International Journal of Computer Vision (2019) 127:1235–1255
https://doi.org/10.1007/s11263-019-01176-2

Which and How Many Regions to Gaze: Focus Discriminative Regions for Fine-Grained Visual Categorization

Xiangteng He[1] · **Yuxin Peng**[1] · **Junjie Zhao**[1]

Received: 29 March 2018 / Accepted: 9 March 2019 / Published online: 23 March 2019

Abstract

Fine-grained visual categorization (FGVC) aims to discriminate similar subcategories that belong to the same superclass. Since the distinctions among similar subcategories are quite subtle and local, it is highly challenging to distinguish them from each other even for humans. So the localization of distinctions is essential for fine-grained visual categorization, and there are two pivotal problems: (1) **Which** regions are discriminative and representative to distinguish from other subcategories? (2) **How many** discriminative regions are necessary to achieve the best categorization performance? It is still difficult to address these two problems *adaptively* and *intelligently*. Artificial prior and experimental validation are widely used in existing mainstream methods to discover *which* and *how many* regions to gaze. However, their applications extremely restrict the *usability* and *scalability* of the methods. To address the above two problems, this paper proposes a **multi-scale and multi-granularity deep reinforcement learning approach (M2DRL)**, which learns multi-granularity discriminative region attention and multi-scale region-based feature representation. Its main contributions are as follows: (1) **Multi-granularity discriminative localization** is proposed to localize the distinctions via a two-stage deep reinforcement learning approach, which discovers the discriminative regions with multiple granularities in a hierarchical manner ("which problem"), and determines the number of discriminative regions in an automatic and adaptive manner ("how many problem"). (2) **Multi-scale representation learning** helps to localize regions in different scales as well as encode images in different scales, boosting the fine-grained visual categorization performance. (3) **Semantic reward function** is proposed to drive M2DRL to fully capture the salient and conceptual visual information, via jointly considering attention and category information in the reward function. It allows the deep reinforcement learning to localize the distinctions in a weakly supervised manner or even an unsupervised manner. (4) **Unsupervised discriminative localization** is further explored to avoid the heavy labor consumption of annotating, and extremely strengthen the *usability* and *scalability* of our M2DRL approach. Compared with state-of-the-art methods on two widely-used fine-grained visual categorization datasets, our M2DRL approach achieves the best categorization accuracy.

Keywords Fine-grained visual categorization · Deep reinforcement learning · Multi-granularity discriminative localization · Multi-scale representation learning · Unsupervised discriminative localization · Semantic reward

1 Introduction

Fine-grained visual categorization (FGVC) (Sfar et al. 2015; Branson et al. 2014b) aims to discriminate numerous similar subcategories that belong to the same basic category, such as the fine distinction of animals (Wah et al. 2011), plants (Nilsback and Zisserman 2008), cars (Krause et al. 2013) and aircraft models (Maji et al. 2013). It is different from the traditional basic-level visual categorization (Gonzalez-Garcia et al. 2018; Zhang et al. 2007), which aims to recognize the basic-level categories. As shown in Fig. 1, basic-level visual categorization only needs to recognize the image as

Communicated by Dr. S.-C. Zhu.

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

✉ Yuxin Peng
pengyuxin@pku.edu.cn

[1] Institute of Computer Science and Technology, Peking University, Beijing 100871, China

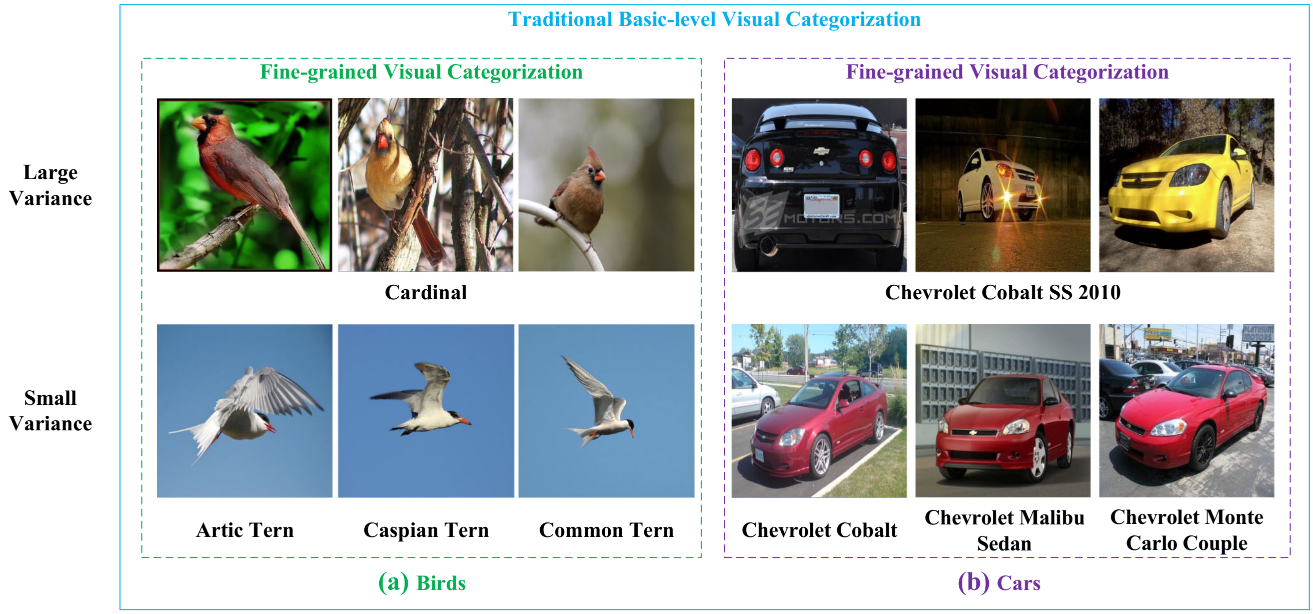


Fig. 1 Illustration of the difference between traditional basic-level visual categorization and fine-grained visual categorization, as well as the two characteristics that fine-grained subcategories have: *large variance* in the same subcategory as shown in the first line, and *small variance* among different subcategories as shown in the second line. Images in **a** birds and **b** cars are from CUB-200-2011 (Wah et al. 2011) and Cars-196 (Krause et al. 2013) datasets respectively

"Birds" or "Cars", rather than recognizing the image as the subcategories of "Artic Tern" or "Caspian Tern", which is the goal of fine-grained visual categorization. The fine-grained subcategories have two characteristics: (1) *Large variance in the same subcategory*. The instances in the same subcategory may look extremely different due to the different postures, different angles of view, or different developmental periods. As shown in Fig. 1, the images in the first row belong to the same subcategory (i.e. "Cardinal" or "Chevrolet Cobalt SS 2010"), but they look quite different, which are easily wrong recognized as different subcategories. (2) *Small variance among different subcategories*. The instances of different subcategories may look similarly in the global appearance, i.e. similar shape or color. As shown in Fig. 1, the images in the second row belong to different subcategories, but look quite the same, which are easily wrong recognized as the same subcategory. Therefore, the pivotal problem is to discover the discriminative regions to distinguish the subcategories, since these regions are the main distinctions among subcategories.

However, it is quite challenging to draw these distinctions even for humans, not to mention the computer. Researches indicate that humans prefer to gaze at the object (Neider and Zelinsky 2006). Eye movements always tend to direct towards the regions with high feature density (Tatler et al. 2006), texture (Itti and Koch 2001), and color contrast (Parkhurst et al. 2002), which can be considered as salient factors affecting object importance. For example, when recognizing an image, humans always first gaze at where the *object* is, and then gaze at those *parts* which and how many are distinct in the object, finally categorize the image, as shown in Fig. 2.

Inspired by the gazes when humans categorize an image, existing fine-grained visual categorization methods focus on localizing the discriminative regions in the image, such as the object and its parts. These regions contain the key distinctions from other subcategories and help to achieve better categorization performance. There are two pivotal problems in the discriminative localization: (1) *"Which problem"*: Which regions are discriminative and representative to distinguish from other subcategories? (2) *"How many problem"*: How many discriminative regions are necessary to achieve the best categorization performance?

Existing methods generally address these problems relying on the artificial prior (i.e. annotated information) or experimental validation, which extremely restrict their usability and scalability. Zhang et al. (2014) utilize R-CNN (Girshick et al. 2014) with geometric constraints to detect the object and its parts. Huang et al. (2016) utilize a fully convolutional neural network to localize the parts of the object. We can conclude that the above methods generally address the "*which*" and "*how many*" problems based on the annotated information, such as the ground truth bounding box of the object and part locations. However, not all the annotated information is significant for the categorization. For example, the "eye" part in CUB-200-2011 dataset (Wah et al. 2011) con-

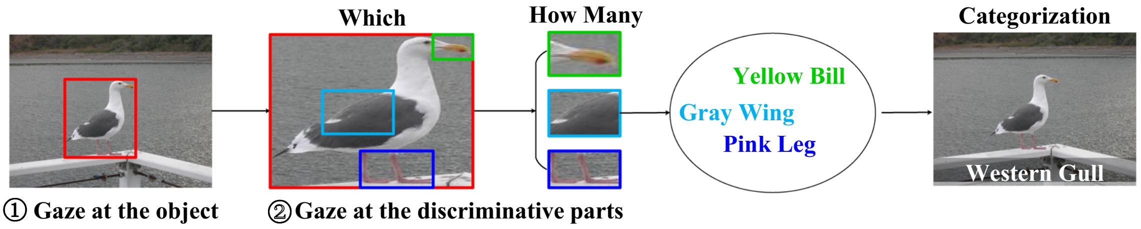


Fig. 2 Illustration of the gazes when humans recognize an image. First, gaze at where the object is, and then gaze at those parts which are distinct in the object, finally categorize the image

tains too little information to draw the distinctions among the similar subcategories, which is not necessary or helpful for categorization. The dependence on artificial prior makes the localization of discriminative regions subjective, and also needs to be customized for different fine-grained visual categorization tasks.

Therefore, researchers begin to study how to automatically localize the discriminative regions in the image without relying on the artificial prior. Xiao et al. (2015) propose a two-level attention method to select discriminative regions without using the object and part annotation, which selects two discriminative regions by experimental validation. Zhang et al. (2016c) incorporate deep convolutional filters for both parts selection and description. In this method, the number of discriminative regions changes for different datasets in order to achieve the best categorization accuracy. They empirically set the discriminative region number by experimental validation, which makes them difficult to scale to other tasks or domains. This increases the complexity and uncertainty of these methods. Besides, they set the same number of discriminative regions for all the subcategories, ignoring the fact that different subcategories, or even different images have different discriminative regions. This greatly reduces the performance of fine-grained visual categorization, and also makes the methods inflexible.

To simultaneously address the "*which problem*" and "*how many problem*" in an adaptive and intelligent manner, this paper proposes a *multi-scale and multi-granularity deep reinforcement learning approach (M2DRL)* for fine-grained visual categorization. It can automatically localize the discriminative regions in a hierarchical manner, as well as discover multiple discriminative regions with a single feed-forward pass by a tree-structured traversing scheme. Besides, due to the tree-structured traversing scheme and stop mechanism in reinforcement learning, our M2DRL approach can determine the discriminative region number in an adaptive manner. Therefore, the usability and scalability are guaranteed. Specifically, it adopts a multi-scale and multi-granularity representation learning architecture via deep reinforcement learning, which is driven by semantic reward function. Images with multiple scales are taken as inputs to the proposed architecture to exploit their comprehensive information. For each scale, a two-stage deep reinforcement learning (DRL) process is applied to exploit the variant granularity information of the discriminative regions in the image. The Stage-I, named *ObjectDRL*, removes the background noise in object alignment, and only reserves the foreground. The Stage-II, named *PartDRL*, further mines the compelling regions of the object, which are variant in numbers and granularities for different subcategories. They provide different but complementary visual information to boost the fine-grained representation learning as well as the categorization accuracy. In the learning process, semantic reward is utilized as tutorial information to guide the model to localize the regions with more salient and conceptual information.

