for categorization, but also which attributes of parts are distinguishing from other subcategories. The part number is determined automatically and adaptively by textual attention.
- **Visual-textual representation learning** is proposed to combine visual and textual information. Visual stream focuses on the locations of the discriminative parts, while textual stream focuses on the discrimination of the regions. It preserves the intra-modality and inter-modality information to generate complementary fine-grained representation, as well as further improves categorization accuracy.

Our previous conference paper CVL [11] proposes a two-stream model combining vision and language for learning the fine-grained representation. Vision stream learns deep representations from visual information and language stream utilizes textual information to encode salient visual aspects for distinguishing subcategories. The main differences between the proposed VTRL approach and CVL can be summarized as the following three aspects: (1) Our VTRL approach employs textual pattern mining to localize textual attention for exploiting the human visual attention transferred into textual information, which indicates how many and which parts are significant and indispensable for categorization. While CVL directly utilizes the whole textual information, does not mine fine-grained textual attention information. (2) Our VTRL approach employs visual pattern mining based on discovered textual patterns to localize discriminative parts, so that

discriminative parts and objects are both exploited to learn multi-grained and multi-level representations for boosting fine-grained categorization. While CVL only exploits the objects, which ignores the complementary and semantic fine-grained clues provided by the discriminative parts. (3) Our VTRL approach employs fine-grained visual-textual pattern mining to discover the discriminative and significant visual-textual pairwise information via jointly modeling vision and text with GANs, which mines the correlation between textual and visual attention. While CVL only combines vision and text, ignoring to exploit their visual and textual attention, as well as their correlation. Compared with state-of-the-art methods on two widely-used fine-grained visual categorization datasets, our VTRL approach achieves the best categorization accuracy.

The remainder of this paper is organized as follows: We briefly review the related works in Section II. In Section III our proposed VTRL approach is presented in detail. Then Section IV reports the experimental results and analyses. Finally, Section V concludes this paper.

II. RELATED WORK

In this section, we briefly review the related works of fine-grained visual categorization, frequent pattern mining and multi-modal analysis.

A. Fine-Grained Visual Categorization

Since the discriminative regions of image is crucial for fine-grained visual categorization, most existing methods [12], [17] first localize the discriminative regions of image, such as the object and its parts, and then extract their discriminative features for fine-grained categorization. Some methods directly use the annotations of the object [21], [22] and parts [23], [24] to localize the discriminative regions. However, it is not available to obtain the annotations in practical applications, some researchers begin to use the annotations of the object and parts only in the training phase. Zhang *et al.* [25] propose the Deformable Part-based Model (DPM) to localize the discriminative regions with the object and part annotations as the supervised information in the training phase. Further more, PG Alignment [26] is proposed to train part detectors only with object annotation, and localize the discriminative parts in an automatic manner in the testing phase.

Only using object annotation is still not promising in the practical applications. Recently, some works [17], [27], [28] are proposed to localize the discriminative regions in a weakly-supervised manner, which means that neither object nor part annotations are used in both training and testing phases. Xiao *et al.* [17] combine the object and part level attentions to select the discriminative image proposals, which is the first work to localize the discriminative regions without using object and part annotations. Yao *et al.* [27] also propose to combine the two complementary object-level and part-level visual descriptions for better performance. A neural activation constellation (NAC) part model [29] is proposed to train part detectors with constellation model. He and Peng [16] integrate two spatial constraints to select more discriminative proposals

and achieve better categorization accuracy. The aforementioned methods mostly set the detector number due to the prior knowledge or experimental validation, which is highly limited in flexibility and difficult for generalizing to the other domains. Therefore, we attempt to automatically learn how many and which parts really make sense to categorization via fine-grained visual-textual pattern mining.

B. Frequent Pattern Mining

Frequent patterns are itemsets, subsequences, or substructures that appear in a data set with frequency no less than a user-specified threshold [30]. For example, diaper and beer appear frequently together in sales data of a supermarket, which is a frequent pattern. Frequent pattern mining is first proposed by Agrawal *et al.* [31] for market basket analysis. Agrawal *et al.* [32] propose Apriori algorithm to mine frequent patterns in a large transaction database. For textual mining, frequent patterns may be sequential patterns, frequent itemsets, or multiple grams. While for visual mining, frequent patterns may be middle-level feature representation or high-level semantic representation. Han *et al.* [33] propose to mine visual patterns using low-level features. Li *et al.* [34] propose to combine CNN features and association rule mining for discovering visual patterns. Li *et al.* [35] propose a novel multi-modal pattern mining method, which takes textual pattern and visual pattern into consideration at the same time. In this paper, we first utilize Apriori algorithm to discover the textual patterns, and then employ generative adversarial networks (GANs) to mine the relationships between part proposals and textual patterns for better categorization accuracy, which discovers visual and textual patterns at the same time as well as mines the intrinsic correlation between them.

C. Multi-Modal Analysis

Nowadays, multi-modal data, e.g. image, text, video and audio, has been widely available on the Internet. They contains different kinds of information, which are complementary to help achieving comprehensive results in many real-world applications. So it is significant to learn multi-modal representation for boosting the signal-modal tasks [36], [37]. Canonical correlation analysis (CCA) [38] is proposed to learn linear projection matrices, which project features of different modalities into the common space and obtain the common representation. It is widely used for modeling multi-modal data [39]–[41]. Zhai *et al.* [76] propose the joint representation learning method (JRL) to learn projection matrices considering the semantic and correlation information. Due to the advances of deep learning, deep learning based methods have been proposed to boost the performance of multi-modal representation learning. Ngiam *et al.* [42] propose the bimodal autoencoders (Bimodal AE) to model multi-modal data via minimizing the reconstruction error, and learn a shared representation across modalities.

