SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

March 2023, Vol. 66 132105:1–132105:13
https://doi.org/10.1007/s11432-021-3493-7

Class attention network for image recognition

Gong CHENG*, Pujian LAI, Decheng GAO & Junwei HAN

School of Automation, Northwestern Polytechnical University, Xi'an 710072, *China*

Received 11 August 2021/Revised 21 March 2022/Accepted 22 April 2022/Published online 18 January 2023

Abstract Visual attention has become a popular and widely used component for image recognition. Although various attention-based methods have been proposed and achieved relatively competitive results, it is observed that the semantic features of each class are likely to entangle with each other, and few studies focus on explicitly extracting category-aware features so far. To address this issue, this paper presents an attention-based image recognition method by using class-specific dictionary learning to disentangle the neural network's outputs into class-dependent features, thus boosting their discrimination abilities. Specifically, we develop a class attention network (CANet) via integrating a simple yet effective class-specific attention encoding (CAE) module on the top of convolutional layers. Given the feature maps of the convolutional neural networks (CNNs), the CAE module learns a class-specific dictionary, which is leveraged to encode attention maps for each category. Then these attention maps are multiplied by the input features for class-wise adaptive feature refinement. Extensive experiments on the PASCAL VOC 2007, PASCAL VOC 2012, MS COCO, and CUB-200-2011 datasets demonstrate the fabulous performance of our method on multiple visual recognition tasks, including multi-label image classification and fine-grained visual classification. In addition, the visualization results testify that CNNs can explicitly learn class-wise feature representations by introducing class-specific dictionary learning.

Keywords visual attention, class-specific attention encoding, class attention network, dictionary learning

Citation Cheng G, Lai P J, Gao D C, et al. Class attention network for image recognition. Sci China Inf Sci, 2023, 66(3): 132105, https://doi.org/10.1007/s11432-021-3493-7

1 Introduction

Image recognition, including multi-label image recognition and single-label image recognition, refers to a fundamental and practical task of automatically assigning multiple possible labels or one possible label to an image based on the visual content. In recent year, convolutional neural networks (CNNs) have been successfully applied to a variety of image recognition tasks [1–12].

However, the representation power of CNNs is still somewhat limited in dealing with challenging image recognition tasks. Take image classification and fine-grained visual categorization (FGVC) as examples, Figures 1(a) and (b) show some example images and their corresponding labels from the PASCAL VOC dataset [13] and the MS COCO dataset [14], respectively. As shown, the large intra-class variances caused by appearance, scale, illumination, occlusion, viewpoint, etc., and the interaction between object categories significantly increase the difficulties of image recognition, making image classification more difficult. Besides, Figure 1(c) illustrates some bird images and their corresponding species from the CUB-200-2011 dataset [15], a challenging dataset of 200 bird species. We can observe that the big intra-class variances brought by pose, scales, location, etc., and the subtle inter-class differences still make FGVC a particularly effortful task. It is natural to throw a question: can we design a method with the capability of enhancing the representation power?

Over the years, continuous efforts have been made to address the aforementioned issues. Visual attention, as an effective way to enhance the representation power of CNNs, has been studied extensively in previous studies [16–23]. However, existing attention-based methods rarely focus on explicitly modeling class-specific feature representations. The performance improvement can be obtained by focusing only on

* Corresponding author (email: gcheng@nwpu.edu.cn)

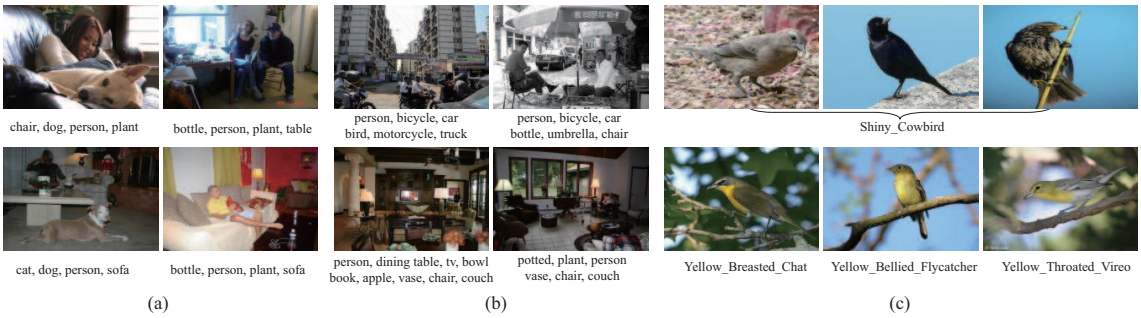

Figure 1 (Color online) Some example images from different datasets. The intra-class variations and the composition and interaction between different object categories make the task of image classification more challenging. (a) PASCAL VOC dataset; (b) MS COCO dataset; (c) CUB-200-2011 dataset.

the regions that are semantically relevant to the considered labels [24], since the class-specific representations can capture and analyze the most crucial information related to a specific category and maintain huge between-class separation. Based on this observation, we wonder if CNNs can acquire the learning ability that expressly models class-specific representations. To approach our goal, there is a key issue to be addressed, which is how to learn class-specific feature representations.

During the past few years, a number of class-specific dictionary learning methods, in which each dictionary atom is assigned to a single class and the dictionary atoms associated with different classes are encouraged to be as independent as possible, have been explored extensively [25, 26]. For instance, Ramirez et al. [25] proposed a structured dictionary learning scheme by promoting the discriminative ability between different class-specific sub-dictionaries. Yang et al. [26] proposed a dictionary learning framework which employs Fisher discrimination criterion to learn class-specific dictionaries. These studies manifest that learning a class-specific dictionary is feasible. Thus, we can use a class-specific dictionary to encode class-specific information.