To the best of our knowledge, our proposed M2DRL approach is the first work to research the fine-grained visual categorization task via deep reinforcement learning. The main contributions of our M2DRL approach can be summarized as follows:

(I) **Multi-granularity localization learning** is proposed to address the "which problem" and "how many problem" in an adaptive manner, instead of based on artificial prior or experimental validation in existing methods (Zhang et al. 2017; Xiao et al. 2015). We propose a two-stage deep reinforcement learning to hierarchically localize discriminative regions in different granularities for the "which problem", such as the object and its parts, and adaptively determine the number of discriminative regions for the "how many problem".

(II) **Multi-scale representation learning** is proposed to avoid the negative influence of variant scales of objects and its parts on fine-grained categorization, which boosts the categorization performance than only considering one scale. It contains two aspects: *Multiple scales of input images*. We apply two types of scales, where the larger one pays more attention to the detailed information, and the smaller one pays more attention to general information. *Multiple scales of discriminative parts*. We obtain multiple discriminative part proposals, which contain the same semantic parts in different scales and provide more information for categorization. Thus multi-scale representation learning boosts the categorization accuracy via localizing more discriminative regions in different scales, as well as jointly integrating information of different scales in the region-based feature representation.

(III) **Semantic reward function** is proposed to enhance the usability and scalability by avoiding the dependence on the annotations of the object and its parts in reinforcement learning. It applies semantic information into M2DRL to take the advantage of the salient and conceptual visual information in the image. It consists of two reward functions: *Attention-based reward function* focuses on localizing the regions with more salient information. *Category-based reward function* focuses on localizing the regions with more conceptual information. They jointly boost the performance of fine-grained localization as well as categorization.

(IV) **Unsupervised discriminative localization** is proposed to exploit the ability of M2DRL to localize discriminative regions in an unsupervised manner, instead of using manual annotations, such as image-level subcategory label, ground truth bounding box, or part locations. It avoids the heavy labor consumption of annotating, and extremely strengthens the usability and scalability of our M2DRL approach, which can facilitate the practical application of fine-grained visual categorization.

Comprehensive experiments on two widely-used fine-grained visual categorization datasets verify the effectiveness of our proposed M2DRL approach, which achieves the best categorization accuracy among state-of-the-art methods.

The rest of this paper is organized as follows: Sect. 2 briefly reviews related works: fine-grained visual categorization, deep reinforcement learning and its application in object detection. Section 3 presents our M2DRL approach in detail, and Sect. 4 introduces the experimental results as well as the experimental analyses. Finally, Sect. 5 presents the conclusion and future works of this paper.

2 Related Work

In this section, we introduce the works of three aspects related to this paper: fine-grained visual categorization, deep reinforcement learning and its application in object detection. Among these, fine-grained visual categorization is our focus, and deep reinforcement learning is the main starting point for our proposed M2DRL approach.

2.1 Fine-Grained Visual Categorization

Most existing fine-grained visual categorization methods follow the pipeline: first localize the object or its parts, and then learn region-based features for categorization. An intuitive idea is to directly utilize the annotations for the locations of object and its parts. For example, CUB-200-2011 dataset (Wah et al. 2011) provides the detailed annotations: a bounding box of the object, and 15 part locations. The bounding box of the object is used in the works of Chai et al. (2013) and Yang et al. (2012) to learn part detectors, and even part locations are used in previous works (Berg and Belhumeur 2013; Xie et al. 2013).

Since the annotations of the testing image are not available in the practical applications, some researchers use the ground truth bounding box or part locations only at training phase, and no knowledge of any annotations at testing phase. Object and part annotations are directly used in training phase to learn a strongly supervised deformable part-based model (Zhang et al. 2013) or fine-tune the pre-trained CNN model (Branson et al. 2014a). Krause et al. (2015) only use object annotation at training phase to learn the part detectors, and then localize the parts automatically in the testing phase. Above methods heavily rely on the time-consuming and labor consuming annotations, which limits their practicability.

Recently, there are some promising works attempting to learn the part detectors in a weakly supervised manner, which means that these works utilize neither object nor part annotations in both training and testing phases. These works make the practical application of fine-grained visual categorization possible. Simon and Rodner (2015) propose a neural activation constellations part model (NAC) to localize parts with constellation model. He and Peng (2017b) propose the part selection model with spatial constraints to localize more discriminative parts. The aforementioned methods mostly set the detector number according to the artificial prior and experimental validation, which is highly limited in flexibility and difficult for generalizing the methods to the other domains. Therefore, we attempt to automatically learn and mine which and how many discriminative regions really make sense to categorization via multi-scale and multi-granularity deep reinforcement learning.

2.2 Deep Reinforcement Learning

Reinforcement learning is the problem faced by an agent that must learn behavior through trial-and-error interactions with a dynamic environment (Kaelbling et al. 1996). As shown in Fig. 3, in the standard reinforcement learning, on each step of interaction, the agent conducts an action, and then the environment changes its state and feeds back a reward to guide the agent to obtain the long-term rewards. Recently, deep learning has achieved great successes in many domains due to its powerful automatic learning ability from a large scale data. DeepMind has pioneered the combination of deep learning and reinforcement learning, called deep reinforce-

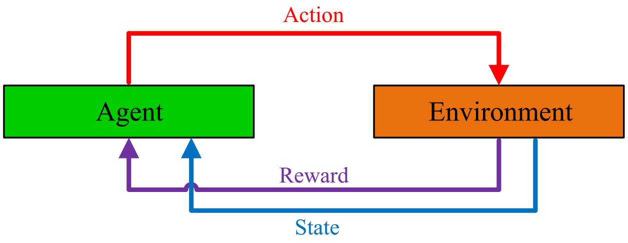


Fig. 3 Illustration of the standard reinforcement learning

ment learning, to achieve human-level performance across many challenging domains.

Mnih et al. (2015) propose the Deep Q-Network (DQN), which ignites the research of deep reinforcement learning. Its key idea is to use deep neural networks to represent the Q-network, and then predict total reward via the Q-network. Inspired by this work, more variations of DQN are proposed. Van Hasselt et al. (2016) propose Double DQN (D-DQN) to address the over-estimate problem in Q-learning, which can be generalized to work with large-scale function approximation. Schaul et al. (2015) propose prioritize experience replay, which replays important transitions more frequently and learn more efficiently. Wang et al. (2016c) propose the dueling network architecture to estimate state value function and state-dependent action advantage function, so that converges faster than Q-learning. As described in LeCun et al. (2015), systems combining deep learning and reinforcement learning are in their infancy, but they already outperform passive vision systems at classification tasks (Ba et al. 2014) and produce impressive results in learning to play many different video games (Mnih et al. 2015).

2.3 Object Detection Based on Deep Reinforcement Learning

Object detection is one of the most fundamental and challenging open problems in computer vision, which not only recognizes the objects but also localizes them in the images. It has achieved great progress due to the application of deep learning. Girshick et al. (2014) propose the method of regions with CNN features (R-CNN), which is a simple yet scalable detection framework and achieves state-of-the-art results in the Pascal and ImageNet benchmarks at that time.

Recently, deep reinforcement learning has been applied into object detection and achieves promising results. Caicedo and Lazebnik (2015) propose an active detection model for localizing objects in scenes. They model the problem of object detection with Markov decision process, and design nine localization actions to help the agent to land a tight bounding box that contains the target object. These actions are organized in four subsets: actions to move the box in the horizontal and vertical axes, actions to change scale, and actions to modify aspect ratio. The reward function is formulated using the Intersection-over-Union (IoU) between the target object and the predicted box at any step. The proposed model obtains the best detection performance among systems that do not use object proposals for object localization. Jie et al. (2016) propose an effective tree-structured reinforcement learning (Tree-RL) approach to sequentially search for objects by fully exploiting both the current observation and historical search paths. Tree-RL has two advantages: (1) Localize multiple objects via a tree-structured traversing scheme in a single feed-forward pass. (2) Localize objects in different scales, which improves its scalability. Mathe et al. (2016) propose a principled sequential models with reinforcement learning, which accumulate evidence collected at a small set of image locations in order to detect objects effectively. Zhao et al. (2017b) propose a convolutional neural network model of visual attention for image classification. The attention is obtained via reinforcement learning, and used to select useful key regions in the image to boost the categorization accuracy. Inspired by these works, we propose multi-scale and multi-granularity deep reinforcement learning to localize discriminative regions to further boost the fine-grained visual categorization accuracy.

3 Multi-scale and Multi-granularity Deep Reinforcement Learning Approach (M2DRL)

In this section, we present the proposed *multi-scale and multi-granularity deep reinforcement learning approach (M2DRL)* for fine-grained visual categorization. It consists of *multi-granularity discriminative localization (MgDL)* and *multi-scale representation learning (MsRL)*, which learn discriminative region attention in multiple granularities and region-based feature representation in multiple scales via deep reinforcement learning. The framework of our M2DRL approach is shown in Fig. 4.

Aiming at utilizing the comprehensive information of the image, M2DRL takes multiple scales of the images as inputs to the architecture. For each scale, a two-stage deep reinforcement learning is applied to exploit the granularity information of the discriminative region attention in the image. The Stage-I deep reinforcement learning, named *ObjectDRL*, aligns object via removing the background noise, and reserves the foreground. The Stage-II deep reinforcement learning, named *PartDRL*, further mines the compelling regions of the object, which are variant in numbers and granularities for each subcategory and each image. In the learning process, semantic reward function is proposed to guide the action of agent to obtain the long-term rewards.

The remainder of this section is organized as follows. First, we briefly introduce the problem formulation in Sect. 3.1.

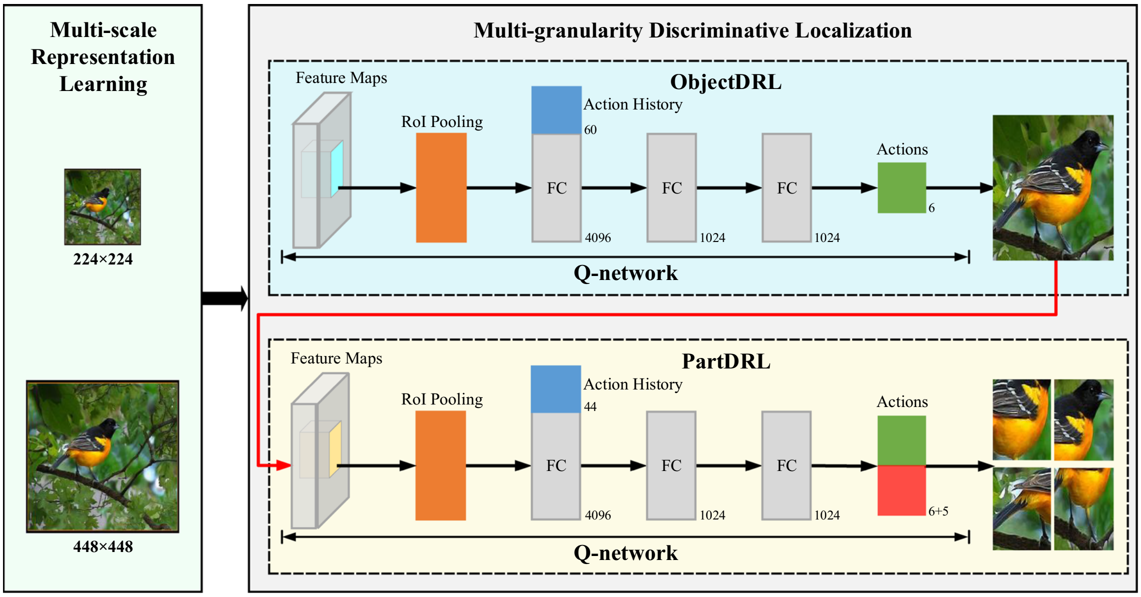


Fig. 4 An overview of the proposed M2DRL approach, which consists of multi-granularity discriminative localization and multi-scale representation learning, which learns multi-granularity discriminative region attention and multi-scale region-based feature representation with deep reinforcement learning. Multi-granularity discriminative localization adopts a two-stage deep reinforcement learning architecture: ObjectDRL and PartDRL

Second, we describe the details of multi-granularity discriminative localization in Sect. 3.2. Third, Sect. 3.5 details the multi-scale representation learning. Finally, prediction pipeline is described in Sect. 3.6.