Recently, image and video captioning, which are types of multi-modal analysis, have achieved great progress. Long Short-Term Memory (LSTM) [43] and character-based convolutional networks [44] are widely used in image and

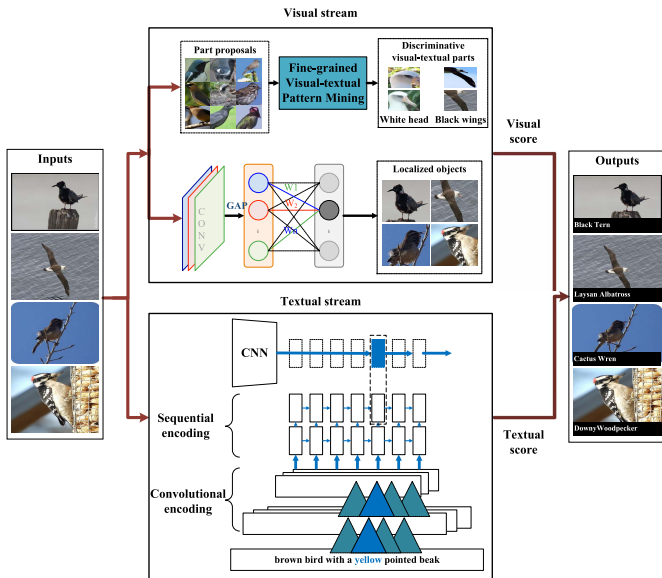


Fig. 3. Overview of our VTRL approach.

video captioning. The architecture of Convolutional and Recurrent Networks (CNN-RNN) is widely used in image and video captioning, and achieves great performance. In this paper, we apply the extension of Convolutional and Recurrent Networks (CNN-RNN) to learn a visual semantic embedding. In this paper, we bring the multi-modal representation learning into fine-grained visual categorization to jointly modeling vision and text for boosting the performance.

III. OUR VTRL APPROACH

A. Overview of Our VTRL Approach

Our approach is based on a very promising and interesting intuition: textual descriptions can point out the discriminative characteristics of images, and provide complementary information with visual information. Therefore, we propose a fine-grained visual-textual representation learning (VTRL) approach, which takes the advantages of visual and textual information jointly as well as exploits the intrinsic correlation between them. Fig. 3 shows our VTRL approach. First, we conduct fine-grained visual-textual pattern mining to discover the discriminative visual-textual parts as shown in Fig. 4. Then, we localize the object region of image to boost the visual analysis. Finally, we propose a visual-textual representation learning approach to jointly model visual and textual streams for better categorization accuracy.

B. Fine-Grained Visual-Textual Pattern Mining

Since human visual attention is described into the form of textual descriptions, we first conduct textual pattern mining to discover the textual attention, which indicates the distinguishing part attributes from other subcategories, such as the shape, size and color of the part. Then, we conduct visual pattern mining to localize the discriminative parts corresponding to the textual patterns discovered by textual pattern mining. The overview of our fine-grained visual-textual pattern mining

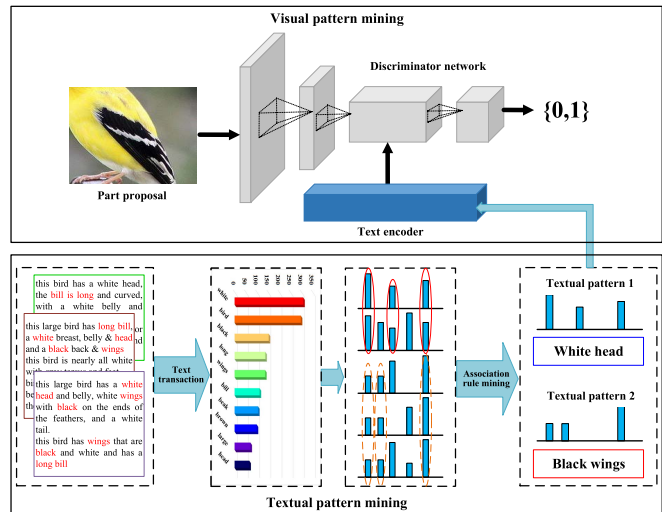


Fig. 4. Overview of our fine-grained visual-textual pattern mining approach. $\{0, 1\}$ denotes the output of the discriminator in GANs, which indicates whether the input part proposal meets the input textual pattern.

approach is shown in Fig. 4. In the following paragraphs, we describe the fine-grained visual-textual pattern mining approach from three aspects: 1) definition of pattern mining, 2) textual pattern mining and 3) visual pattern mining via GANs.

1) *Definition of Pattern Mining*: We first introduce the basic definitions for pattern mining. Assume that there is a set of n items, which is denoted as $X = \{x_1, x_2, \dots, x_n\}$, and the transaction T is a subset of X , i.e. $T \subseteq X$. We also define a transaction database $D = \{T_1, T_2, \dots, T_K\}$ that contains K transactions. Our goal is to discover a particular subset T^* of transactions database X , which can predict the presence of some target item $y \in T_y$, and $T^* \subset T_y$ as well as $y \cap T^* = \emptyset$. T^* refers to frequent itemset in pattern mining literature. The support of T^* denotes how often T^* appears in D and its definition is as follow:

$$supp(T^*) = \frac{|\{T_y | T^* \subseteq T_y, T_y \in D\}|}{K} \quad (1)$$

An association rule $T^* \rightarrow y$ defines a relationship between T^* and a certain item y . Therefore, we aim to find patterns that appear in a transaction there is a high likelihood that y . We define the confidence as follow:

$$conf(T^* \rightarrow y) = \frac{supp(T^* \cup y)}{supp(T^*)} \quad (2)$$

2) *Textual Pattern Mining*: In this paper, we devote to discovering textual patterns, which contain the human visual attention information. First, we remove stop words and punctuations from each textual description. Then we select the words, which appear in at least 10 textual descriptions in our dataset. Build a vocabulary with these selected words, which is used for generating transactions. It is noted that there are no duplicate words in the vocabulary. In order to generate transaction for each textual description, we map each word back to its corresponding word in the vocabulary, then include that corresponding word index in the transaction.

After obtaining the transactions, we perform association rule mining to find the words that frequently appear in textual descriptions, which also means that these words can represent the characteristics of this subcategory. Specifically, we utilize the Apriori algorithm [32] to find a set of patterns P through association rule mining. Each pattern $p \in P$ must satisfy the following criteria:

$$\text{supp}(p) > \text{supp}_{\min} \quad (3)$$

$$\text{conf}(p \rightarrow c) > \text{conf}_{\min} \quad (4)$$

where supp_{\min} and conf_{\min} are thresholds for the support value and confidence value respectively, and c means the image-level subcategory label. After association rule mining, each discovered pattern p contains a set of words.

We want to find some patterns that point out the discriminative parts of the image, which have the semantic meaning. Therefore, we conduct distance constraint on association rule mining as follow:

$$\text{dis}(w_i, w_j) < \text{dis}_{\min} \quad (5)$$

where w_i and w_j mean the i -th and j -th words in the same textual description, and $\text{dis}(\cdot)$ means the interval between the i -th and j -th words. The distance function ensures that the discovered patterns have the semantic meaning. The actual threshold in distance function is set to 4 in the experiments, which is set by the cross-validation method following [12]. Finally, we discover a set of patterns P , i.e. textual attention in the textual descriptions, which contains the information of human visual attention.