The core of encoding class-specific information is that the semantic features of each class can be disentangled from the original features of a neural network's outputs. Recently, several attempts have been made to decompose the class-specific features. Zeiler et al. [27] demonstrated that high-layer convolutional filters extract high-level semantic features which might relate to certain classes to some extent. Prakash et al. [28] showed that the redundant overlap between the features extracted by different filters makes it possible to learn specialized filters. Chu et al. [29] decomposed the features of each class into a class-generic component and a class-specific component. Inspired by these studies and class-specific dictionary learning methods, this paper proposes a class-specific attention encoding (CAE) module by introducing a dictionary with a set of atom groups, where each atom group encodes a class-wise attention map with explicit class semantic information. Then, we plug CAE into a CNN to construct an end-to-end class attention network (CANet) for visual recognition. Using the class-specific attention maps generated via CAE module, we can guide the network to learn more discriminative feature representations. To train CANet, we introduce an attention loss as a regularization term to enforce the attention maps to better correspond to different semantic classes.

To sum up, the main contributions of this paper are as follows.

- We propose CAE module to enforce the CNNs to explicitly encode class attentions. The CAE module can be conveniently embedded into current CNNs to boost their discrimination abilities.
- With the CAE module and its corresponding attention loss, we construct an end-to-end CANet to extract highly category-related feature representations.
- We extensively evaluate our method on multiple visual recognition tasks, including multi-label classification and fine-grained visual classification. The experiment results demonstrate the effectiveness of our method. Moreover, the visualization results prove that convolutional neural networks can explicitly learn class-wise feature representations by introducing class-specific dictionary learning.

The rest of this paper is organized as follows. We first review the related work in Section 2. Then, Section 3 describes our method in detail. After that, Section 4 presents the comprehensive experiments on four datasets. Finally, a brief conclusion of this paper is summarized in Section 5.

2 Related work

2.1 Multi-label image classification

Recently, multi-label image classification problem has attracted a lot of attention. Several state-of-the-art single-label image classification networks have been adopted to address this task [30–32]. Besides, Wang et al. [33] proposed a proposal-free approach which uses stochastic scaling and cropping for better capturing the detailed appearances and spatial layout information to improve the performance. Hypotheses-CNN-Pooling [34] made predictions on each proposal and then aggregated all the predictions as the final output through category-wise max-pooling. Ref. [35] proposed a multi-view and multi-instance framework by incorporating local information to enhance features for handling the problem of multi-label classification.

To better consider the correlations between labels instead of treating each label independently, a series of studies were introduced, such as MLGCN [36], DSDL [37], conditional graphical Lasso [38], and CNN-RNN [39]. More recently, Refs. [40, 41] further took advantage of visual attention mechanism to search local discriminative regions and Chen et al. [42] computed a probabilistic matrix as the relation edge between each label in a graph to aid multi-label image classification.

2.2 Fine-grained visual classification

Some algorithms [10,43–45] guide the training of deep CNN models for fine-grained visual categorization, relying on object annotations or even dense part annotations. For example, SPDA-CNN [45] proposed a network consisting of detection and classification sub-networks.

The above approaches are labor-wasting, so more and more methods that only require image-level annotations have been developed. Lin et al. proposed bilinear pooling [10] and improved bilinear pooling [46], where two features are combined at each location using the outer product. Fu et al. [47] developed RA-CNN to recursively learn discriminative region attention to obtain region-based feature representations at multiple scales in a mutually reinforced way. To generate multiple attention locations at the same time, Zheng et al. [48] proposed multi-attention CNN, which simultaneously locates several body parts. MAMC [49] regulated multiple object parts among different input images by using multiple attention region features of each input image. WS-DAN [50] combined weakly supervised learning with data augmentation to promote the model to extract more discriminative features from multiple local regions.

2.3 Visual attention

Visual attention networks have been extensively proposed to automatically mine relevant and informative regions for image recognition in recent years. SENet [18] proposed a novel unit, termed squeeze-and-excitation (SE) block to model the channel relationship for improving visual classification. In [51], the cascade attention method was proposed and the features of different CNN layers were concatenated to gain discriminative representation as the input of final linear classifier. ACNet [52] focused on different discriminative regions using the attention transformer inserted into the convolutional operations along the edges of the tree. SRN [48] learned attention heatmaps to specify spatial relations between labels. Refs. [40, 53–55] aimed to learn accurate attention regions to strengthen their relevance. Ref. [16] introduced a new attention consistency loss which was combined with multi-label image classification loss for network training. ADD-GCN [56] adopted a dynamic graph convolutional network to model the relation of content-aware category representations that are generated by a semantic attention module.

However, the current attention methods for image recognition just concentrate on enhancing key regions' feature representation, without explicitly learning class-wise attentions for each of the visual classes. Differently, this paper proposes a class-specific attention module, CAE, which not only focuses on the key regions in the input images but also is specific for each category. NetVLAD [57] and EncNet [58] shared a similar philosophy with our method. To be specific, NetVLAD proposed a new method to learn the vector of locally aggregated descriptors (VLAD) in an end-to-end manner. It assigns each C-dimensional pixel vector coming from the output features of the backbone to multiple clusters in a soft-assignment manner. In contrast to NetVLAD, our method learns a more compact set of atom groups based on the input features of the CAE module to encode class-specific representations. EncNet proposed a context encoding module to capture the semantic context of scenes and selectively highlight class-dependent feature maps. Essentially, EncNet predicts the scaling factors of feature channels. Different from EncNet, our method learns the scaling factors of spatial-wise features.