3.1 Problem Formulation

For a given image I, we formulate the discriminative localization as the problem of maximizing a confidence score function $f_c : B \to \mathbb{B}$ over the set of image region candidates B:

$$b^* = arg \max_{b \in B} f_c(b) \tag{1}$$

We address the problem via Markov decision process (MDP), which is well suitable for modeling the discrete time sequential decision making process. The MDP consists of a set of actions A, a set of states S, and a reward function R. They are defined differently for ObjectDRL and PartDRL, and introduced in detail as follows.

3.2 Multi-granularity Discriminative Localization

The given image I is considered as the environment, and the goal of agent is to localize discriminative regions in the image. The agent localizes a region at each step by conducting one action of A. Then, the state of the agent changes based on the conducted action, which contains the information of the current localized region and the past action history. Simultaneously, a corresponding reward will be fed back to the agent at the training phase, which may be positive or negative. The reward guide the agent to obtain the long-term rewards for better optimization. In the following paragraphs, the details of actions, states and reward are described.

3.2.1 Discriminative Localization Actions

Inspired by Tree-RL (Jie et al. 2016), we define the discriminative localization actions A as two action groups according to their different effects, as shown in Fig. 5.

(I) Action Group 1

Action group 1 consists of five cropping actions to localize discriminative region, and one special action to terminate the localization process. Each cropping action crops the current region to a certain sub-region with the cropping ratio α, corresponding to cropping the current region to the top left corner, bottom left corner, top right corner, bottom right corner and the center respectively, where $\alpha \in [0, 1]$. The cropping actions can localize the regions with different scales, which guarantees the scalability of the localization.

(II) Action Group 2

Action group 2 consists of four local translation actions and one action to terminate the localization process. Each local translation action moves the region downward, upward, towards the right, and towards the left respectively by β times

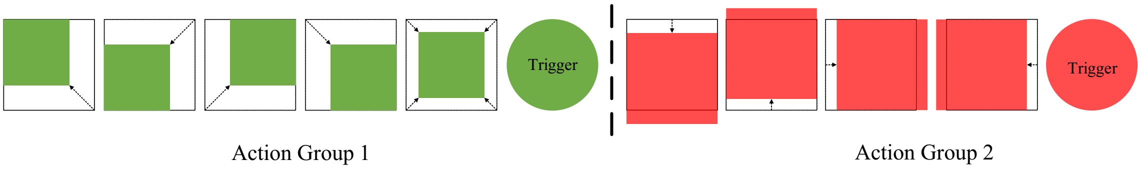


Fig. 5 An overview of the discriminative localization actions. It consists of two action groups, one is cropping action group with six actions, as shown in green color, and the other is local translation action group with five actions, as shown in red color (Color figure online)

of the current region size, where $\beta \in [0, 1]$. The local translation actions can drive the agent to modify the localization process as well as discover different discriminative regions.

It is noted that ObjectDRL and PartDRL adopt different action groups. Only action group 1 is adopted in ObjectDRL. In PartDRL, we hope that the agent can localize multiple regions with different characteristics, which may be the distinctions from other similar subcategories and contribute to discriminating the similar subcategories. So, we follow Tree-RL (Jie et al. 2016) to adopt a tree-structured search scheme in PartDRL, which has two branches: one only conducts actions in group 1, and the other only conducts actions in group 2, as shown in Fig. 6.

3.2.2 States

Once an action is conducted on the current region, the content of the region will be changed deterministically, which means that the state is changed. At action step t, the state of the agent is represented as $S_t = (v_t, h_t)$, where v_t denotes the feature vector of the current localized region in the image, and h_t denotes the history vector of the past conducted actions. The following paragraphs introduce the details of v_t and h_t.

The feature vector v_t is extracted from the CNN model, which is pre-trained on the ImageNet dataset (Deng et al. 2009). In our experiment, feature maps are extracted from the layer “conv5_3” of the 16-layer VGGNet (Simonyan and Zisserman 2014) as the initial features, followed by a fully-connected layer to generate the final 4096-dimensional feature vector. Inspired by Fast R-CNN (Girshick 2015), RoI Pooling layer is applied to accelerate feature extraction for each localized region.

The history vector $h_t = \{H_1, H_2, \ldots, H_N\}$ is a binary vector, and indicates the past conducted actions, where N denotes a pre-defined maximal action execution number per image, and $N = N_{step}$ in ObjectDRL as well as $N = N_{level}$ in PartDRL. H_i donates a one-hot encoding of the ith conducted action, whose dimension is 6 in ObjectDRL and 11 in PartDRL, corresponding to the number of actions respectively. It is noted that the elements after tth element are all zero vectors at action step t.

3.2.3 Semantic Reward Function

The reward function R reflects the effect of the conducted action to the localization accuracy, where a positive reward means that the conducted action is a good decision to make the localization more accurate, while a negative reward means a non-ideal decision. We propose a semantic reward function to fully learn the discriminative and conceptual visual information via considering both attention-based reward and category-based reward.

(I) Attention-based Reward Function

Intersection-over-Union (IoU) between the current localized region and the ground truth bounding box of target discriminative region, e.g. object and its parts, is widely used to measure the effect of the conducted action for localization (Jie et al. 2016). The reward function $RA_a(s, s')$ denotes the reward received when the state of the agent changes from s to s' by conducting an action a, and its definition is as follows:

$$RA_a(s, s') = sign(IoU(b', g) - IoU(b, g)) \tag{2}$$

where b denotes the current region, b' denotes the region obtained by conducting action a on the current region b, g denotes the ground truth bounding box. $IoU(b, g) = area(b \cap g)/area(b \cup g)$, similar is $IoU(b', g)$. The above reward function $RA_a(s, s')$ relies on the ground truth bounding box, whose labeling is expensive.

Therefore, we propose a new reward function based on the attention information, which avoids the heavy labor consumption for labeling. Recent works (Zhou et al. 2015, 2016) have shown that the neurons in the convolutional layers actually have the ability to localize the object without supervision of object annotation. Therefore, we first extract the attention map M of the image, which indicates the representative and significant regions for CNN to identify the subcategory of image.

Given an image I, the activation of neuron u in the last convolutional layer at spatial location (x, y) is defined as $f_u(x, y)$. The attention value at spatial location (x, y) is computed as follows:

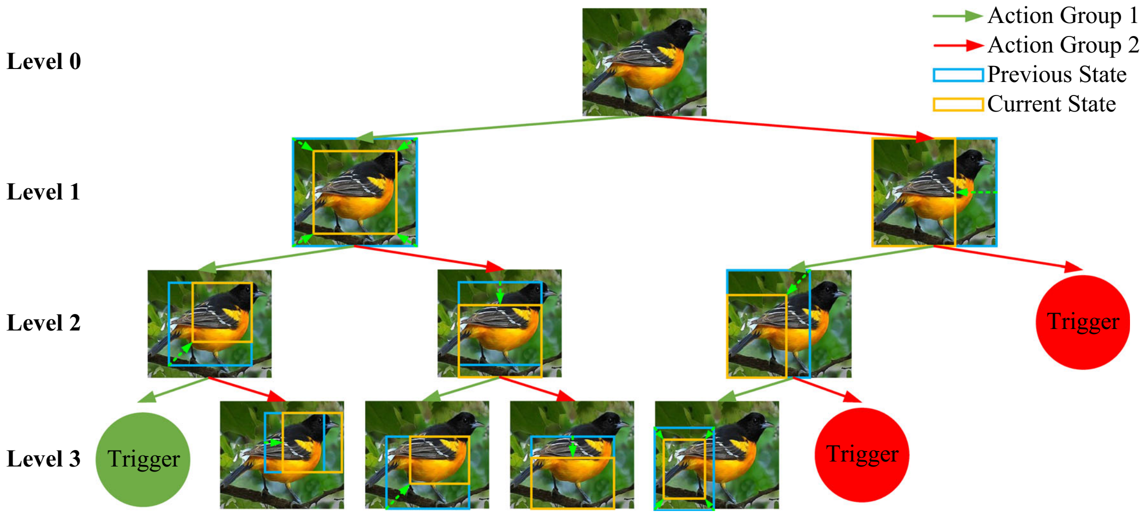


Fig. 6 Illustration of the tree structure scheme. At each node, the left branch conducts action group 1 (as shown in green color), and the right branch conducts action group 2 (as shown in red color). For each image in the figure, the blue rectangle denotes the previous state, and the yellow rectangle denotes the current state (Color figure online)

$$M(x, y) = \frac{1}{|u|} \sum_{u} f_u(x, y) \tag{3}$$

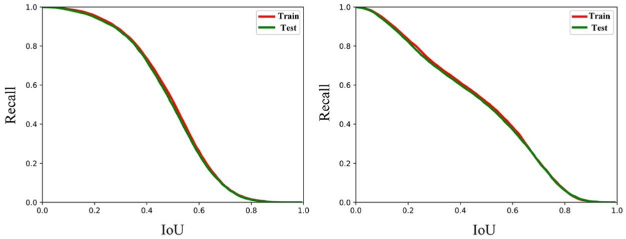
where $M(x, y)$ directly indicates the importance of activation at spatial location (x, y) for categorizing the image. As attention map has significant impact on the final categorization performance, we show the effectiveness of localization. The curves of recall versus IoU overlap ratio are shown in Fig. 7. For CUB-200-2011 dataset, the area under curve (AUC) values of training and testing sets are 0.494 and 0.487 respectively. For Cars-196 dataset, the AUC values of training and testing sets are 0.478 and 0.471 respectively. Considering that CAM is not trained with ground truth bounding box, the localization results are promising.

It is noted that we design different attention-based reward functions for ObjectDRL and PartDRL, which are presented as follows.

Attention-based reward function for ObjectDRL. We perform binarization operation on the attention map with OTSU algorithm (Otsu 1979), and take the bounding box that covers the largest connected area as g_{atten}. Therefore, the attention-based reward function is defined as follows:

$$RA_a(s, s') = sign(IoU(b', g_{atten}) - IoU(b, g_{atten})) \tag{4}$$

The attention-based reward function fully utilizes the attention information of the image, without relying on the ground truth bounding box, and guides the agent to localize the region with the highest saliency, corresponding to the region of the target object.



(a) CUB-200-2011 **(b)** Cars-196

Fig. 7 Recall versus IoU overlap ratio of CAM (Zhou et al. 2016) on CUB-200-2011 and Cars-196 datasets

Attention-based reward function for PartDRL. We define the reward function RA_a as follows:

$$RA_a(s, s') = sign(Mean(b') - Mean(b)) \tag{5}$$

where function $Mean(\cdot)$ denotes the mean value of the attention map of the current region. Through the tree-structured search scheme and the attention-based reward function, we can localize different regions of the object, which can boost the diversity of the feature representation.