3) *Visual Pattern Mining via GANs*: After obtaining the textual attention, we devote to mining the relationship between visual and textual attention, i.e. localize the discriminative parts of images via the guidance of textual attention. Due to the great progress made by generative adversarial networks (GANs), which can generate images based on textual information. In this paper, we employ GANs to break through the gap between visual and textual information, and localize the discriminative parts corresponding to the discovered textual patterns. Specifically, the network architecture follows GAN-CLS [45]. The original training images and their annotated textual descriptions are used to train the GAN-CLS model. We take the alternating strategy to update the generator and discriminator networks, and use ADAM solver [46] to train the model. The training settings, such as learning rate and momentum, are configured following GAN-CLS [45].

It is noted that part proposals and textual patterns are not used to train the GAN-CLS model, as it is unavailable to obtain their matching labels. Reed *et al.* [20] point out that the text embedding based on textual descriptions covers the visual attributes, i.e. textual patterns, such as shape, size and color of the part. GAN-CLS follows [20] to obtain a visually-discriminative vector representation of text descriptions, by using deep convolutional and recurrent text encoders that learn a correspondence function with images. Even using images and textual descriptions in the training phase, GAN-CLS can still learn the correlation between the part proposals and textual patterns. As described in GAN-CLS,

the generator has learned to generate plausible images, and also learned to align them with the conditioning information, and likewise the discriminator must learn to evaluate whether samples from generator meet this conditioning constraint. So we first train GAN-CLS on the datasets in our paper, and then apply the discriminator in GAN-CLS to select the corresponding part proposals for the specific textual patterns, where we take one part proposal as the sample input, and one textual pattern as the conditioning constraint input. The selected part proposals contain discriminative information that helps to distinguish similar subcategories. In the following paragraphs, we introduce the visual pattern mining approach in details.

First, for each image we perform bottom-up process to generate part proposals. In this paper, we utilize selective search method [47] to generate 1000 part proposals for each image. Then we take the part proposals and discovered textual patterns as the inputs of discriminator network, to relate the discovered textual patterns with the corresponding part proposals. For each part proposal, discriminator network outputs a score vector that refers to the degree of correlations between part proposal and textual patterns. We select the part proposal with highest score for each textual pattern, which is one of the most discriminative parts for categorization. They will be utilized as the inputs of visual-textual representation learning.

C. Object Localization

For better categorization performance, we apply an automatic object localization method based on CAM [48] to localize the object in a weakly-supervised manner, which means that neither object nor part annotations are used in both training and testing phases. Through CAM, we can generate a subcategory activation map M_c for each subcategory c , in which the spatial value is calculated as follow:

$$M_c(x, y) = \sum_k w_k^c f_k(x, y) \quad (6)$$

where $f_k(x, y)$ denotes the activation of unit k in the last convolutional layer at spatial location (x, y) , and w_k^c is the weight corresponding to subcategory c for unit k . The subcategory label information is not available in testing phase, so we set subcategory c by the predicted subcategory. After obtaining the activation map for each image, we conduct OTSU algorithm [49] to binarize the image and take the bounding box that covers the largest connected area as the localization of object. The localized object is utilized as the inputs of visual-textual representation learning along with the localized discriminative parts via fine-grained visual-textual pattern mining. Examples of object localization results are shown in Fig. 5. It is noted that we use a variant of VGGNet [50] as CAM following [48]. In order to get a higher spatial resolution, the layers after conv5_3 are removed, resulting in a mapping resolution of 14×14 . Besides, a convolutional layer of size 3×3 , stride 1, pad 1 with 1024 neurons is added, followed by a global average pooling layer and a softmax layer.

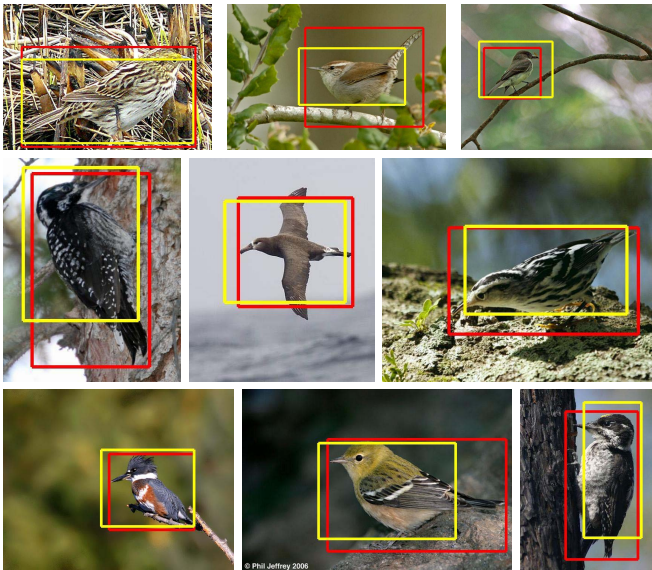


Fig. 5. Examples of object localization results in this paper. The red rectangles indicate the ground truth object annotations, i.e. bounding boxes of objects, and the yellow rectangles indicate the object regions localized by our approach.

D. Visual-Textual Representation Learning

Since visual content and textual descriptions provide complementary information, we jointly model them with a two-stream model for learning visual-textual representations to boost the categorization performance. The two-stream model consists of: 1) visual stream and 2) textual stream.

1) *Visual Stream*: We apply CNN model, e.g. VGGNet [50] in our experiments, as the visual categorization function f . The CNN model is pre-trained on the ImageNet 1 K dataset [51], and then fine-tuned on the fine-grained visual categorization dataset.

For a given image I , we first conduct object localization and fine-grained visual-textual pattern mining respectively to obtain the object b and its n discriminative parts $Pa = \{Pa_1, Pa_2, \dots, Pa_n\}$. Then the object and discriminative parts are cropped from the original image, and saved as images I_b and $I_{Pa} = \{I_{Pa_1}, I_{Pa_2}, \dots, I_{Pa_n}\}$. We feed the original image I and its object image I_b as well as its part images $I_{Pa} = \{I_{Pa_1}, I_{Pa_2}, \dots, I_{Pa_n}\}$ to the CNN model to obtain the predicted visual scores. For the part images, we calculate their mean value as the final part prediction. Finally, we calculate the weighted mean of original prediction, object prediction and part prediction as the final visual prediction.

2) *Textual Stream*: In textual stream, we aim to measure the similarity between visual and textual information. We first apply the deep structured joint embedding method [20] to jointly embed vision (i.e. images) and text (i.e. natural language descriptions for images), which learns a compatibility function of vision and text.