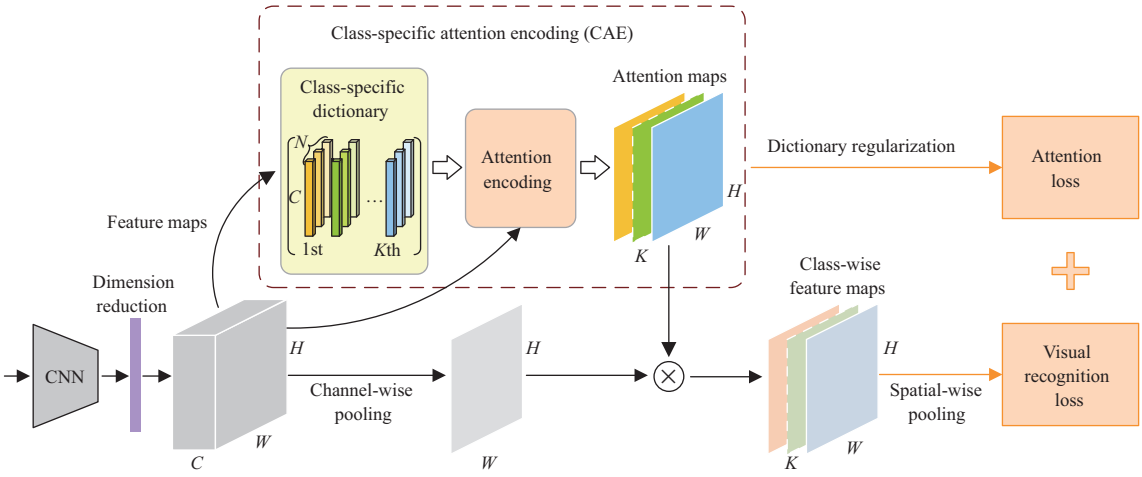

Figure 2 (Color online) The architecture of our proposed class attention network (CANet). The CANet consists of two key parts: a class-specific attention encoding (CAE) module and the attention loss function. The CAE module encodes class-wise attention maps. The attention loss is used as the dictionary regularization term to enforce CANet to better learn class-wise semantic information. In addition, dimension reduction is conducted by a set of 1×1 convolutions. Here, K denotes the number of object categories, N denotes the number of atoms for each class in the dictionary, and C represents the dimension of each atom.

3 Methodology

As shown in Figure 2, the framework of our CANet contains two main parts. (i) A class-specific attention encoding module that learns a class-specific dictionary to encode class attention maps. Then the attention maps are encoded into the network to acquire class-aware feature maps. (ii) An attention loss that is a new regularization term to encourage CANet for learning class-wise semantic information. By jointly optimizing the attention loss and visual recognition loss, we can better achieve the tasks of visual recognition such as multi-label image classification and FGVC.

3.1 Class-specific dictionary

Image recognition is typically based on learning a synthesis dictionary which yields the representations of each image as a sparse linear combination of the atoms of the learned dictionary [59]. Given a training data $\boldsymbol{X}$, the dictionary learning can be formulated as

$$\min_{\boldsymbol{D},\boldsymbol{Z}} \|\boldsymbol{X} - \boldsymbol{D}\boldsymbol{Z}\|^2, \tag{1}$$

where $\boldsymbol{D}$ is the dictionary to be learned and $\boldsymbol{Z}$ consists of the loading coefficients.

In this paper, we propose a class-specific dictionary learning. The intuition of the concept of class-specific dictionary is to learn disentangled representations for each class. Let $\boldsymbol{D} \in \mathbb{R}^{C\times K\times N}$ denote a class-specific dictionary, where K denotes the number of atom groups, which is the same as the number of object categories, N denotes the number of atoms for each class in the dictionary, and C represents the dimension of each atom vector, which is consistent with the number of channels of the output feature maps from the convolutional neural networks. It is worth noting that each atom group encodes a class-wise attention map with explicit class semantic information. Furthermore, the class-specific dictionary should be initialized before the training processing of the network.

In this way, for each pixel in the input feature maps, we use different atom groups to respectively encode the information and get K attention maps, as displayed in Figure 2. We call this process attention encoding, which will be described next.

3.2 Attention encoding

To preserve as much information as possible for subsequent attention encoding and image recognition, and considering the capacity of graphics processing unit (GPU) resources, we first perform a dimension reduction operation on the last convolution features with C 1×1 convolutions to obtain $\boldsymbol{X} \in \mathbb{R}^{C\times H\times W}$ used for the input of CAE module, as shown in Figure 2. Here, the value of C is typically much smaller than the channel number of the last convolution features, and H and W denote the height and width of

the feature maps. Given each C-dimensional pixel vector $\boldsymbol{x}_i \in \mathbb{R}^C$, we can calculate its similarity with each atom vector in the class-specific dictionary.

Suppose $\boldsymbol{D} = \left\{\boldsymbol{d}_l \in \mathbb{R}^C \mid l = 1, \ldots, K \times N\right\}$ is the collection of $K \times N$ atoms, where $\boldsymbol{d}_l$ is an atom vector in the dictionary. For each pixel vector $\boldsymbol{x}_i$ and atom vector $\boldsymbol{d}_l$, the calculation of their similarity is formulated as

$$S_{il} = \sigma(\boldsymbol{x}_i, \boldsymbol{d}_l), \tag{2}$$

where $\sigma(\cdot)$ denotes a kernel function. We use a_{il} to denote the response of the pixel vector $\boldsymbol{x}_i$ obtained from the l-th atom vector $\boldsymbol{d}_l$. Thus, a_{il} can be formulated as follows:

$$a_{il} = \frac{\sigma\left(\boldsymbol{x}_i, \boldsymbol{d}_l\right)}{\sum_{j=1}^{K \times N} \sigma\left(\boldsymbol{x}_i, \boldsymbol{d}_j\right)}. \tag{3}$$

In order to avoid complex computation, we need to choose a proper kernel function. There are several choices, such as inner dot kernel $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{b}$, radial basis function (RBF) kernel $\exp\left(-\|\boldsymbol{a}-\boldsymbol{b}\|_2^2/\sigma^2\right)$, sigmoid kernel $\tanh\left(\beta \boldsymbol{a}^{\mathrm{T}}\boldsymbol{b} + \theta\right)$, and so on. Considering the computational efficiency, we adopt the inner dot kernel in exponential form. It is formulated as

$$\sigma\left(\boldsymbol{x}_i, \boldsymbol{d}_l\right) = \exp\left(-\boldsymbol{d}_l^{\mathrm{T}}\boldsymbol{x}_i\right). \tag{4}$$

And now, Eq. (2) can be reformulated into a more general form:

$$a_{il} = \frac{\exp\left(-\boldsymbol{d}_l^{\mathrm{T}}\boldsymbol{x}_i\right)}{\sum_{j=1}^{K \times N} \exp\left(-\boldsymbol{d}_j^{\mathrm{T}}\boldsymbol{x}_i\right)}. \tag{5}$$