(II) Category-based Reward Function

As is known to all, the category label directly provides the conceptual information. It can guide the agent to localize the region that is actually helpful for the categorization. Therefore, we propose the category-based reward function as follows:

$$RC_a(s, s') = sign(P_c(b') - P_c(b)) \quad (6)$$

where $P_c(\cdot)$ indicates the predicted score of the corresponding region that is categorized as subcategory c, and c is the annotated image-level subcategory label.

The semantic reward function R jointly considers the attention and category information, and its definition is as follows:

$$R_a(s, s') = RA_a(s, s') + RC_a(s, s') \quad (7)$$

It is noted that we define a different reward function for the trigger following (Caicedo and Lazebnik 2015), which is to indicate that the current region contains the target object or discriminative region. In ObjectDRL, its definition is as follows:

$$RO_{trigger}(s, s') = \begin{cases} +\eta, & if\ IoU(b, g_{atten}) \geq \tau \\ -\eta, & otherwise \end{cases} \quad (8)$$

where η is the trigger reward, and the trigger will be conducted when the IoU value is over the threshold τ. η and τ are set to 3 and 0.6 respectively in our experiments. In PartDRL, its definition is as follows:

$$RO_{trigger}(s, s') = \begin{cases} +\eta, & if\ Mean(b) \geq \tau \\ -\eta, & otherwise \end{cases} \quad (9)$$

where η and τ are set to 3 and 0.5 respectively in our experiments.

Through the ObjectDRL, the object region is obtained. To further represent the object with more local and discriminative information, we learn to mine finer-granularity regions on the localized object in the PartDRL.

3.3 Q-Learning for Discriminative Localization

We apply reinforcement learning to learn the discriminative policy of maximizing the sum of the received rewards of running an episode starting from the original image. Deep Q-network (Mnih et al. 2015) is applied to solve the problem of reinforcement learning. The detailed architecture of our Q-network is shown in Fig. 8. There are three streams in our Q-network, where the first one is for action prediction, the second one is for attention-based reward calculation, and the third one is for category-based reward calculation. Specifically, the feature of each proposal object or part bounding box is extracted by RoI Pooling to reduce the cost of computation, and then fed into the streams of category-based reward calculation (i.e. $P_c(\cdot)$ in Eq. (6) is computed as the softmax vector output from the third stream of the Q-network). Before RoI Pooling, the feature maps are used to generate the attention map as described in Sect. 3.2.3. In this way, the attention-based and category-based rewards can be calculated for guiding the action prediction. For the stream of action prediction, we concatenate the feature vector and the action history vector, and then feed them into the fully-connected layers. Finally, mean squared error (MSE) is used to estimate the predicted values of the localization actions. Different from (Jie et al. 2016; Caicedo and Lazebnik 2015), we apply the fine-tuned CNN as the feature extractor at each action step. The CNN is first pre-trained on the ImageNet dataset (Deng et al. 2009), and then fine-tuned on the specific fine-grained dataset, such as CUB-200-2011 dataset (Wah et al. 2011). It is because that fine-tuned CNN can obtain a better attention map for each image, and extract more powerful and discriminative features. At training phase, the parameters of Q-network are updated by the agent running multiple episodes, whose behavior is ϵ-greedy (Sutton and Barto 1998). At each step, the agent randomly selects an action from the whole action set with probability ϵ, and selects the best action according to the learned Q-network in action group 1 for ObjectDRL with probability $1-\epsilon$, a random action from the two best actions in action group 1 and 2 respectively for PartDRL with probability $1-\epsilon$.

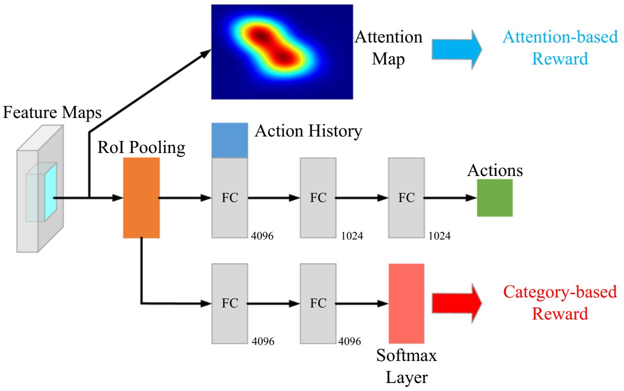


Fig. 8 Architecture of Q-network

3.4 Unsupervised Discriminative Localization

In this subsection, we explore the discriminative localization in an unsupervised manner, without using any annotations. From Sect. 3.2.1, we know that attention map can tell which region is discriminative and significant for categorization. Besides, we know that the CNN pre-trained on ImageNet dataset has good generalization. Considering the attention map extracted from pre-trained CNN has bad ability to reflect the region of object but corresponds to some discriminative local regions, we only explore the PartDRL in the unsupervised manner.

Specifically, in unsupervised discriminative localization, localization actions and states are the same as PartDRL, which are described in Sects. 3.2.2 and 3.2.1. To avoid using the annotation information, we design the semantic reward function RU with attention-based reward, and its definition is the same as RA_a in PartDRL.

$$RU(s, s') = sign(Mean(b') - Mean(b)) \quad (10)$$

But the CNN, which is used to extract the attention map for each image, is not fine-tuned on the specific fine-grained visual categorization dataset. It is a pre-trained CNN on the ImageNet dataset, which is widely used in the computer vision tasks. So fine-grained subcategory label is not used. Besides, for the Q-learning in discriminative localization, we initialize the parameters of convolutional layers with the pre-trained CNN, and initialize the parameters of each fully-connected layer from a zero-mean normal distribution with standard deviation 0.01. Its training process is same as PartDRL as described in Sect. 3.3, which is guided by the semantic reward function RU. Thus, there is no annotation used in the whole learning process.

3.5 Multi-scale Representation Learning

After multi-granularity discriminative localization, we obtain a variable number of discriminative regions for each image. These discriminative regions, i.e. both object and its parts, are in multiple scales, in which some regions are in small scale. The problem of small scale causes that these regions are difficult to localize, and have little detailed information for CNNs to generate good feature representation. To address this problem, we apply multi-scale representation learning. It has two aspects: First, we crop the original images into different scales, and take them as the inputs of multi-granularity discriminative localization. Different scales can provide different but complementary information, where large scale pays more attention to the detailed information, such as fine texture, and small scale pays more attention to general information, such as holistic shape. Different scales make our M2DRL approach obtain more discriminative regions and extract better feature representation. In our experiments, we choose the scales of 224 × 224 and 448 × 448. Second, in PartDRL, we not only utilize the localized regions on the leaf nodes, as shown in Fig. 6, but also utilize the other regions on the tree except the region on the root node. Since regions in different levels of tree structure are in different scales, they provide more useful information for categorization.

3.6 Final Prediction

For a given image I, no more than $N_{step} - 1$ regions are obtained that correspond to the target object in ObjectDRL, and no more than $2^{N_{level}} - 2$ regions are obtained that correspond to the discriminative parts of the object in PartDRL. Each region is fed to the fine-tuned CNN, i.e. 19-layer VGGNet with batch normalization (Ioffe and Szegedy 2015), and received its prediction vector. For the regions obtained by ObjectDRL, we select the region with highest predicted score, denoted as $\max(SO)$. For the regions obtained by PartDRL, we select the region with highest predicted score for each level of the tree structure, denoted as $\max(SP_l)$. Finally, the final prediction is obtained by fusing the above predictions via the following equation:

$$Score = \lambda \max(SO) + (1 - \lambda)\frac{1}{N_{level}}\sum_{l=1}^{N_{level}} \max(SP_l) \quad (11)$$

where λ is selected via k-fold cross-validation method.

4 Experiments

In this section, we conduct experiments on two widely-used datasets for fine-grained visual categorization: CUB-200-2011 (Wah et al. 2011) and Cars-196 (Krause et al. 2013), taking more than fifteen state-of-the-art methods for comparison to verify the effectiveness of our proposed M2DRL approach. Besides, comprehensive experimental analyses are presented including baseline experiments, localization analyses, as well as unsupervised discriminative localization to verify the contribution of each component in our proposed M2DRL approach.

4.1 Datasets

Here we briefly introduce two widely-used fine-grained visual categorization datasets adopted in the experiments, including CUB-200-2011 and Cars-196 datasets. We can observe that images in the same basic-level category are very similar in global appearance, which make the fine-grained categorization highly challenging. Each dataset is divided into two subsets, namely training set and testing set.

(I) **CUB-200-2011** (Wah et al. 2011)[1]: It is the most widely-used dataset for fine-grained visual categorization, and contains 11,788 images of 200 different bird subcategories, which is divided as follows: 5994 images as training set and 5794 images as testing set. For each subcategory, about 30 images are selected for training and 11–30 images for testing. Each image has detailed annotations as follows: an image-level subcategory label, a bounding box of the object, 15 part locations and 312 binary attributes. All attributes are visual in nature, pertaining to color, pattern, or shape of a particular part. In

[1] http://www.vision.caltech.edu/visipedia/CUB-200-2011.html.

our experiments, only image-level subcategory label is utilized in the training phase.

(II) **Cars-196** (Krause et al. 2013)[2]: It contains 16,185 images of 196 car subcategories, and is divided as follows: 8144 images as training set and 8041 images as testing set. For each subcategory, 24–84 images are selected for training and 24–83 images for testing. Each image is annotated with an image-level subcategory label and a bounding box of the object. The same as CUB-200-2011, only image-level subcategory label is utilized in the training phase.

4.2 Evaluation Metric

Here we introduce the evaluation metrics used in our experiments to verify the effectiveness of our proposed M2DRL approach, namely *accuracy* and *Intersection-over-Union*.

(I) **Accuracy** is adopted as the evaluation metric to evaluate the categorization accuracy of our proposed M2DRL approach compared with state-of-the-art methods, which is widely used for evaluating the performance of fine-grained visual categorization (Zhang et al. 2014, 2016c, e). Its definition is as follows:

$$Accuracy = \frac{|I_r|}{|I|} \tag{12}$$

where $|I|$ means the number of images in testing set, and $|I_r|$ counts the number of images which are correctly categorized in testing set.

(II) **Intersection-over-Union (IoU)** (Everingham et al. 2015) is adopted to evaluate the overlap between the predicted bounding box of discriminative region and the target region, and its definition is as follow:

$$IoU = \frac{area(b \cap g)}{area(b \cup g)} \tag{13}$$

where b denotes the predicted bounding box of discriminative region, g denotes the ground truth bounding box of the target region, such as the object, $b \cap g$ denotes the intersection of the predicted and ground truth bounding boxes, and $b \cup g$ denotes their union.