We define the training data as $D = (v_n, t_n, y_n)$, $n = 1, \dots, N$, where $v \in V$ and $t \in T$ denote the vision and text, and $y \in Y$ denotes their subcategory labels. Then we apply the

empirical risk to learn the visual and textual classifier functions $f_v : V \rightarrow Y$ and $f_t : T \rightarrow Y$ as follows:

$$\frac{1}{N} \sum_{n=1}^N \Delta(y_n, f_v(v_n)) + \Delta(y_n, f_t(t_n)) \quad (7)$$

where $\Delta : y \times y \rightarrow \mathbb{R}$ is the 0-1 loss and

$$f_v(v) = \arg \max_{y \in Y} \mathbb{E}_{t \sim T(y)} [F(v, t)] \quad (8)$$

$$f_t(t) = \arg \max_{y \in Y} \mathbb{E}_{v \sim V(y)} [F(v, t)] \quad (9)$$

The compatibility function $F : V \times Y \rightarrow \mathbb{R}$ is defined as the inner product of features from the learnable encoder functions as follows:

$$F(v, t) = \theta(v)^T \phi(t) \quad (10)$$

where $\theta(v)$ is the visual encoder, and $\phi(t)$ is the textual encoder. The visual and textual encoders are implemented by GoogleNet [52] and Convolutional Recurrent Net (CNN-RNN) [20] respectively in our approach. The CNN-RNN model consist of a mid-level temporal CNN hidden layer and a recurrent network. The outputs of the hidden unit over the textual sequence is averaged as the textural features. Then the textual predicted score is defined as a linear accumulation of evidence for compatibility with the image which needs to be recognized.

E. Training Process

In this subsection, we summarize our training process. We train three models for original images, objects and parts respectively. Their detailed training processes are shown in Algorithm 1.

F. Final Prediction

For a given image I , we obtain the visual predicted score from the view of the visual information, and obtain the textual predicted score via measuring the visual and textual information with the shared compatibility function. Due to the fact that joint learning of visual and textual information preserves the intra-modality and inter-modality information to generate complementary information, we fuse the visual and textual predicted results as the final prediction via the follow equation:







$$f(I) = f_v(v) + \beta * f_t(t) \quad (11)$$

where $f_v(v)$ and $f_t(t)$ are the visual and textual predicted scores as mentioned above. β is selected by the cross-validation method following [12], and its value is 2 in our experiments on the two fine-grained datasets.

IV. EXPERIMENTS

A. Datasets

This subsection presents two fine-grained visual categorization datasets adopted in the experiments, including CUB-200-2011 and Oxford Flowers-102 datasets, and their detailed information is described as follows:

Subcategory	Vision	Text	Subcategory	Vision	Text
Heermann Gull		(1)A large bird with different shades of grey all over its body, white and black tail feathers, and a long sharp orange beak. (2)This bird is gray and black in color, with a orange beak. (3)This bird has black outer retices and white inner retices and an orange beak. ...	Primula		(1)This flower is white and yellow in color, with petals that are heart shaped. (2)This white color flower has the simple row of heart shaped petals shaded with orange color at the end. (3)This flower has thick heart shaped white petals and a very yellow star shaped center. ...
Red Legged Kittiwake		(1)This bird has a white head, breast and belly with gray wings, red feet and thighs, and a red beak. (2)This is a white bird with gray wings, red webbed feet and a red beak. (3)Long bird with an orange beak and white feathers with grey colored wings. ...	Silverbush		(1)The flower has petals that are arc white with yellow centers. (2)This flower has large, flat white petals that connect to each other and have a yellow center. (3)This flower is white and yellow in color, with petals that are connected to each other. ...
Bohemian Waxwing		(1)This bird is light gray with a light orange patch on its under-tail coverts, neck and crown, and a black malar stripe and nape. (2)This is a grey bird with a red and yellow tail and a red face. (3)This bird has wings that are gray and black and has a red crown ...	Tree Poppy		(1)This flower has a yellow center surrounded by several large, overlapping white petals with ruffled edges. (2)This flower is white and yellow in color, with petals that are ruffled on the edges. (3)This pretty little flower has large white petals and a yellow center ...

CUB-200-2011

Oxford Flowers-102

Fig. 6. Some examples of vision and text in CUB-200-2011 dataset and Oxford Flowers-102 dataset.

Algorithm 1 Training Process

Input: The training images I and their corresponding textual descriptions T .

Output: The model M .

- 1: Set $M = \{M_{ori}, M_{object}, M_{part}\}$
- 2: Use I to fine-tune the CNN model, which is pre-trained on ImageNet, obtaining the model M_{ori}
- 3: Conduct object localization as described in Section III-C, to get the object regions b of I
- 4: Crop b from I and save as images I_b
- 5: Use I_b to fine-tune M_{ori} , obtaining the model M_{object}
- 6: Follow [45] to train GAN-CLS using minibatch SGD with I and T as pairwise constraints
- 7: Conduct selective search [47] on each image to get part proposals S
- 8: Conduct textual pattern mining to obtain the discriminative textual patterns P for each subcategory
- 9: **for** $k = 1, \dots, n; j = 1, \dots, d$ **do**
- 10: Take k -th part proposal S_k and j -th textual pattern P_j as the input of the generator \mathcal{G} of GAN-CLS
- 11: Perform a feed-forward pass, and output the correlation score of S_k and P_j
- 12: For P_j we select one part proposal with the highest correlation score
- 13: **end for**
- 14: Use the selected part proposals to fine-tune M_{object} , obtaining the model M_{part}
- 15: **return** M .

- **CUB-200-2011.** It is the most widely-used dataset for fine-grained visual categorization task. The visual information comes from the original dataset of CUB-200-2011 [1]. 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. The textual information comes from [20]. They 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 Fig. 6. The natural language descriptions are collected through the Amazon Mechanical Turk (AMT) platform, and are required at least 10 words, without any information of subcategories and actions.

- **Oxford Flowers-102.** Same with CUB-200-2011 dataset, textual information comes from [20], and visual information comes from the original dataset of Oxford Flowers-102 [4], as shown in Fig. 6. It has 8,189 images of 102 subcategories belonging to flowers, 1,020 for training, 1,020 for validation and 6,149 for testing. Each subcategory consists of between 40 and 258 images.