Thus, we use (4) to calculate the response for each atom vector on the input feature maps. Let $\boldsymbol{A} \in \mathbb{R}^{K \times N \times H \times W}$ denote the obtained response matrices. We conduct intra-group average pooling operation along the second dimension (N) to get class-wise attention maps $\boldsymbol{A}^{\mathrm{att}} \in \mathbb{R}^{K \times H \times W}$. Let $a_{k,h,w}^{\mathrm{att}} \in \boldsymbol{A}^{\mathrm{att}}$, which is conducted as follows:

$$a_{k,h,w}^{\mathrm{att}} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{A}_{k,i,h,w}. \tag{6}$$

Notably, if we directly multiply $\boldsymbol{X} \in \mathbb{R}^{C \times H \times W}$ with the attention matrix $\boldsymbol{A}^{\mathrm{att}} \in \mathbb{R}^{K \times H \times W}$ to get the attention-guided feature maps, it would result in high computational complexity and more GPU consumption. To avoid this problem, a channel-wise average pooling operation is firstly implemented on the feature $\boldsymbol{X} \in \mathbb{R}^{C \times H \times W}$ to obtain $\boldsymbol{X'} \in \mathbb{R}^{1 \times H \times W}$. Then, we get the attention-guided feature maps $\boldsymbol{X}^{\mathrm{att}} \in \mathbb{R}^{K \times H \times W}$ by multiplying $\boldsymbol{X'} \in \mathbb{R}^{1 \times H \times W}$ with the attention matrix $\boldsymbol{A}^{\mathrm{att}} \in \mathbb{R}^{K \times H \times W}$, as shown in Figure 2. The procedure can be formulated as

$$\boldsymbol{X}_k^{\mathrm{att}} = \boldsymbol{A}_k^{\mathrm{att}} \otimes \boldsymbol{X'}, \tag{7}$$

$$\boldsymbol{X}^{\mathrm{att}} = \mathrm{cat}\left(\boldsymbol{X}_1^{\mathrm{att}}, \boldsymbol{X}_2^{\mathrm{att}}, \ldots, \boldsymbol{X}_K^{\mathrm{att}}\right), \tag{8}$$

where $\boldsymbol{A}_k^{\mathrm{att}} \in \mathbb{R}^{H \times W}$ denotes the k-th ($k = 1, 2, \ldots, K$) attention map belonging to the k-th class, $\boldsymbol{X}_k^{\mathrm{att}} \in \mathbb{R}^{H \times W}$ denotes the attention-guided feature map for the k-th class, $\otimes$ represents the hadamard product, and $\mathrm{cat}(\cdot)$ is a concatenation operation. Reviewing the whole process, we can see that the attention-guided feature maps $\boldsymbol{X}^{\mathrm{att}} \in \mathbb{R}^{K \times H \times W}$ are class specific.

3.3 Training loss

Our goal of optimization is that if a certain class emerges in an image, the attention map corresponding to that class should have a higher activation value than other attention maps. Therefore, we introduce an attention loss as a regularization term, which is an auxiliary loss to enforce CAE module and CANet to better learn class-wise semantic information. To be specific, following the symbol definition in Subsection 3.2, we perform global max pooling operation on the attention maps $\boldsymbol{A}^{\mathrm{att}} \in \mathbb{R}^{K \times H \times W}$ as follows:

$$\begin{cases} v_k = \max_{i,j}\left(\boldsymbol{A}_{i,j,k}^{\mathrm{att}}\right), \\ \boldsymbol{v} = \left(v_k\right)_{1 \times K}, \end{cases} \tag{9}$$

where $\boldsymbol{v} \in \mathbb{R}^K$ is a feature vector used to compute attention loss. In brief, given a training dataset $\{\boldsymbol{I}_i, \boldsymbol{y}_i\}_{i=1}^Z$, where $\boldsymbol{I}_i$ is the i-th image, $\boldsymbol{y}_i = \left\{y_i^1, y_i^2, \ldots, y_i^K\right\}$ is its corresponding ground-truth label vector. The attention loss for multi-label classification can be formulated as

$$J_{\text{att}} = \sum_{i=1}^{Z}\sum_{j=1}^{K} y_i^j \log\left(v_i^j\right) + \left(1 - y_i^j\right)\log\left(1 - v_i^j\right), \tag{10}$$

where v_i^j represents the predicted probability of the i-th image belonging to the j-th category. As for single-label classification, the attention loss is formulated as

$$J_{\text{att}} = \sum_{i=1}^{Z}\sum_{j=1}^{K} y_i^j \log\left(\frac{v_i^j}{\sum_{j=1}^{K} v_i^j}\right). \tag{11}$$

For the visual recognition loss, denoted by J_{cls}, it changes with different tasks. Taking the tasks of multi-label classification and fine-grained visual classification as examples, which are also the tasks we perform in this paper, they share the same formulation as attention loss. However, they use different features for loss computation. The attention loss utilizes the features pooled from class-specific attention maps $\boldsymbol{A}^{\text{att}} \in \mathbb{R}^{K\times H\times W}$, while the features for visual classification loss are pooled from class-wise feature maps $\boldsymbol{X}^{\text{att}} \in \mathbb{R}^{K\times H\times W}$. Taking into account the visual recognition loss and the attention loss, the overall loss can be defined as follows:

$$J = J_{\text{cls}} + \lambda J_{\text{att}}, \tag{12}$$

where λ is a trade-off coefficient for balancing the overall loss. By jointly optimizing these two loss terms, we can better perform the tasks of visual recognition.

4 Experiments

4.1 Datasets

We evaluate our proposed CANet on two different image classification tasks including multi-label classification and fine-grained visual classification. The experiments are conducted on four widely-used datasets. They are the PASCAL VOC 2007 and 2012 datasets [13], the MS COCO dataset [14], and the CUB-200-2011 dataset [15]. The former three datasets are for multi-label classification and the fourth dataset is for fine-grained visual classification.

For the PASCAL VOC 2007 dataset, 5011 images are used for training and validation, and 4952 images are used for testing. The PASCAL VOC 2012 dataset contains 22531 images in total, of which 11540 images for training and validation, and 10991 images for testing.