4.3 Implementation Details

We describe the details of M2DRL in the following five aspects: (1) For actions, the ratios of cropping action and local translation actions are set to 0.9 and 0.1 respectively. To make a trade-off between localization speed and accuracy, we set the maximal action execution number $N_{step} = 10$ in ObjectDRL. The value of N_{step} reserves the same for different datasets. In ObjectDRL, cropping actions are conducted by the agent, which crop the current region to a certain subregion with the cropping ratio 0.9. If the agent conducts N_{step} steps, the region at the N_{step}th step is $0.9^9 = 0.387$ of the original images. While the object in the image is generally not smaller than 0.387 of the original images. It is enough for localization. Similarly in PartDRL, the level of tree structure, $N_{level} = 4$ is enough for localizing the discriminative regions of the object. (2) For semantic reward function, the trigger reward η and threshold τ are set to 3 and 0.5 respectively. (3) For Q-learning, the architecture of Q-network is shown in Fig. 8. The region features are computed via RoI Pooling layer with the shape of $512 \times 7 \times 7$, and then concatenated with the action history vector to be fed into fully-connection layers. We initialize the parameters of convolutional layers with the fine-tuned CNN, and initialize the parameters of each fully-connected layer from a zero-mean normal distribution with standard deviation 0.01. In the fine-tuning phase, all layers are updated. The fine-tuned CNN is trained with the original whole images and the image patches generated by the data augmentation, which is to select relevant image patches by object-level attention in Xiao et al. (2015). In the training phase, the parameter ϵ starts with 1.0 and decreases by 0.1 for each epoch. It is finally fixed to 0.1 after the first 10 epochs to let the agent focus on learning from experiences generated by its own model. The optimization process of Q-network follows Tree-RL (Jie et al. 2016). We use 16-layer VGGNet (Simonyan and Zisserman 2014) as the CNN model. We follow CAM (Zhou et al. 2016) to modify the architecture of VGG network. Specifically, 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. In this way, the network can identify the discriminative regions easily in a single forward pass. (4) For multiple scales, we crop the original images into scales of 224 and 448×448. (5) For the training of ObjectDRL and PartDRL, they are trained in a separate and parallel manner. First, we initialize the parameters of convolutional layers with the fine-tuned CNN, and initialize the parameters of each fully-connected layer from a zero-mean normal distribution with standard deviation 0.01. Then we train them separately and parallel, which accelerates the training speed. Thus we obtain the model of M2DRL.

4.4 Comparisons with State-of-the-Art Methods

This subsection presents the experimental results and analyses of our M2DRL approach compared with the state-of-the-art methods on CUB-200-2011 and Cars-196 datasets, as shown in Tables 1 and 2. For fair comparison, the anno-

[2] http://ai.stanford.edu/~jkrause/cars/car_dataset.html.

Table 1 The results of categorization accuracy for our proposed M2DRL approach and the state-of-the-art methods on CUB-200-2011 dataset (Wah et al. 2011)

Method	Training annotation		Testing annotation		Accuracy (%)
	Object	Parts	Object	Parts	
Our M2DRL approach					87.21
OPAM (Peng et al. 2018)					85.83
CVL (He and Peng 2017a)					85.55
RA-CNN (Fu et al. 2017)					85.30
HCA (Cai et al. 2017)					85.30
PNA (Zhang et al. 2017)					84.70
TSC (He and Peng 2017b)					84.69
FOAF (Zhang et al. 2016d)					84.63
PD (Zhang et al. 2016c)					84.54
LRBP (Kong and Fowlkes 2017)					84.21
STN (Jaderberg et al. 2015)					84.10
Bilinear-CNN (Lin et al. 2015b)					84.10
Multi-grained (Wang et al. 2015)					81.70
NAC (Simon and Rodner 2015)					81.01
PIR (Zhang et al. 2016e)					79.34
TL Atten (Xiao et al. 2015)					77.90
MIL (Xu et al. 2017)					77.40
VGG-BGLm (Zhou and Lin 2016)					75.90
InterActive (Xie et al. 2016)					75.62
Dense Graph Mining (Zhang et al. 2016b)					60.19
Coarse-to-Fine (Yao et al. 2016)	√				82.50
Coarse-to-Fine (Yao et al. 2016)	√		√		82.90
PG Alignment (Krause et al. 2015)	√		√		82.80
VGG-BGLm (Zhou and Lin 2016)	√		√		80.40
Triplet-A (64) (Cui et al. 2016)	√		√		80.70
Triplet-M (64) (Cui et al. 2016)	√		√		79.30
Webly-supervised (Xu et al. 2018)	√	√			78.60
PN-CNN (Branson et al. 2014a)	√	√			75.70
Part-based R-CNN (Zhang et al. 2014)	√	√			73.50
SPDA-CNN (Zhang et al. 2016a)	√	√	√		85.14
Deep LAC (Lin et al. 2015a)	√	√	√		84.10
SPDA-CNN (Zhang et al. 2016a)	√	√	√		81.01
PS-CNN (Huang et al. 2016)	√	√	√		76.20
PN-CNN (Branson et al. 2014a)	√	√	√	√	85.40
Part-based R-CNN (Zhang et al. 2014)	√	√	√	√	76.37
POOF (Berg and Belhumeur 2013)	√	√	√	√	73.30
HPM (Xie et al. 2013)	√	√	√	√	66.35

"Object" and "Parts" denote the annotation utilized in the training phase and testing phase of our proposed M2DRL approach as well as the compared methods, where "Object" denotes the ground truth bounding box of the object and "Parts" denotes the annotated part locations. It is noted that neither "Object" nor "Parts" is used in our proposed M2DRL approach

tations utilized in the training and testing phases are listed, where "Object" denotes the ground truth bounding box of the object and "Parts" denotes the annotated part locations. If the column is empty, it means that the annotation is not used. It is noted that neither the ground truth bounding box nor part locations are used in our proposed M2DRL approach, only image-level subcategory label is used.

On CUB-200-2011 dataset, our approach achieves the best categorization accuracy among all the methods, as shown in Table 1. The best result of compared methods is achieved

Table 2 The results of categorization accuracy for our proposed M2DRL approach and the state-of-the-art methods on Cars-196 dataset (Krause et al. 2013)

Method	Training annotation		Testing annotation		Accuracy (%)
	Object	Parts	Object	Parts	
Our M2DRL approach					93.25
RA-CNN (Fu et al. 2017)					92.50
OPAM (Peng et al. 2018)					92.19
Bilinear-CNN (Lin et al. 2015b)					91.30
TL Atten (Xiao et al. 2015)					88.63
DVAN (Zhao et al. 2017a)					87.10
FT-HAR-CNN (Xie et al. 2015)					86.30
HAR-CNN (Xie et al. 2015)					80.80
PG Alignment (Krause et al. 2015)	√				92.60
ELLF (Krause et al. 2014)	√				73.90
R-CNN (Girshick et al. 2014)	√				57.40
PG Alignment (Krause et al. 2015)	√		√		92.80
BoT(CNN With Geo) (Wang et al. 2016a)	√		√		92.50
DPL-CNN (Wang et al. 2016b)	√		√		92.30
VGG-BGLm (Zhou and Lin 2016)	√		√		90.50
LLC (Wang et al. 2010)	√		√		69.50
BB-3D-G (Krause et al. 2013)	√		√		67.60

“Object” and “Parts” denote the annotation utilized in the training phase and testing phase of our proposed M2DRL approach as well as the compared methods, where “Object” denotes the ground truth bounding box of the object and “Parts” denotes the annotated part locations. It is noted that neither “Object” nor “Parts” is used in our proposed M2DRL approach

by OPAM (Peng et al. 2018), which integrates object-level attention and part-level attention, the number of discriminative regions is set to 3, including one localized object and two discriminative parts. Our M2DRL approach brings a 1.38% categorization accuracy improvement. CVL (He and Peng 2017a), which jointly models visual and textual information, uses both the original images and the object. Besides, it also utilizes extern textural descriptions of the image in the training phase. However, our M2DRL approach still outperforms it by 1.66%. RA-CNN (Fu et al. 2017) achieves the categorization accuracy of 85.30%, which utilizes three regions in different scales. While it only achieves the categorization accuracy of 84.70% with two regions in different scales, which is 0.60% lower than using three regions. PNA trains 11 part detectors to localize the discriminative regions. TSC (He and Peng 2017b) localizes three discriminative regions to achieve the better categorization accuracy, including one object and two discriminative parts. From the above analyses, we can see that the number of discriminative regions is significant for the categorization accuracy, but it is generally set due to the artificial prior or experimental validation, which leads that method should be customized for different tasks, as well as restricts the usability and scalability of fine-grained visual categorization.

Our M2DRL approach tries to address this problem, via adaptively localizing and determining “which” and “how many” regions are discriminative in the image, boosting the categorization accuracy as well as enhancing the usability and scalability of fine-grained visual categorization. The number of discriminative regions is set adaptively in the process of multi-granularity discriminative localization. The number is different not only for each subcategory but also for each image, as shown in Fig. 12, which will be analyzed in Sect. 4.5.2. Our M2DRL achieves the best categorization accuracy based on the adaptively localized discriminative regions.

Even compared with the methods which utilize the ground truth bounding box in training phase or even in testing phase, our M2DRL approach achieves better categorization accuracy. Furthermore, compared with the methods that utilize the part locations, our M2DRL approach still achieves better categorization accuracy.

Besides, the results of categorization accuracy on Cars-196 dataset are shown in Table 2. The trend is similar as CUB-200-2011 dataset, our proposed M2DRL approach achieves the best categorization accuracy among state-of-the-art methods, and brings a 0.75% improvement than the best result of compared methods, which verifies the effectiveness of our proposed M2DRL approach.

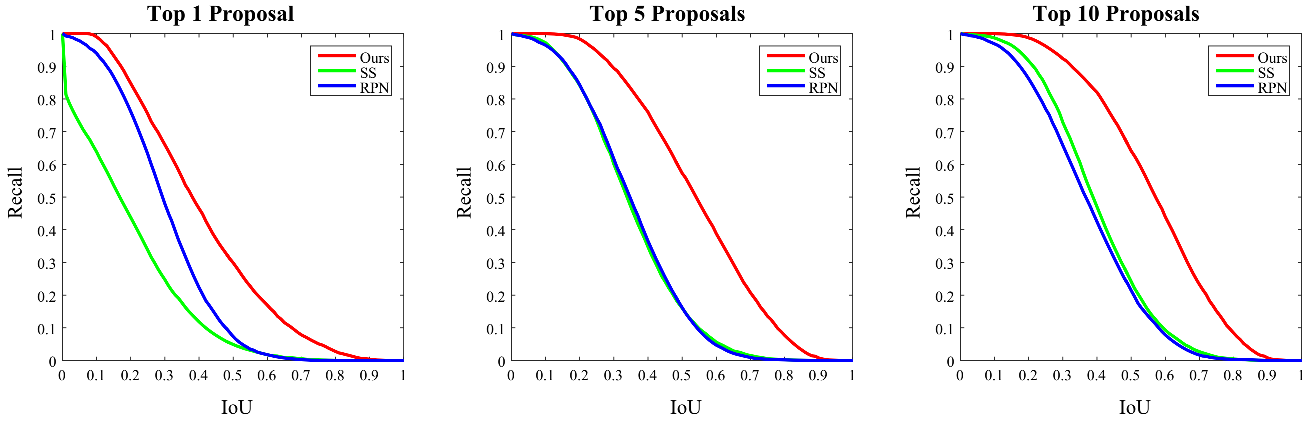


Fig. 9 Recall versus IoU overlap ratio on the CUB-200-2011 dataset. Here we show the results of using Top 1, 5 and 10 proposals, as well as compare with selective search (SS) (Uijlings et al. 2013), and region proposal network (RPN) (Ren et al. 2015)

4.5 Effectiveness of Discriminative Localization

In our M2DRL approach, ObjectDRL and PartDRL are conducted in sequence to localize the discriminative regions in two granularities: object and its parts. In this subsection, we discuss the effectiveness of localization on the CUB-200-2011 dataset with the scale of 224 × 224.