B. Evaluation Metric

Accuracy is adopted to comprehensively evaluate the categorization performances of our VTRL approach as well as compared state-of-the-art methods, which is widely used in fine-grained visual categorization [8], [12], and its definition is as follow:

$$Accuracy = \frac{R_a}{R} \quad (12)$$

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

C. Implementation Details

1) *Fine-Grained Visual-Textual Pattern Mining:* First, we calculate the frequency of each word in the textual descriptions for each subcategory, and select the top-10 words as keywords, and then discover textual frequent patterns via Apriori algorithm [32]. It is noted that we conduct textual pattern mining for each subcategory respectively rather than

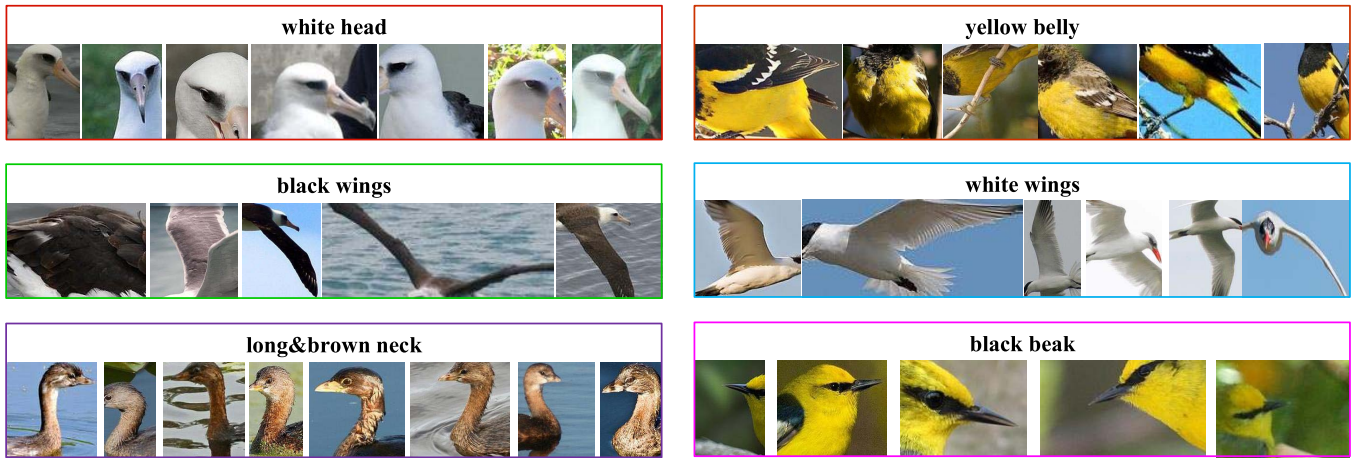


Fig. 7. Examples of the matching between textual patterns and visual patterns in our fine-grained visual-textual pattern mining approach.

all subcategories together, which guarantees that the frequent textual patterns tend to be the descriptions of the discriminative parts, such as “white head”, “black wings” and “long bill”. Second, we conduct selective search [47] on each image to generate part proposals. Finally, we employ discriminator network to relate textual patterns to part proposals, then select the proposal with highest score as the discriminative part for each textual pattern. For each subcategory, the number of parts is set automatically and adaptively based on the discovered textual patterns. Fig. 7 shows some matching examples between textual pattern and visual pattern, which are the discriminative characteristics of the subcategory, such as “long&brown neck”, “yellow belly” and “black beak”.

2) *Visual-Textual Representation Learning*: For textual stream, we apply CNN-RNN [20] as the text encoder to learn a correspondence function with images. In the training phase, we follow Reed *et al.* [20]. For visual stream, we apply the widely-used model 19-layer VGGNet [50] with batch normalization. The model is first pre-trained on ImageNet 1K dataset, and then fine-tuned on the fine-grained visual categorization dataset. Inspired by the strategy adopted by Xiao *et al.* [17], we utilize the pre-trained CNN model as a filter net to select proposals relevant to the object from the generated image proposals by selective search method. We further fine-tune the pre-trained model with the selected image proposals.

D. Comparisons With State-of-the-Art Methods

In this subsection, we present the experimental results of our proposed approach as well as all the compared state-of-the-art methods, as shown in Tables I and II, which demonstrate the effectiveness of our proposed VTRL approach. As shown in Table I, our proposed VTRL approach improves the categorization accuracy from 85.65% to 86.31% on CUB-200-2011 dataset. We divide the compared methods into three groups due to the usage of object and part annotations in these methods.

- *Neither object nor part annotations are used.* Nowadays, researchers focus on how to get better categorization

accuracy under the weakly-supervised setting, which means neither object nor part annotations are used. Most of these methods utilize the attention property of convolutional neural layers to localize the discriminative parts of object for better accuracy, such as Fused CN-Nets [28], RA-CNN [18], PNA [8], TSC [16] and TL Atten [17]. They simulate human visual attention mechanism only from visual information. In our approach, we exploit visual and textual attention simultaneously as well as mine the complementary information between them, which make our proposed approach more effective and obtain a 0.66% higher accuracy than the best performing result of Fused CN-Nets [28]. We also compare with our previous conference work, i.e. CVL [11]. We can see that our VTRL approach brings improvements than CVL by 0.76% and 0.67% respectively on CUB-200-2011 and Oxford Flowers-102 datasets. It is mainly because that the VTRL approach exploits the textual attention to localize discriminative regions, while CVL directly uses the whole textual descriptions and does not consider the discriminative regions in the images.

- *Only one of object and part annotations is used.* These methods utilize object annotation (i.e. bounding box) to train an object detector or learn part detectors, which are to learn more representative features for categorization. In our approach, we utilize CAM [48] to automatically localize the object region of image, which avoids using object annotation. The result of object localization can be seen in Fig. 2. Even using object annotation, these methods achieve lower accuracies than our proposed VTRL approach.
- *Both object and part annotations are used.* In order to obtain better categorization accuracy, some methods utilize both object and part annotations at training phase as well as testing phase. However, these annotations are heavy labor-consuming. In our approach, we get object region and discriminative parts automatically via object localization and fine-grained visual-textual pattern mining respectively without using any annotations. We promote

TABLE I

COMPARISONS WITH STATE-OF-THE-ART METHODS ON CUB-200-2011, SORTED BY AMOUNT OF ANNOTATION USED. “BBOX” AND “PARTS” INDICATE THE OBJECT AND PART ANNOTATIONS (I.E. BOUNDING BOX AND PARTS LOCATIONS) PROVIDED BY THE DATASET