The MS COCO dataset is another widely-used dataset for multi-label image classification. It contains 80 object classes, where 82081 images as the training set and 40504 images as the validation set.

The CUB-200-2011 dataset, containing a total of 11788 pictures, is used for fine-grained visual classification task. It is divided into two parts: a training set and a test set. The training set has 5994 images and the test set contains 5794 images.

4.2 Implementation details

We adopt ResNet-101 [2] as our base architecture to implement the CANet because of its excellent performance for visual recognition tasks. We employ SGD as an optimizer to train our network, with the momentum of 0.9 and the weight decay of 0.0001. The initial learning rate is set to 0.05 for the class-specific attention encoding module and 0.005 for the backbone. The number of epochs is set to 50. The learning rate is multiplied by a factor of 0.1 at the 12th, 25th, and 40th epochs, respectively. The batch size of each GPU is 16 with a total of 2 GPUs. During training, we randomly crop and resize the input images into 448×448. Also, we use random horizontal flips as data augmentation. Following other methods, when testing, the input images are resized into 576×576. And also, we use the model trained on MS COCO as the pre-train model for the PASCAL VOC dataset, so that our model can converge quickly.

4.3 Evaluation metrics

Following conventional settings [42], the average per-class precision (CP), average per-class recall (CR), average per-class F1 (CF1), the average overall precision (OP), average overall recall (OR), and the average overall F1 (OF1) are adopted as the performance evaluation metrics for the MS COCO dataset. Their formulations are as follows:

$$\begin{cases} \mathrm{OP} = \frac{\sum_{i=1}^{K} N_i}{\sum_{i=1}^{K} N_i^{\mathrm{p}}}, & \mathrm{CP} = \frac{1}{K}\sum_{i=1}^{K} \frac{N_i}{N_i^{\mathrm{p}}}, \\ \mathrm{OR} = \frac{\sum_{i=1}^{K} N_i}{\sum_{i=1}^{K} N_i^{\mathrm{gt}}}, & \mathrm{CR} = \frac{1}{K}\sum_{i=1}^{K} \frac{N_i}{N_i^{\mathrm{gt}}}, \\ \mathrm{OF1} = \frac{2\times \mathrm{OP}\times \mathrm{OR}}{\mathrm{OP}+\mathrm{OR}}, & \mathrm{CF1} = \frac{2\times \mathrm{CP}\times \mathrm{CR}}{\mathrm{CP}+\mathrm{CR}}, \end{cases} \tag{13}$$

where K denotes the number of object categories, N_i is the number of images which are correctly predicted for the i-th label, N_i^{p} is the number of predicted images for the i-th label, and N_i^{gt} is the number of ground truth images for the i-th label.

For each image, its labels are predicted as positive if the confidences are greater than 0.5. The above metrics of the Top-3 labels are also reported for fair comparisons. In addition, we also compute and report the mean average precision (mAP). For the PASCAL VOC dataset, we report average precision (AP) for per-class and mAP for all classes. Notably, we use Top-1 accuracy as a performance evaluation metric for the fine-grained visual classification task.

4.4 Ablation studies

In this subsection, we investigate the influence of the parameter N which is the number of atoms per class and the influence of the weight coefficient λ used to balance loss.

Influence of parameter N. In order to explore the effects of different values of N on the experimental results, we change its value in the set of $\{1, 2, 4, 6, 8, 10, 20\}$. The results of the ablation study on the PASCAL VOC 2007, PASCAL VOC 2012, MS COCO, and CUB-200-2011 datasets are reported in Figures 3(a) and (b). Note that, we set the weight coefficient λ to 1 for this ablation study.

As shown in Figure 3(a), (i) on the PASCAL VOC 2007 dataset, with the increase of N from 1 to 6, the accuracy measured in terms of mAP improves significantly from 87.1% to 94.3%. And then, we change N from 8 to 20, and the accuracy tends to be stable, just with very slight mAP gains from 94.5% to 94.8%. (ii) On the PASCAL VOC 2012 dataset, the mAP shows an upward trend with the number of N gradually increasing, which is similar to the trend on the PASCAL VOC 2007 dataset. (iii) On the MS COCO dataset, the mAP also shows a rising trend with the number of N gradually increasing. Specifically, with the increase of N from 1 to 10, the mAP improves significantly from 77.2% to 82.2%. As shown in Figure 3(b), on the CUB-200-2011 dataset, with N in the range of [1, 4], our method markedly boosts the Top-1 accuracy from 85.9% to 87.2%, but with the increase of N from 8 to 20, the accuracy just fluctuates merely from 87.9% to 88.1%. It can be seen that the performance varies with the number of atoms per class, and when the number of atoms reaches a certain value, the improvement will get stuck. One possible explanation for this phenomenon might be that a certain number of atoms are enough to encode category-specific semantic information. The performance tends to saturate when the number of atoms is larger than a certain value as over-complete representations will not bring remarkable gain to the performance. Consequently, considering the computation complexity, we set $N = 10$ in this work to balance the accuracy and the efficiency for all four datasets.

Influence of the weight coefficient λ**.** The weight coefficient λ is used to balance our proposed attention loss. We conduct ablation study on the PASCAL VOC 2007, PASCAL VOC 2012, MS COCO and CUB-200-2011 datasets by changing the values of λ in the set of $\{0, 0.1, 0.5, 1, 2, 10\}$. Here, the ablation study of λ was conducted with the parameter N set to 10.