4.5.1 Effectiveness of ObjectDRL

ObjectDRL distinguishes the object from the background, and represents the features of the global appearance. Here we compute the recall of proposals at different IoU overlap ratios with the ground truth bounding boxes, the same as Ren et al. (2015). Figure 9 shows the results of using Top 1, 5, 10 proposals. We compare with selective search (SS) (Uijlings et al. 2013), and region proposal network (RPN) (Ren et al. 2015). The training of RPN is the same with (Ren et al. 2015). We apply nine anchors with three scales and three aspect ratios as Faster R-CNN. For training RPN, a binary class label of being an object or not is assigned to each anchor, which depends on the IoU overlap with the ground truth bounding box. Top N proposals are selected based on the confidence scores generated by these methods. Our M2DRL approach can generate 10 proposals for each image, and Top N proposals are selected based on the sequence of conducted actions. From Fig. 9, we can see that recalls of SS and RPN are much lower than our M2DRL approach, which verify the effectiveness of localization of ObjectDRL. Figure 10 shows the results of ObjectDRL with Top 1, 5 and 10 proposals. It is noted that the proposals are generated automatically by ObjectDRL, thus the number of proposals is variant for each image. The recall curve of Top 10 proposals is the best that verifies the effectiveness of ObjectDRL. In Fig. 11, we show the localized regions observed and localization action sequence conducted by the agent in ObjectDRL. We can see that the agent tries to guarantee the discriminative region, i.e. the object, in the center of the predicted box via conducting the most suitable action in each step. The red rectangle shows the final localization results, which verifies the effectiveness of ObjectDRL. We also calculate AUCs of the ObjectDRL, which are 0.501 and 0.508 on CUB-200-2011 and Cars-196 datasets, while the AUCs of g_atten are 0.494 and 0.487 respectively, as mentioned in Sect. 3.2.3. It verifies the effectiveness of the proposed ObjectDRL, which boosts the localization performance by reinforcement learning based on semantic reward function.

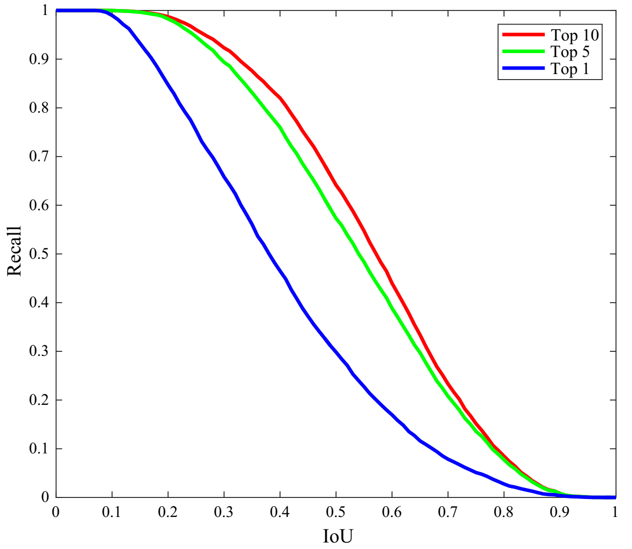


Fig. 10 Recall versus IoU overlap ratio on the CUB-200-2011 dataset. Here we show the results of our M2DRL approach with Top 1, 5 and 10 proposals

4.5.2 Effectiveness of PartDRL

PartDRL discovers the characteristics of the object, and ties to draw the distinctions between similar subcategories.

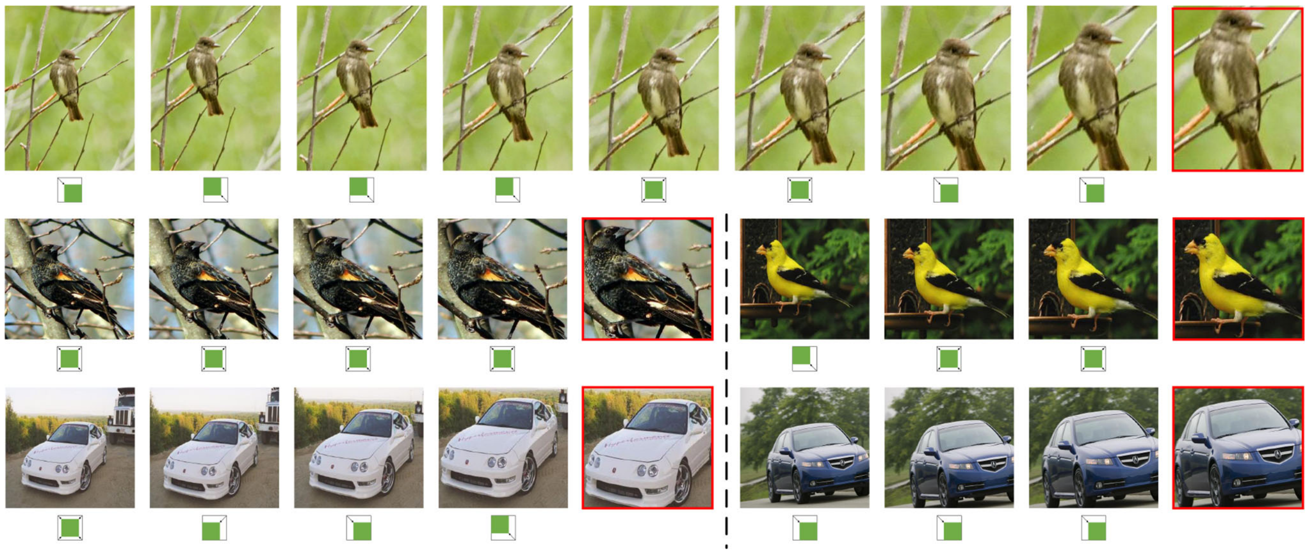


Fig. 11 Examples of localized regions observed by the agent, as shown in the upper line, and localization action sequence conducted by the agent in ObjectDRL, as shown in the lower line. The red rectangle shows the final localization results by ObjectDRL, which will be fed forward PartDRL to further explore more discriminative regions. The images in first and second lines are from CUB-200-2011 dataset, and those in third line are from Cars-196 dataset (Color figure online)

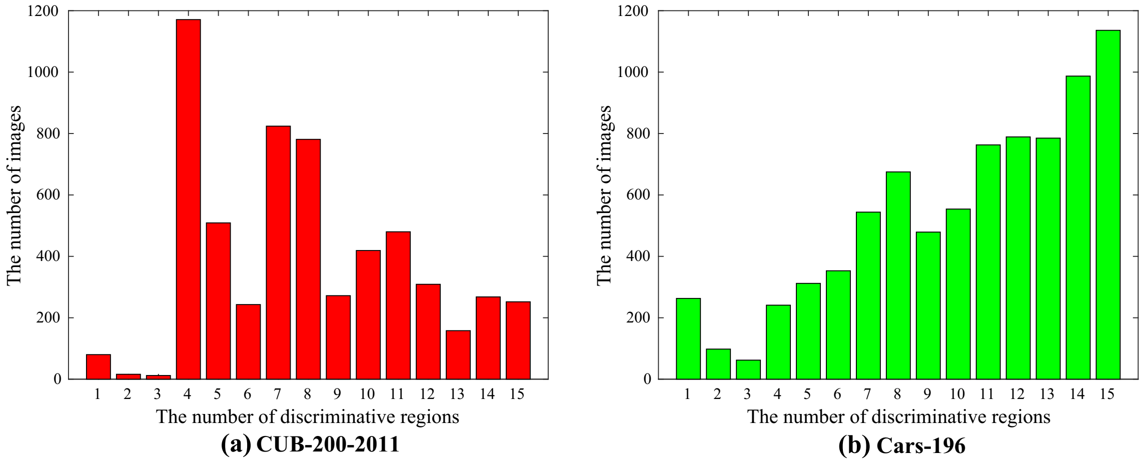


Fig. 12 Results of the number of discriminative regions localized by our proposed M2DRL approach for each image in the testing set of CUB-200-2011 and Cars-196 datasets. The coordinate axes denote the number of discriminative regions and its corresponding number of images in the testing set respectively

Figure 12 shows the results of the number of discriminative regions localized by our proposed M2DRL approach for each image in the testing set of CUB-200-2011 and Cars-196 datasets. Our M2DRL aims to address the “which problem” and “how many problem” automatically and adaptively, thus the number of the localized discriminative regions is different for each image. In our experiments, the number of discriminative regions is from 1 to 15. Due to the adaptive and flexible number of discriminative regions, our M2DRL approach achieves the best categorization accuracy, outperforms state-of-the-art methods that set the number of discriminative regions based on artificial prior or experimental validation results. In Fig. 13, we show the localized regions by the agent at each level of tree structure in PartDRL. It is noted that the images of “Level 0” are the localized object by ObjectDRL, not the original images. We can see that the agent tries to discover different discriminative regions, and discover regions in multiple scales at each level, which point the important characteristics to boost the categorization accuracy. The yellow and red rectangles show the localization results by action group 1 and 2 respectively, and they are different regions, which verifies the effectiveness of the tree structure in PartDRL.

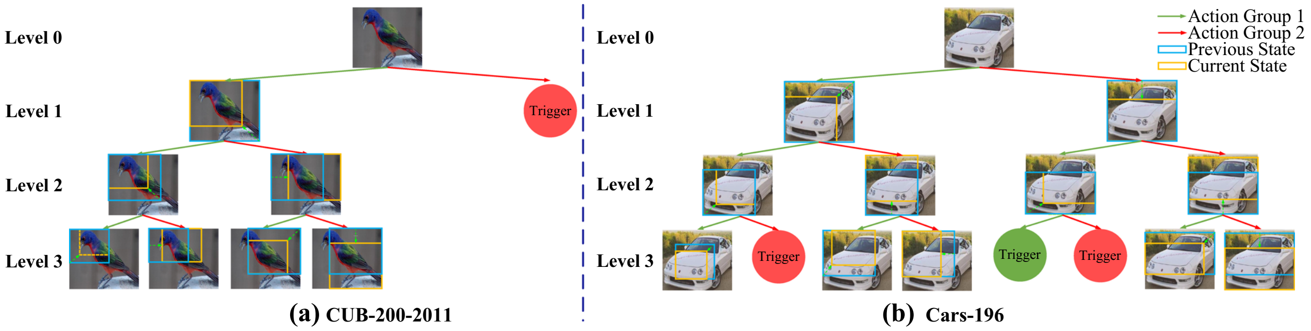


Fig. 13 Examples of localized regions by the agent at each level of tree structure in PartDRL. The yellow rectangles show the localization results by action group 1, and red rectangles show the localization results by action 2. Here we can see that the number of localized discriminative regions is different for each image (Color figure online)

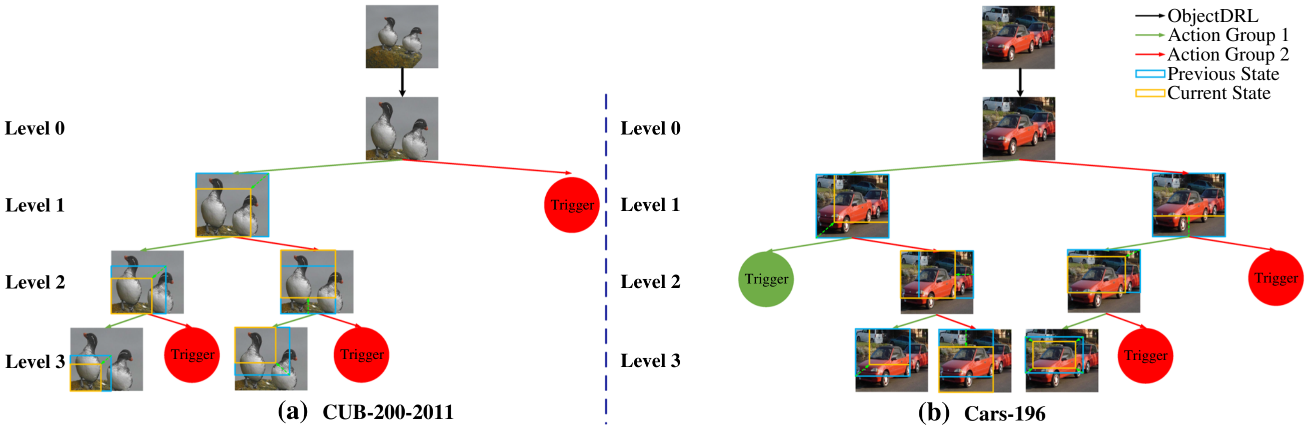


Fig. 14 Illustration of our M2DRL approach to handle images with multiple object instances

4.5.3 Discussion of Handling Multiple Object Instances

Actually, our proposed M2DRL approach can handle images with multiple object instances by the collaboration of ObjectDRL and PartDRL. Images with multiple object instances are existed in the two datasets, as shown in Fig. 14, where the first image contains multiple birds from the same subcategory, and the second image contains multiple cars from different subcategories. We can see that ObjectDRL first localize the region of the object in the image, as shown by the first two lines of Fig. 14. The images of "Level 0" show the localized object region by ObjectDRL, covering multiple object instances when the images contain multiple object instances. Then PartDRL is conducted on the object region, to further localize the regions of one single object or discriminative parts. For the example of CUB-200-2011, as shown in the left of Fig. 14, PartDRL can localize the head and foot of the bird. For the example of Cars-196, as shown in the right of Fig. 14, PartDRL can localize the region of one single car and the roof of the car. Through ObjectDRL and PartDRL, the discriminative regions of the images can be discovered to distinguish from other subcategories.