Method	Train Annotation		Test Annotation		Accuracy (%)
	Bbox	Parts	Bbox	Parts	
Our VTRL Approach					86.31
Fused CN-Nets [28]					85.65
CVL [11]					85.55
RA-CNN [18]					85.30
PNA [8]					84.70
TSC [16]					84.69
FOAF [53]					84.63
Low-rank Bilinear [54]					84.21
Spatial Transformer [55]					84.10
Bilinear-CNN [56]					84.10
Multi-grained [57]					81.70
AutoBD [27]					81.60
NAC [29]					81.01
PIR [58]					79.34
TL Atten [17]					77.90
MIL [59]					77.40
VGG-BGLm [60]					75.90
Dense Graph Mining [61]					60.19
Coarse-to-Fine [62]	✓				82.50
PG Alignment [26]	✓		✓		82.80
Triplet-A (64) [63]	✓		✓		80.70
Webly-supervised [64]	✓	✓			78.60
PN-CNN [65]	✓	✓			75.70
Part-based R-CNN [12]	✓	✓			73.50
SPDA-CNN [66]	✓	✓	✓		85.14
Deep LAC [67]	✓	✓	✓		84.10
PBC [68]	✓	✓	✓		83.70
SPDA-CNN [69]	✓	✓	✓		81.01
PS-CNN [14]	✓	✓	✓		76.20
PN-CNN [65]	✓	✓	✓	✓	85.40

TABLE II

COMPARISONS WITH STATE-OF-THE-ART METHODS ON OXFORD FLOWERS-102

Method	Accuracy (%)
Our VTRL Approach	96.89
CVL [11]	96.21
PBC [68]	96.10
NAC [29]	95.34
RIIR [70]	94.01
Deep Optimized [71]	91.30
SDR [71]	90.50
MML [72]	89.45
CNN Feature [73]	86.80
Generalized Max Pooling [74]	84.60
Efficient Object Detection [3]	80.66

the categorization performance through discovering the discriminative and representative object and its parts.

Besides, categorization results on Oxford Flowers-102 dataset are shown in Table II, and also have the similar trend as CUB-200-2011 dataset, while our proposed VTRL approach still keeps the best.

E. Effects of Components in Our VTRL Approach

In this subsection, we conduct two baseline experiments to verify the separate contribution of each component in our proposed VTRL approach. Tables III to V show the accuracies of our proposed VTRL approach as well as the baseline

TABLE III

EFFECTS OF FINE-GRAINED PATTERN MINING AND OBJECT LOCALIZATION FOR VISUAL STREAM

Method	Accuracy (%)
VTRL-visual	85.54
VTRL-visual(w/o object)	83.21
VTRL-visual(w/o parts)	84.79
VTRL-visual(w/o object&parts)	80.82

TABLE IV

EFFECTS OF DIFFERENT COMPONENTS OF OUR PROPOSED APPROACH ON CUB-200-2011

Method	Accuracy (%)
Our VTRL Approach	86.31
VTRL(w/o object)	85.17
VTRL(w/o parts)	85.83
VTRL(w/o object&parts)	84.05

approaches on CUB-200-2011 dataset at the following two aspects.

1) *Effects of Fine-Grained Visual-Textual Pattern Mining and Object Localization:* In our VTRL approach, fine-grained visual-textual pattern mining and object localization generate discriminative parts and object for promoting the categorization accuracy. They make sense to the visual stream and then further impact whole approach. Tables III and IV show the effects of fine-grained visual-textual pattern mining

TABLE V
EFFECTS OF DIFFERENT COMPONENTS OF OUR PROPOSED APPROACH ON CUB-200-2011

Method	Accuracy (%)
Our VTRL Approach	86.31
VTRL-textual	81.81
VTRL-visual	85.54
VTRL(only original image)	80.82
Co-attention [75]	73.90


Subcategory	Vision	Text Rank List(Top3)
Sooty Albatross		(1)This bird has wings that are grey and has a black bill . (2)This bird is gray in color, with a large curved beak. (3)This bird is white and brown in color, and has a black beak .
California Gull		(1)This bird has large feet, a short yellow bill , and a black and white body . (2)This bird has wings that are grey and has a white belly and yellow bill . (3)This bird has a yellow beak as well as a white belly .
Cerulean Warbler		(1)A little bird with a short, grey bill , blue crown , nape , white breast . (2)The bird has a white abdomen, black breast and white throat, blue specks. (3)This bird is blue and white in color with a black beak, and black eye rings.

Fig. 8. Some results of the textual stream.

and object localization to visual stream and our proposed VTRL approach respectively. In the tables, “object” means that object localization is conducted, and “parts” means that fine-grained visual-textual pattern mining is employed. We can observe that considering object localization can achieve better categorization accuracy than considering fine-grained visual-textual pattern mining. This is because that objects contain the global and local features simultaneously, while discriminative parts focus subtle and local characteristics. However, jointly considering object localization and fine-grained visual-textual pattern mining can further improve the categorization accuracy.

Fine-grained visual-textual pattern mining aims to select the part proposals that corresponding to the discovered textual patterns. The relations between part proposals and textual patterns ensure the discrimination and representativeness of selected parts. Some examples of discovered visual-textual patterns are shown in Fig. 7.

2) *Effectiveness of Visual-Textual Representation Learning:* We also present the baseline experiment to verify the effectiveness of visual-textual representation learning. The results are shown in Table V, where “VTRL-textual” means textual stream, “VTRL-visual” means visual stream and “VTRL(only original image)” means only a fine-tuned CNN model is used. We can observe that categorization result of textual stream is promising. From the first line of each row in Fig. 8, we can find that textual description with the highest score always points out the discriminative characteristics of the object, as the red words shows. Combining visual and textual information can further achieve more accurate categorization result, which demonstrates that the two types of information are complementary: visual information focuses on the global and local features, and textual information further points the importance of these features. Fig. 9 shows some example results where the textual and visual streams are complementary. Visual stream is effective for dealing with those images, which have few

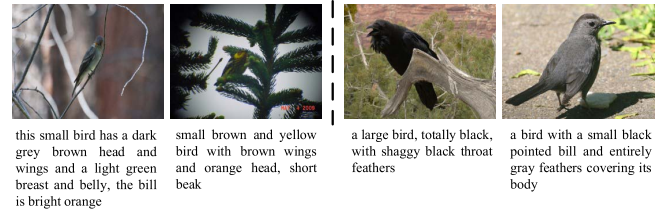


Fig. 9. Some example results where the textual and visual streams are complementary. The left two images are rightly categorized by textual stream, but wrongly categorized by visual stream. The right two images are just the opposite.

discriminative characteristics. Humans can only describe them in a rough way, but cannot describe them in detail, which leads that the textual information carries less useful information to distinguish it from other subcategories. Examples are shown as the right two images. Textual stream is effective for dealing with those images, whose foreground and background are hard to be distinguished by visual stream. But they can be described in details by text, which carries the information of the discriminative characteristics and be helpful for the categorization. Examples are shown as the left two images.