The results of this ablation study are illustrated in Figures 3(c) and (d). It is worth noting that setting λ to 0 denotes that our CANet model is trained without using the attention loss (just with the conventional visual recognition loss). As shown in Figure 3(c), (i) on the PASCAL VOC 2007 dataset, with the increase of λ from 0 to 1, our proposed attention loss contributes substantially to the accuracy, from 92.7% to 94.7%. That is, our attention loss could boost the accuracy with 2.0% mAP. When we further increase the weight coefficient from 1 to 10, the mAP values drop marginally. (ii) On the PASCAL VOC 2012 dataset and the MS COCO dataset, the mAP shows an upward trend with λ gradually increasing in the range of [0, 1], while the mAP shows a down trend which is similar to the trend on the PASCAL VOC

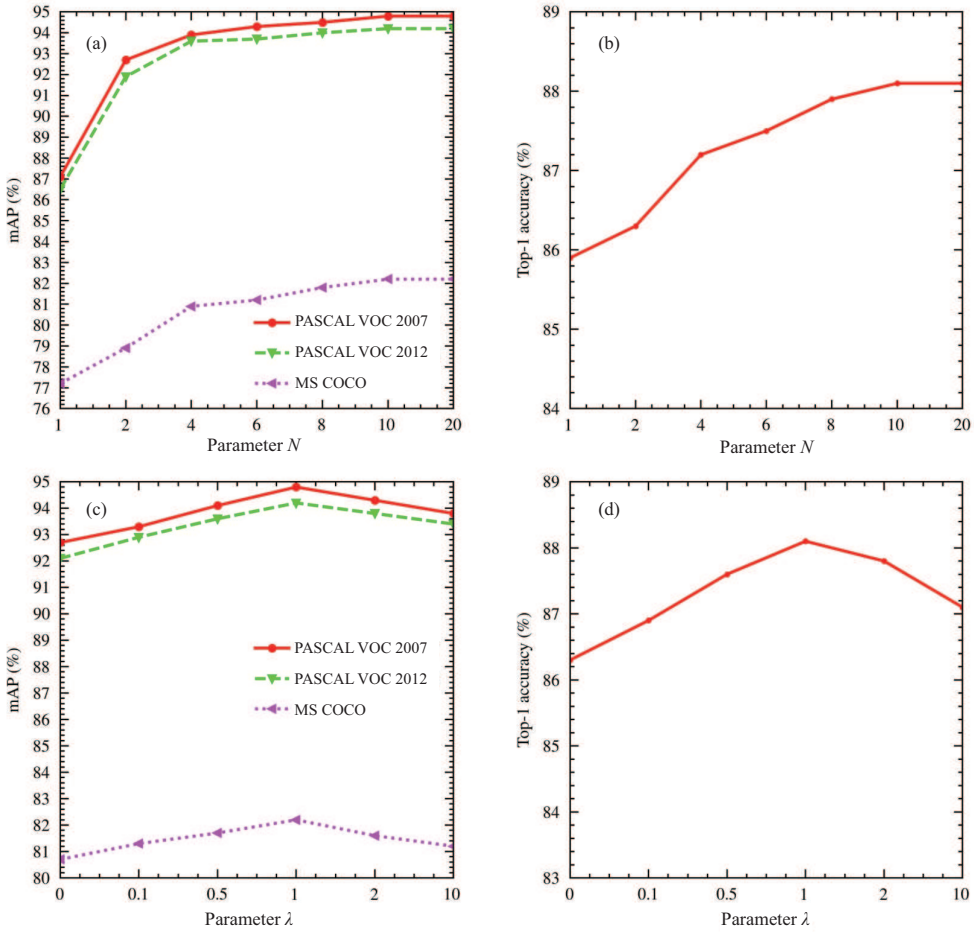

Figure 3 (Color online) The ablation studies with different values of N on the multi-label image classification datasets (a) and CUB-200-2011 dataset (b), and different values of λ on the multi-label image classification datasets (c) and CUB-200-2011 dataset (d).

2007 dataset. As shown in Figure 3(d), with the increase of λ from 0 to 1, the Top-1 accuracy improves from 86.3% to 88.1%. And then, we change λ from 1 to 10, the Top-1 accuracy value begins to fall. The best results are obtained with $\lambda = 1$ for all four datasets. We follow these parameter settings of $\lambda = 1$ and $N = 10$ on all datasets for subsequent experiments.

4.5 Comparisons with other current methods

Experimental results on the PASCAL VOC 2007 dataset. We compare our CANet with 12 multi-label image classification methods including CNN-RNN [39], RLSD [60], CoP [32], VeryDeep [30], ResNet-101 [2], FeV+LV [35], DSDL [37], RNN-Attention [41], ML-GCN [36], RARL [40], HCP [31], and RCP [33]. The comparison results are reported in Table 1. The AP values for each class and the mAP for all classes are represented. The previous attention-based method, RNN-Attention [41], introduced a spatial transformer layer to locate attentional regions on the convolutional maps, achieving an mAP of 91.9%. In contrast, we use the CAE module to obtain class-specific convolutional maps. Our proposed CANet achieves an mAP of 94.8%, which outperforms the RNN-Attention [41] method by 2.9%. We can also see that, compared with the baseline ResNet-101 model [2], our method obtains a big mAP gain of 4.0%. Also, our method outperforms the DSDL [37], a new method for multi-label image classification, by 0.4% in terms of mAP. Last but not least, our method shows good performance for most of the object categories including those challenging classes such as "bottle", "table", and "sofa".

Experimental results on the PASCAL VOC 2012 dataset. On the PASCAL VOC 2012 dataset, we also report the AP for each category and the mAP for all categories. The experimental results are represented in Table 2. Six representative methods are selected for comparison. They are RMIC [61],