4.6 Effectiveness of Unsupervised Discriminative Localization

In this subsection, we explore the effectiveness of unsupervised discriminative localization (denoted as "UDL" in Table 3) in fine-grained visual categorization task. From Table 3, we can see that the application of unsupervised discriminative localization achieves a promising performance. It is an interesting and significant phenomenon that UDL achieves the similar categorization accuracy with PartDRL, while PartDRL utilizes the category label information. This is owing to the good generation of CNN model trained on ImageNet dataset. Unsupervised discriminative localization even outperforms the methods using the ground truth bounding box, such as Coarse-to-Fine (82.50% and 82.90%) (Yao et al. 2016) and PG Alignment (82.80%) (Krause et al. 2015) on CUB-200-2011 dataset, as shown in Table 1. This inspires us to further explore the study and application of unsupervised discriminative localization.

Table 3 Effectiveness of unsupervised discriminative localization

Methods	CUB-200-2011	Cars-196
MgDL	86.61	90.98
UDL	83.29	90.34
PartDRL	83.23	88.98

Table 4 Effectiveness of multi-scale representation learning

Methods	CUB-200-2011	Cars-196
M2DRL	87.21	93.25
MgDL (224 × 224)	86.61	90.98
MgDL (448 × 448)	86.42	92.82

"MgDL" denotes the multi-granularity discriminative localization, "224 × 224" and 448 × 448 denote different inputs with different scales

Table 5 Effectiveness of each stage in multi-granularity discriminative localization (MgDL)

Methods	CUB-200-2011	Cars-196
MgDL	86.61	90.98
ObjectDRL	85.29	89.93
PartDRL	83.23	88.98
Baseline	80.82	86.79

Table 6 Comparison between ObjectDRL and CAM

Methods	CUB-200-2011	Cars-196
ObjectDRL	85.29	89.93
Baseline w/bbox	84.97	91.36
CAM (Zhou et al. 2016)	83.74	88.79
Baseline	80.82	86.79

4.7 Effectiveness of Each Component in M2DRL

We conduct comprehensive experiments on CUB-200-2011 and Cars-196 datasets to verify the separate contribution of each component in our proposed M2DRL approach. Detailed experiments and analyses are as follows:

4.7.1 Effectiveness of Multi-scale Representation Learning

Here we verify the effectiveness of multi-scale representation learning. Different input images in multiple scales are applied in our M2DRL, which provide different but complementary information to boost the categorization accuracy. From Table 4, we can observe that integration of multiple scale information can facilitate the category accuracy by at least 0.6% on the two datasets.

4.7.2 Effectiveness of Multi-granularity Discriminative Localization

Here we conduct experiments to verify the effectiveness of multi-granularity discriminative localization (MgDL) with the input of 224 × 224 scale on CUB-200-2011 and Cars-196 datasets, as shown in Table 5. "Baseline" denotes recognizing the original images with the fine-tuned 19-layer VGGNet. "ObjectDRL" denotes that the localized objects are considered, without considering the localized discriminative parts of the object. "PartDRL" denotes that the localized discriminative parts are considered. "MgDL" denotes that both the object and parts are considered. We can observe that:

(I) Compared with the "Baseline", considering the localized discriminative parts can improve 2.41% and 2.19% on CUB-200-2011 and Cars-196 datasets respectively. It is because the good ability of PartDRL to localize the discriminative regions, which is also in multiple scales. These regions point out the subtle and local distinctions that are distinguished from other similar subcategories. PartDRL enhances the feature representation with more variances and discrimination.

(II) ObjectDRL boosts the categorization accuracy significantly, which brings 4.47% and 3.14% improvements compared with "Baseline" on CUB-200-2011 and Cars-196 datasets respectively. The categorization accuracies are also 2.06% and 0.95% higher than PartDRL. It is because that the localized region of ObjectDRL contains both the global features reflecting the appearance, and the local features reflecting the salient visual information. To further verify the effectiveness of our ObjectDRL approach, we compare it with salient object detection method (i.e. CAM) and baseline method with ground truth bounding box, as shown in Table 6. We use the CAM to generate discriminative regions that are related to the objects, and use them for categorization, achieving improvement than the baseline method. We also show the results of baseline method with ground truth bounding box, whose accuracies are 1.23% and 2.57% higher than CAM. However, considering that CAM does not use ground truth bounding box, its categorization results are promising. Our ObjectDRL outperforms CAM by 1.55% and 1.14% on CUB-200-2011 and Cars-196 dataset respectively, which is mainly because of the different learning strategies of our OjectDRL and CAM. In the training phase, CAM only learns from the original images. While our ObjectDRL can learn more discriminative features with multiple scales and granularities from the sub-regions of the image, which are generated by conducting specific actions. So our ObjectDRL can achieve better classification accuracy. It is noted that, even without using ground truth bounding box, our proposed ObjectDRL still achieves nearly similar or even better performance than the baseline method using ground truth bounding box, which verifies the effectiveness of ObjectDRL.

(III) The integration of ObjectDRL and PartDRL can further achieve more accurate result than only one-stage DRL, e.g. 86.61% versus 85.29% and 83.23% on CUB-200-2011 dataset. Compared with "Baseline", an improvement of 5.79% is achieved. It shows the complementarity of ObjectDRL and PartDRL as well as the effectiveness of the two-stage deep reinforcement learning architecture. ObjectDRL and PartDRL have different but complementary gazes at different regions of the image, providing more salient and variant visual information to boost the fine-grained representation learning as well as the categorization.

4.7.3 Effectiveness of Semantic Reward Function

We conduct experiments to show the effectiveness of the proposed semantic reward function with the input of 224 × 224 scale on CUB-200-2011 and Cars-196 datasets. In Table 7, "RA" denotes the attention-based reward functions, and "RC" denotes the category-based reward function. From Table 7, we can observe that:

(I) Attention-based reward and category-based reward achieve similar categorization accuracy, which shows that the attention information and category information play similar roles in the fine-grained visual categorization.

(II) The joint application of attention-based and category-based reward functions further improve the categorization accuracy due to the fact that the two reward functions focus on different but complementary aspects: attention-based reward provides the discriminative visual information, and category-based reward provides the conceptual visual information.

Table 7 Effectiveness of semantic reward function

Methods	CUB-200-2011	Cars-196
MgDL	86.61	90.98
RA	85.79	90.37
RC	85.23	90.00

"RA" denotes the attention-based reward functions, and "RC" denotes the category-based reward function

4.7.4 Effects of Using Grounding Truth Bounding Box

To further verify the effectiveness our proposed approach, we conduct experiments in the following aspects:

(I) We use ground truth bounding box (bbox) in training phase. Specifically, in ObjectDRL, we use the bbox based reward function, i.e. Eq. (2), instead of attention-based reward function. In PartDRL, attention information is still applied. The results are shown in Table 8. We can see that using the bbox improves the categorization accuracy only by 0.34% and 0.14% on CUB-200-2011 and Cars-196 datasets respectively. Even without using ground truth bounding box, our ObjectDRL can achieve similar performance, which is mainly because that the attention information points out the regions with discriminative and significant information for categorization. In Fig. 15, we show the original images and their attention maps, as well as the bounding boxes generated based on their attention maps (green rectangles, denoted as "$bbox_{am}$"), and the ground truth bounding boxes (red rectangles, denoted as "$bbox_{gt}$"). For further comparing the attention maps of different objects, we divide them into five groups by different intersection over union (IoU) of $bbox_{am}$ and $bbox_{gt}$, i.e. 0–0.2, 0.2–0.4, 0.4–0.6, 0.6–0.8, 0.8–1. We can observe that the attention maps show the discriminative regions of the images. When IoU>0.4, the $bbox_{am}$ can cover the most regions of the objects, and there are 72.3% and 60.2% of the testing data with IoU>0.4 in CUB-200-2011 and Cars-196 datasets respectively. Even when IoU<0.4, the $bbox_{am}$ can still cover the main discriminative regions of the objects, such as heads or bodies, which are significant to distinguish from other subcategories. In the stage of ObjectDRL, our target is to localize the main discriminative regions of the images, which are not always the regions of the entire objects. Therefore, we can use the attention map to approximate the ground-truth in ObjectDRL.

(II) We use ground truth bounding box as the input of PartDRL, and show the results in Table 9. We can see that using the bbox can improve the categorization accuracy by 0.28% and 1.03% on CUB-200-2011 and Cars-196 datasets respectively. We also observe the results of localized parts, and find that more discrim-

Table 8 Effects of using ground truth bounding box based reward function instead of attention-based reward function

Methods	CUB-200-2011	Cars-196
MgDL	86.61	90.98
MgDL w/bbox	86.95	91.12

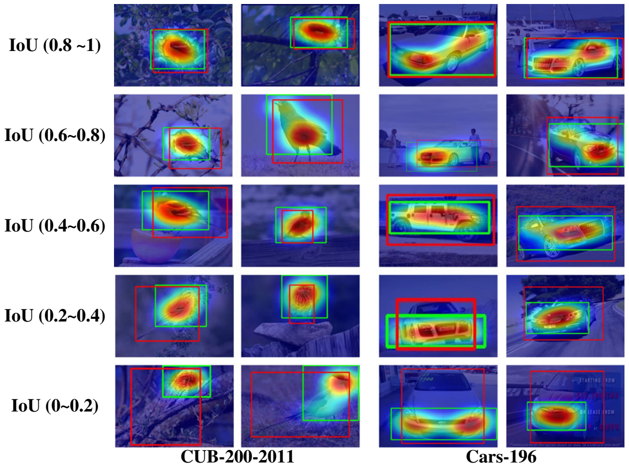


Fig. 15 Attention maps of different objects in CUB-200-2011 and Cars-196 datasets

Table 9 Effects of using ground truth bounding box on PartDRL

Methods	CUB-200-2011	Cars-196
PartDRL	83.23	88.98
PartDRL w/bbox	83.51	90.01

inative parts can be localized. However, considering that the labeling of ground truth bounding box is time-consuming and labor-consuming, as well as it is not available in the real-word applications, it is not suitable to use ground truth bounding box in the test phase. Even without using ground truth bounding box, our PartDRL can achieve similar performance, which verifies that the attention information is helpful for localizing discriminative regions as well as fine-grained visual categorization.