Besides, we also compare our VTRL approach with method based on both textual and visual attention, such as Co-attention [75]. It only achieves the accuracy of 73.90%, which is lower than our VTRL approach. It is mainly because that our VTRL approach discovers the fine-grained visual-textual patterns, which are key hints to the fine-grained categorization.

From the above baseline results, the separate contribution of each component in our proposed VTRL approach can be verified. First, object localization and fine-grained pattern mining discover the discriminative and representative information of image via visual-textual attention. Second, the complementarity between visual and textual information is fully captured by visual-textual representation learning.

V. CONCLUSIONS

In this paper, the fine-grained visual-textual representation learning approach has been proposed. Based on textual attention, we employ fine-grained visual-textual pattern mining to discover discriminative information for categorization through jointly modeling vision and text with GANs. Then, visual-textual representation learning jointly considers visual and textual information, which preserves the intra-modality and inter-modality information to generate complementary fine-grained representation, and further improve categorization performance. Experimental results on two widely-used fine-grained visual categorization datasets demonstrate the superiority of our approach compared with state-of-the-art methods.

As for the future work, we will focus on the following two aspects: First, we will attempt to extend the current two-stream framework into an end-to-end framework for simplifying the process. Second, we will exploit exact and effective methods on relating textual attention and visual attention for more accurate discriminative parts localization as well as better categorization performance.

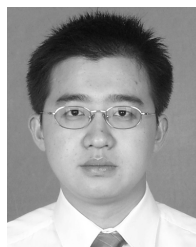
REFERENCES

- [1] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200-2011 dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011.
- [2] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *Proc. CVPR Workshop Fine-Grained Vis. Categorization (FGVC)*, vol. 2, Jun. 2011, p. 1.
- [3] A. Angelova and S. Zhu, "Efficient object detection and segmentation for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2013, pp. 811–818.
- [4] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. 6th Indian Conf. Comput. Vis., Graph. Image Process.*, Dec. 2008, pp. 722–729.
- [5] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 554–561.
- [6] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi. (2013). "Fine-grained visual classification of aircraft." [Online]. Available: <https://arxiv.org/abs/1306.5151>
- [7] C. Huang, Z. He, G. Cao, and W. Cao, "Task-driven progressive part localization for fine-grained object recognition," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2372–2383, Dec. 2016.
- [8] X. Zhang, H. Xiong, W. Zhou, W. Lin, and A. Tian, "Picking neural activations for fine-grained recognition," *IEEE Trans. Multimedia*, vol. 19, no. 12, pp. 2736–2750, Dec. 2017.
- [9] Y. Wang, S. Li, and A. C. Kot, "DeepBag: Recognizing handbag models," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2072–2083, Nov. 2015.
- [10] Y. Wang, S. Li, and A. C. Kot, "On branded handbag recognition," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1869–1881, Sep. 2016.
- [11] X. He and Y. Peng, "Fine-grained image classification via combining vision and language," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7332–7340.
- [12] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based R-CNNs for fine-grained category detection," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 834–849.
- [13] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 580–587.
- [14] S. Huang, Z. Xu, D. Tao, and Y. Zhang, "Part-stacked CNN for fine-grained visual categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1173–1182.
- [15] X. Zhang, H. Xiong, W. Zhou, W. Lin, and Q. Tian, "Picking deep filter responses for fine-grained image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1134–1142.
- [16] X. He and Y. Peng, "Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2017, pp. 4075–4081.
- [17] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 842–850.
- [18] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4476–4484.
- [19] J. R. Anderson, *Cognitive Psychology and Its Implications*. San Francisco, CA, USA: Freeman, 1985.
- [20] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 49–58.
- [21] Y. Chai, V. Lempitsky, and A. Zisserman, "Symbiotic segmentation and part localization for fine-grained categorization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 321–328.
- [22] S. Yang, L. Bo, J. Wang, and L. G. Shapiro, "Unsupervised template learning for fine-grained object recognition," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 3122–3130.
- [23] T. Berg and P. Belhumeur, "POOF: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 955–962.
- [24] L. Xie, Q. Tian, R. Hong, S. Yan, and B. Zhang, "Hierarchical part matching for fine-grained visual categorization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 1641–1648.
- [25] N. Zhang, R. Farrell, F. Iandola, and T. Darrell, "Deformable part descriptors for fine-grained recognition and attribute prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 729–736.
- [26] J. Krause, H. Jin, J. Yang, and L. Fei-Fei, "Fine-grained recognition without part annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5546–5555.
- [27] H. Yao, S. Zhang, C. Yan, Y. Zhang, J. Li, and Q. Tian, "AutoBD: Automated bi-level description for scalable fine-grained visual categorization," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 10–27, Jan. 2018.
- [28] H. Yao, S. Zhang, Y. Zhang, J. Li, and Q. Tian, "One-shot fine-grained instance retrieval," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 342–350.
- [29] M. Simon and E. Rodner, "Neural activation constellations: Unsupervised part model discovery with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1143–1151.
- [30] J. Han, H. Cheng, D. Xin, and X. Yan, "Frequent pattern mining: Current status and future directions," *Data Mining Knowl. Discovery*, vol. 15, no. 1, pp. –86, Aug. 2007.
- [31] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," *ACM SIGMOD Rec.*, vol. 22, no. 2, pp. 207–216, Jun. 1993.
- [32] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. Int. Conf. Very Large Data Bases*, vol. 1215. 1994, pp. 487–499.
- [33] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," *ACM SIGMOD Rec.*, vol. 29, no. 2, pp. 1–12, 2000.
- [34] Y. Li, L. Liu, C. Shen, and A. van den Hengel, "Mid-level deep pattern mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 971–980.
- [35] H. Li, J. G. Ellis, H. Ji, and S.-F. Chang, "Event specific multimodal pattern mining for knowledge base construction," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 821–830.
- [36] R. Hong, J. Tang, H.-K. Tan, C.-W. Ngo, S. Yan, and T.-S. Chua, "Beyond search: Event-driven summarization for Web videos," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 7, no. 4, Nov. 2011, Art. no. 35.
- [37] R. Hong, L. Li, J. Cai, D. Tao, M. Wang, and Q. Tian, "Coherent semantic-visual indexing for large-scale image retrieval in the cloud," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4128–4138, Sep. 2017.
- [38] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, nos. 3–4, pp. 321–377, Dec. 1936.
- [39] H. Bredin and G. Chollet, "Audio-visual speech synchrony measure for talking-face identity verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, vol. 2, Apr. 2007, pp. II-233–II-236.
- [40] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, Dec. 2004.
- [41] B. Klein, G. Lev, G. Sadeh, and L. Wolf, "Associating neural word embeddings with deep image representations using Fisher vectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4437–4446.
- [42] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and Y. A. Ng, "Multimodal deep learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2011, pp. 689–696.
- [43] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [44] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 649–657.
- [45] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. Int. Conf. Mach. Learn. (ICML)*, May 2016, pp. 1060–1069.
- [46] D. P. Kingma and J. Ba. (2014). "Adam: A method for stochastic optimization." [Online]. Available: <https://arxiv.org/abs/1412.6980>
- [47] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Sep. 2013.
- [48] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2016, pp. 2921–2929.
- [49] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man, Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.
- [50] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: <https://arxiv.org/abs/1409.1556>