Table 1 Comparisons of our CANet with other methods on the PASCAL VOC 2007 dataset. The best results are marked in bold.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
CNN-RNN [39]	96.7	83.1	94.2	92.8	61.2	82.1	89.1	94.2	64.2	83.6	70.0	92.4	91.7	84.2	93.7	59.8	93.2	75.3	99.7	78.6	84.0
RLSD [60]	96.4	92.7	93.8	94.1	71.2	92.5	94.2	95.7	74.3	90.0	74.2	95.4	96.2	92.1	97.9	66.9	93.5	73.7	97.5	87.6	88.5
VeryDeep [30]	98.9	95.0	96.8	95.4	69.7	90.4	93.5	96.0	74.2	86.6	87.8	96	96.3	93.1	97.2	70.0	92.1	80.3	98.1	87.0	89.7
ResNet-101 [2]	99.1	97.3	96.2	94.7	68.3	92.9	95.9	94.6	77.9	89.9	85.1	94.7	96.8	94.3	98.1	80.8	93.1	79.1	98.2	91.1	90.8
HCP [31]	98.6	97.1	98.0	95.6	75.3	94.7	95.8	97.3	73.1	90.2	80.0	97.3	96.1	94.9	96.3	78.3	94.7	76.2	97.9	91.5	90.9
RNN-Attention [41]	98.6	97.4	96.3	96.2	75.2	92.4	96.5	97.1	76.5	92.0	87.7	96.8	97.5	93.8	98.5	81.6	93.7	82.8	98.6	89.3	91.9
FeV+LV [35]	98.2	96.9	97.1	95.8	74.3	94.2	96.7	96.7	76.7	90.5	88.0	96.9	97.7	95.9	98.6	78.5	93.6	82.4	98.4	90.4	92.0
RARL [40]	98.6	97.1	97.1	95.5	75.6	92.8	96.8	97.3	78.3	92.2	87.6	96.9	96.5	93.6	98.5	81.6	93.1	83.2	98.5	89.3	92.0
RCP [33]	99.3	97.6	98.0	96.4	79.3	93.8	96.6	97.1	78.0	88.7	87.1	97.1	96.3	95.4	99.1	82.1	93.6	82.2	98.4	92.8	92.5
ML-GCN [36]	99.5	98.5	**98.6**	98.1	80.8	94.6	97.2	98.2	82.3	95.7	86.4	98.2	98.4	96.7	99.0	84.7	96.7	84.3	98.9	93.7	94.0
CoP [32]	**99.9**	98.4	97.8	**98.8**	81.2	93.7	97.1	**98.4**	82.7	94.6	87.1	98.1	97.6	96.2	98.8	83.2	96.2	84.7	99.1	93.5	93.8
DSDL [37]	99.8	98.7	98.4	97.9	81.9	95.4	97.6	98.3	**83.3**	95.0	88.6	98.0	97.9	95.8	99.0	**86.6**	95.9	86.4	98.6	**94.4**	94.4
CANet (ours)	99.6	**98.8**	97.7	97.9	**83.7**	**96.4**	**97.8**	98.0	82.4	**97.3**	**89.7**	**98.5**	**98.6**	**97.3**	**99.0**	84.6	**98.4**	**86.7**	**99.4**	94.1	**94.8**

Table 2 Comparisons of our CANet with other methods on the PASCAL VOC 2012 dataset. The best results are marked in bold.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
RMIC [61]	98.0	85.5	92.6	88.7	64	86.8	82.0	94.9	72.7	83.1	73.4	95.2	91.7	90.8	95.5	58.3	87.6	70.6	93.8	83.0	84.4
VeryDeep [30]	99.1	88.7	95.7	93.9	73.1	92.1	84.8	97.7	79.1	90.7	83.2	97.3	96.2	94.3	96.9	63.4	93.2	74.6	97.3	87.9	89.0
FeV+LV [35]	98.4	92.8	93.4	90.7	74.9	93.2	90.2	96.1	78.2	89.8	80.6	95.7	96.1	95.3	97.5	73.1	91.2	75.4	97	88.2	89.4
HCP [31]	99.1	92.8	97.4	94.4	79.9	93.6	89.8	98.2	78.2	94.9	79.8	97.8	97.0	93.8	96.4	74.3	94.7	71.9	96.7	88.6	90.5
RCP [33]	99.3	92.2	97.5	94.9	82.3	94.1	92.4	98.5	83.8	93.5	83.1	98.1	97.3	96.0	**98.8**	77.7	95.1	79.4	97.7	92.4	92.2
DSDL [37]	99.4	95.3	97.6	95.7	83.5	94.8	93.9	98.5	85.7	94.5	83.8	98.4	97.7	95.9	98.5	80.6	95.7	82.3	98.2	93.2	93.2
CANet (ours)	**99.7**	**95.4**	**98.1**	**96.2**	**85.0**	**95.4**	**94.7**	**98.8**	**85.7**	**97.4**	**85.7**	**98.8**	**99.0**	**96.4**	98.6	**83.3**	**98.2**	**85.0**	**98.4**	**93.5**	**94.2**

Table 3 Comparisons of our CANet with other methods on the MS COCO dataset. The best results are marked in bold.

Method	All							Top-3					
	mAP	CP	CR	CF1	OP	OR	OF1	CP	CR	CF1	OP	OR	OF1
CNN-RNN [39]	61.2	–	–	–	–	–	–	66.0	55.6	60.4	69.2	**66.4**	67.8
RNN-Attention [41]	–	–	–	–	–	–	–	79.1	58.7	67.4	84.0	63.0	72.0
Order-Free RNN [62]	–	–	–	–	–	–	–	71.6	54.8	62.1	74.2	62.2	67.7
ML-ZSL [63]	–	–	–	–	–	–	–	74.1	**64.5**	69.0	–	–	–
SRN [64]	77.1	81.6	65.4	71.2	82.7	69.9	75.8	85.2	58.8	67.4	87.4	62.5	72.9
ResNet-101 [2]	77.3	80.2	66.7	72.8	83.9	70.8	76.8	84.1	59.4	69.7	89.1	62.8	73.6
ML-GCN [36]	80.3	81.1	70.1	75.2	83.8	**74.2**	78.7	84.9	61.3	71.2	88.8	65.2	75.2
CoP [32]	81.1	81.2	**70.8**	75.8	83.6	73.3	78.1	86.4	62.9	72.7	88.7	65.1	75.1
DSDL [37]	81.7	84.1	70.4	**76.7**	85.1	73.9	**79.1**	88.1	62.9	**73.4**	89.6	65.3	**75.6**
CBAM [23]	80.3	86.5	63.5	73.2	**90.3**	67.4	77.2	89.2	57.9	70.2	**93.3**	62.5	74.9
SE-Net [18]	79.9	**87.4**	61.7	72.3	90.1	66.8	77.0	89.1	57.0	69.5	93.0	61.6	74.1
ECA-Net [65]	80.8	85.3	66.0	74.4	89.7	69.6	78.4	87.2	60.5	71.4	92.1	63.6	75.3
CANet (ours)	**82.2**	87.3	66.2	75.3	90.2	70.4	**79.1**	**89.5**	60.1	71.9	92.9	63.6	75.5

DSDL [37], VeryDeep [30], FeV+LV [35], HCP [31], and RCP [33]. It can be found from Table 2 that our method achieves an mAP of 94.2%, outperforming the competitive DSDL method [37] by 1%.