5 Conclusion

To address the "which problem" and "how problem", this paper proposes the M2DRL approach for fine-grained visual categorization. First, multi-granularity discriminative localization localizes discriminative regions in different granularities hierarchically ("which problem"), and determines the number of discriminative regions adaptively ("how many problem"). Then, multi-scale representation learning helps to localize objects in different scales and encode images in different scales, boosting the categorization performance. Semantic reward function drives M2DRL to fully capture the discriminative and conceptual visual information, via jointly integrating the attention-based reward and category-based reward. Furthermore, unsupervised discriminative localization avoids the heavy labor consumption of labeling, and extremely strengthens the *usability* and *scalability* of our M2DRL approach. Compared with state-of-the-art methods on two widely-used fine-grained visual categorization datasets, our M2DRL approach achieves the best categorization accuracy. Besides, the effectiveness of unsupervised discriminative localization is also verified on these two datasets, which achieves promising performance.

In the future, we devote to improving this work in the following two aspects: First, integrate discriminative localization and fine-grained visual categorization in the same network, rather than separated processing by deep reinforcement learning and convolutional neural network, which will further improve their performance in a complementary manner. Second, unsupervised discriminative localization achieves promising results, which should be further explored to bring more improvement in categorization accuracy as well as more scalability and usability of fine-grained visual categorization, marching forward the practical application. Both of these two aspects will be employed to further improve the fine-grained visual categorization performance.

References

Ba, J., Mnih, V., & Kavukcuoglu, K. (2014). Multiple object recognition with visual attention. In *International conference on learning representations (ICLR)*. arXiv:1412.7755.

Berg, T., & Belhumeur, P. (2013). 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)* (pp. 955–962).

Branson, S., Van Horn, G., Belongie, S., & Perona, P. (2014a). Bird species categorization using pose normalized deep convolutional nets. arXiv:1406.2952.

Branson, S., Van Horn, G., Wah, C., Perona, P., & Belongie, S. (2014b). The ignorant led by the blind: A hybrid human-machine vision system for fine-grained categorization. *International Journal of Computer Vision (IJCV)*, *108*(1–2), 3–29.

Cai, S., Zuo, W., & Zhang, L. (2017). Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 511–520).

Caicedo, J. C., & Lazebnik, S. (2015). Active object localization with deep reinforcement learning. In *International conference of computer vision (ICCV), IEEE* (pp. 2488–2496).

Chai, Y., Lempitsky, V., & Zisserman, A. (2013). Symbiotic segmentation and part localization for fine-grained categorization. In *International conference of computer vision (ICCV)* (pp. 321–328).

Cui, Y., Zhou, F., Lin, Y., & Belongie, S. (2016). Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1153–1162).

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 248–255).

Everingham, M., Eslami, S. A., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision (IJCV)*, *111*(1), 98–136.

Fu, J., Zheng, H., & Mei, T. (2017). Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Girshick, R. (2015). Fast R-CNN. In *International conference of computer vision (ICCV)*.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 580–587).

Gonzalez-Garcia, A., Modolo, D., & Ferrari, V. (2018). Do semantic parts emerge in convolutional neural networks? *International Journal of Computer Vision (IJCV)*, *126*(5), 476–494.

He, X., & Peng, Y. (2017a). Fine-grained image classification via combining vision and language. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

He, X., & Peng, Y. (2017b). Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification. In: *AAAI conference on artificial intelligence (AAAI)* (pp. 4075–4081).

Huang, S., Xu, Z., Tao, D., & Zhang, Y. (2016). Part-stacked cnn for fine-grained visual categorization. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1173–1182).

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203.

Jaderberg, M., Simonyan, K., & Zisserman, A. (2015). Spatial transformer networks. In *Neural information processing systems (NIPS)* (NIPS) (pp. 2017–2025).

Jie, Z., Liang, X., Feng, J., Jin, X., Lu, W., & Yan, S. (2016). Tree-structured reinforcement learning for sequential object localization. In *Neural information processing systems (NIPS)* (pp. 127–135).

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research (JAIR)*, *4*, 237–285.

Kong, S., & Fowlkes, C. (2017). Low-rank bilinear pooling for fine-grained classification. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 7025–7034). IEEE.

Krause, J., Stark, M., Deng, J., & Fei-Fei, L. (2013). 3d object representations for fine-grained categorization. In *International conference of computer vision (ICCV)* (pp. 554–561).

Krause, J., Gebru, T., Deng, J., Li, L. J., & Fei-Fei, L. (2014). Learning features and parts for fine-grained recognition. In *International conference on pattern recognition (ICPR)* (pp. 26–33).

Krause, J., Jin, H., Yang, J., Fei-Fei, L. (2015). Fine-grained recognition without part annotations. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 5546–5555).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lin, D., Shen, X., Lu, C., & Jia, J. (2015a). Deep lac: Deep localization, alignment and classification for fine-grained recognition. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1666–1674).

Lin, T. Y., RoyChowdhury, A., & Maji, S. (2015b). Bilinear CNN models for fine-grained visual recognition. In: *International conference of computer vision (ICCV)* (pp. 1449–1457).

Maji, S., Rahtu, E., Kannala, J., Blaschko, M., & Vedaldi, A. (2013). Fine-grained visual classification of aircraft. arXiv:1306.5151.

Mathe, S., Pirinen, A., & Sminchisescu, C. (2016). Reinforcement learning for visual object detection. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2894–2902).

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529.

Neider, M. B., & Zelinsky, G. J. (2006). Searching for camouflaged targets: Effects of target-background similarity on visual search. *Vision Research*, *46*(14), 2217–2235.

Nilsback, M. E., & Zisserman, A. (2008). Automated flower classification over a large number of classes. In *Sixth Indian conference on computer vision, graphics & image processing* (pp. 722–729).

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, *9*(1), 62–66.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, *42*(1), 107–123.

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

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In: *Advances in neural information processing systems (NIPS)* (pp. 91–99).

Schaul, T., Quan, J., Antonoglou, I., & Silver, D. (2015). Prioritized experience replay. arXiv:1511.05952.

Sfar, A. R., Boujemaa, N., & Geman, D. (2015). Confidence sets for fine-grained categorization and plant species identification. *International Journal of Computer Vision (IJCV)*, *111*(3), 255–275.

Simon, M., & Rodner, E. (2015). Neural activation constellations: Unsupervised part model discovery with convolutional networks. In *International conference of computer vision (ICCV)* (pp. 1143–1151).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1). Cambridge: MIT Press.

Tatler, B. W., Baddeley, R. J., & Vincent, B. T. (2006). The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task. *Vision Research*, *46*(12), 1857–1862.

Uijlings, J. R., van de Sande, K. E., Gevers, T., & Smeulders, A. W. (2013). Selective search for object recognition. *International Journal of Computer Vision (IJCV)*, *104*(2), 154–171.

Van Hasselt, H., Guez, A., & Silver, D. (2016). Deep reinforcement learning with double q-learning. In *AAAI conference on artificial intelligence (AAAI)* (Vol. 16, pp. 2094–2100).

Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). The caltech-ucsd birds-200-2011 dataset. California Inst. Technol., Pasadena, CA, USA, Tech. Rep (CNS-TR-2011-001).

Wang, D., Shen, Z., Shao, J., Zhang, W., Xue, X., & Zhang, Z. (2015). Multiple granularity descriptors for fine-grained categorization. In *International conference of computer vision (ICCV)* (pp. 2399–2406).

Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., & Gong, Y. (2010). Locality-constrained linear coding for image classification. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3360–3367).

Wang, Y., Choi, J., Morariu, V., & Davis, L. S. (2016a). Mining discriminative triplets of patches for fine-grained classification. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1163–1172).

Wang, Y., Morariu, V. I., & Davis, L. S. (2016b). Weakly-supervised discriminative patch learning via CNN for fine-grained recognition. arXiv:1611.09932.

Wang, Z., Schaul, T., Hessel, M., Hasselt, H., Lanctot, M., & Freitas, N. (2016c). Dueling network architectures for deep reinforcement learning. In *International conference on machine learning (ICML)* (pp. 1995–2003).

Xiao, T., Xu, Y., Yang, K., Zhang, J., Peng, Y., & Zhang, Z. (2015). 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)* (pp. 842–850).

Xie, L., Tian, Q., Hong, R., Yan, S., & Zhang, B. (2013). Hierarchical part matching for fine-grained visual categorization. In *International conference of computer vision (ICCV)* (pp. 1641–1648).

Xie, L., Zheng, L., Wang, J., Yuille, A. L., & Tian, Q. (2016). Interactive: Inter-layer activeness propagation. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 270–279).

Xie, S., Yang, T., Wang, X., & Lin, Y. (2015). Hyper-class augmented and regularized deep learning for fine-grained image classification. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2645–2654).

Xu, Z., Huang, S., Zhang, Y., & Tao, D. (2018). Webly-supervised fine-grained visual categorization via deep domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *40*(5), 1100–1113.

Xu, Z., Tao, D., Huang, S., & Zhang, Y. (2017). Friend or foe: Fine-grained categorization with weak supervision. *IEEE Transactions on Image Processing (TIP)*, *26*(1), 135–146.

Yang, S., Bo, L., Wang, J., & Shapiro, L. G. (2012). Unsupervised template learning for fine-grained object recognition. In *Neural information processing systems (NIPS)* (pp. 3122–3130).

Yao, H., Zhang, S., Zhang, Y., Li, J., & Tian, Q. (2016). Coarse-to-fine description for fine-grained visual categorization. *IEEE Transactions on Image Processing (TIP)*, *25*(10), 4858–4872.

Zhang, H., Xu, T., Elhoseiny, M., Huang, X., Zhang, S., Elgammal, A., & Metaxas, D. (2016a). Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1143–1152).

Zhang, J., Marszałek, M., Lazebnik, S., & Schmid, C. (2007). Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision (IJCV)*, *73*(2), 213–238.

Zhang, L., Yang, Y., Wang, M., Hong, R., Nie, L., & Li, X. (2016b). Detecting densely distributed graph patterns for fine-grained image categorization. *IEEE Transactions on Image Processing (TIP)*, *25*(2), 553–565.

Zhang, N., Farrell, R., Iandola, F., & Darrell, T. (2013). Deformable part descriptors for fine-grained recognition and attribute prediction. In *International conference of computer vision (ICCV)* (pp. 729–736).

Zhang, N., Donahue, J., Girshick, R., & Darrell, T. (2014). Part-based R-CNNs for fine-grained category detection. In *International conference on machine learning (ICML)* (pp. 834–849).

Zhang, X., Xiong, H., Zhou, W., Lin, W., & Tian, Q. (2016c). Picking deep filter responses for fine-grained image recognition. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1134–1142).

Zhang, X., Xiong, H., Zhou, W., & Tian, Q. (2016d). Fused one-vs-all features with semantic alignments for fine-grained visual categorization. *IEEE Transactions on Image Processing (TIP)*, *25*(2), 878–892.

Zhang, X., Xiong, H., Zhou, W., Lin, W., & Tian, Q. (2017). Picking neural activations for fine-grained recognition. *IEEE Transactions on Multimedia (TMM)*, *19*(12), 2736–2750.

Zhang, Y., Wei, X. S., Wu, J., Cai, J., Lu, J., Nguyen, V. A., et al. (2016e). Weakly supervised fine-grained categorization with part-based image representation. *IEEE Transactions on Image Processing (TIP)*, *25*(4), 1713–1725.

Zhao, B., Wu, X., Feng, J., Peng, Q., & Yan, S. (2017a). Diversified visual attention networks for fine-grained object classification. *IEEE Transactions on Multimedia (TMM)*, *19*(6), 1245–1256.

Zhao, D., Chen, Y., & Lv, L. (2017b). Deep reinforcement learning with visual attention for vehicle classification. *IEEE Transactions on Cognitive and Developmental Systems*, *9*(4), 356–367.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2015). Object detectors emerge in deep scene CNNs. In *International conference on learning representations (ICLR)*.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2921–2929).

Zhou, F., & Lin, Y. (2016). Fine-grained image classification by exploring bipartite-graph labels. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1124–1133).