- [51] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 248–255.
- [52] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
- [53] X. Zhang, H. Xiong, W. Zhou, and Q. Tian, "Fused one-vs-all features with semantic alignments for fine-grained visual categorization," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 878–892, Feb. 2016.
- [54] S. Kong and C. Fowlkes, "Low-rank bilinear pooling for fine-grained classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7025–7034.
- [55] M. Jaderberg, K. Simonyan, A. Zisserman, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 2017–2025.
- [56] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Jul. 2015, pp. 1449–1457.
- [57] D. Wang, Z. Shen, J. Shao, W. Zhang, X. Xue, and Z. Zhang, "Multiple granularity descriptors for fine-grained categorization," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2399–2406.
- [58] Y. Zhang *et al.*, "Weakly supervised fine-grained categorization with part-based image representation," *IEEE Trans. Image Process.*, vol. 25, no. 4, pp. 1713–1725, Apr. 2016.
- [59] Z. Xu, D. Tao, S. Huang, and Y. Zhang, "Friend or foe: Fine-grained categorization with weak supervision," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 135–146, Jan. 2017.
- [60] F. Zhou and Y. Lin. (2015). "Fine-grained image classification by exploring bipartite-graph labels." [Online]. Available: <https://arxiv.org/abs/1512.02665>
- [61] L. Zhang, Y. Yang, M. Wang, R. Hong, L. Nie, and X. Li, "Detecting densely distributed graph patterns for fine-grained image categorization," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 553–565, Feb. 2016.
- [62] H. Yao, S. Zhang, Y. Zhang, J. Li, and Q. Tian, "Coarse-to-fine description for fine-grained visual categorization," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4858–4872, Oct. 2016.
- [63] Y. Cui, F. Zhou, Y. Lin, and S. Belongie. (2015). "Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop." [Online]. Available: <https://arxiv.org/abs/1512.05227>
- [64] Z. Xu, S. Huang, Y. Zhang, and D. Tao, "Webly-supervised fine-grained visual categorization via deep domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1100–1113, May 2018.
- [65] S. Branson, G. V. Horn, S. Belongie, and P. Perona. (2014). "Bird species categorization using pose normalized deep convolutional nets." [Online]. Available: <https://arxiv.org/abs/1406.2952>
- [66] H. Zhang, T. Xu, M. Elhoseiny, X. Huang, S. Zhang, A. Elgammal, and D. Metaxas, "SPDA-CNN: Unifying semantic part detection and abstraction for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1143–1152.
- [67] D. Lin, X. Shen, C. Lu, and J. Jia, "Deep LAC: Deep localization, alignment and classification for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1666–1674.
- [68] C. Huang, H. Li, Y. Xie, Q. Wu, and B. Luo, "PBC: Polygon-based classifier for fine-grained categorization," *IEEE Trans. Multimedia*, vol. 19, no. 4, pp. 673–684, Apr. 2017.
- [69] H. Zhang, T. Xu, M. Elhoseiny, X. Huang, S. Zhang, A. Elgammal, and D. Metaxas, "SPDA-CNN: Unifying semantic part detection and abstraction for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1143–1152.
- [70] L. Xie, J. Wang, W. Lin, B. Zhang, and Q. Tian, "Towards reversal-invariant image representation," *Int. J. Comput. Vis.*, vol. 123, no. 2, pp. 226–250, Jun. 2017.
- [71] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "From generic to specific deep representations for visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 36–45.
- [72] Q. Qian, R. Jin, S. Zhu, and Y. Lin, "Fine-grained visual categorization via multi-stage metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2015, pp. 3716–3724.
- [73] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2014, pp. 806–813.
- [74] N. Murray and F. Perronnin, "Generalized max pooling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 2473–2480.
- [75] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 289–297.
- [76] X. Zhai, Y. Peng, and J. Xiao, "Learning cross-media joint representation with sparse and semisupervised regularization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 6, pp. 965–978, 2014.



Xiangteng He received the B.S. degree in computer science and technology from Nankai University in 2014. He is currently pursuing the Ph.D. degree with the Institute of Computer Science and Technology, Peking University. His current research interests include image analysis and deep learning.



Yuxin Peng received the Ph.D. degree in computer application from Peking University, Beijing, China, in 2003. After that, he was an Assistant Professor with the Institute of Computer Science and Technology (ICST), Peking University. He was promoted to an Associate Professor and a Professor at Peking University in 2005 and 2010, respectively, where he is currently a Professor with ICST and the Chief Scientist with the National Hi-Tech Research and Development Program of China (863 Program). In 2006, he was authorized by the Program for New Star in Science and Technology of Beijing and the Program for New Century Excellent Talents in University. He has authored over 130 papers in refereed international journals and conference proceedings, including IJCV, TIP, TMM, TCSVT, TCYB, TOMM, ACM MM, ICCV, CVPR, IJCAI, and AAAI. He has submitted 38 patent applications and received 23 of them. His current research interests mainly include cross-media analysis and reasoning, image and video analysis and retrieval, and computer vision. He led his team to participate in the TREC Video Retrieval Evaluation (TRECVID) many times. In 2009, his team received four first places on four sub-tasks of the High-Level Feature Extraction task and Search task. In 2012, his team received four first places on all four sub-tasks of the Instance Search (INS) task and the Known-Item Search task. In 2014, his team received the first place in the Interactive Instance Search task. His team also received first places in both the INS task of TRECVID from 2015 to 2018, respectively. He received the First Prize of the Beijing Science and Technology Award in 2016 (ranking first).