Experimental results on the MS COCO dataset. We also evaluate the performance of our method on the MS COCO dataset with different metrics, as depicted in Table 3 [62, 63]. Our method achieves the mAP of 82.2%, improving those of the previous best methods by 0.5%. Particularly, we pay attention to the comparison among attention-based methods. RNN-Attention [41] and SRN [64] are two popular attention-based methods for image classification: RNN-Attention takes advantage of visual attention mechanism to search local discriminative regions while SRN explores label relations based on the learned attention maps. Compared with these two attention-based methods, our method is simpler and overpasses them with all evaluation metrics. Moreover, our method also outperforms SE-Net [18], CBAM [23], and ECA-Net [65] in terms of mAP metric by 2.3%, 1.9%, and 1.4%, respectively. These results imply that our proposed approach is effective.

Table 4 Comparisons of our CANet with other methods on the CUB-200-2011 dataset. The baseline and baseline* follow the setting in SnapMix [67]. Here, baseline* indicates incorporating mid-level features in performance evaluation. The best results are marked in bold.

Method	Annotations	Top-1 accuracy (%)
PS-CNN [9]	✓	76.6
Part-RCNN [66]	✓	76.4
Mask-CNN [8]	✓	85.7
SPDA-CNN [45]	✓	85.1
STN [19]	×	84.1
Improved B-CNN [46]	×	85.8
MAMC [49]	×	86.2
OPAM [44]	×	85.8
Bilinear CNN [10]	×	84.1
PDFR [67]	×	84.5
AutoBD [68]	×	81.6
RACNN [47]	×	85.3
MACNN [48]	×	86.5
iSQRT-COV-Net with ResNet-101 [70]	×	88.7
ResNet-101 (baseline) [69]	×	85.7
ResNet-101 (baseline*) [69]	×	87.8
CANet (baseline+CAE)	×	88.1
CANet (baseline*+CAE)	×	**88.9**

Experimental results on the CUB-200-2011 dataset. The comparison results of our CANet with 14 FGVC approaches on the CUB-200-2011 dataset are presented in Table 4 [19,66–69] . As shown, our CANet obtains the Top-1 accuracy of 88.1%, which dramatically surpasses the baseline method with ResNet-101 [2] by 2.4%. In contract, the supervised method Mask-CNN [8] that uses both the object and the part-level annotations produces 85.7% Top-1 accuracy. Particularly, compared with several attention-based methods specifically designed for FGVC, including MAMC [49], MACNN [48], and RACNN [47], our proposed CANet outperforms them by 1.6%, 1.9%, and 2.8%, respectively. More notably, CANet is slightly lower than iSQRT-COV-Net with ResNet-101 [70] by 0.6%. A possible explanation for this result may be that global covariance pooling designed in iSQRT-COV-Net generates a richer representation. When CANet incorporates mid-level features, CANet obtains the Top-1 accuracy of 88.9%, which surpasses iSQRT-COV-Net with ResNet-101 by 0.2%.

4.6 Visualization

We visualize some class-specific attention maps in order to intuitively and qualitatively illustrate the effectiveness of our method. These example images are from the PASCAL VOC 2007 dataset. The original images and their corresponding class-specific attention maps are shown in Figure 4. The color of the bar in the most-right side denotes the intensity of the attention maps generated through our class-specific attention encoding module: dark red has the strongest activation while dark blue denotes the weakest intensity. We visualize the attention maps according to the object classes appearing in each image in Figure 4. As can be seen, each class-wise attention map can well locate the object instances belonging to the same class, no matter how many objects are included in the images, such as the cars and sheep in the second row, the persons in the fourth row and the bikes in the fifth row. Taking the last image in the first row as an example, our method can well capture the location of the "pottedplant" even if the image has a complex background.

Particularly, for each image in the third row containing multiple object classes, each attention map only activates its class-specific objects (with dark red color) and meanwhile suppresses the expression of other classes (with dark blue color). These results suggest that our CANet can indeed generate highly category-related attention maps with better interpretability, thus guiding the network to learn better feature representations, which are very consistent with our motivations.

Figure 4 (Color online) Example of class-specific attention maps. Here, dark red regions have the strongest activations while dark blue regions have the weakest intensity. The first row and the third row are the input images. Each image in the first row contains one object class, and each image in the third row includes multiple object classes. The second row shows the results of the images of the first row. The fourth and fifth rows present the results of the images of the third row. As can be seen, each class-wise attention map can well locate the object instances belonging to the same class, no matter how many objects are included in the images. Also, for the images containing multiple object classes, each attention map only activates its class-specific objects (with dark red color) and suppresses the expression of other classes (with dark blue color).

5 Conclusion

In this paper, we presented an attention method, termed CANet for image recognition. This is achieved by designing a CAE module, which introduces a class-specific dictionary to encode class-aware attention maps, to explicitly learn class-wise feature representations. For CANet training, we also presented the attention loss, an auxiliary loss to encourage CAE module and CANet to better learn class-wise semantic information. Extensive experiments conducted on several challenging datasets, including MS COCO, PASCAL VOC 2007, PASCAL VOC 2012, and CUB-200-2011 datasets, show the effectiveness of CANet. More notably, the proposed CAE module is easy to adopt and can be plugged into existing convolutional neural networks conveniently. Inspired by the feasibility of the CAE module on the classification task, we will explore the CAE's potential for other visual recognition tasks in the future like weakly supervised object localization and semantic segmentation.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61772425, U20B20-65), Shaanxi Science Foundation for Distinguished Young Scholars (Grant No. 2021JC-16), and Fundamental Research Funds for the Central Universities.

References

1 Cheng G, Gao D, Liu Y, et al. Multi-scale and discriminative part detectors based features for multi-label image classification. In: Proceedings of International Joint Conference on Artificial Intelligence, 2018. 649–655
2 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 770–778
3 Sun X, Cheng G, Pei L, et al. Query-efficient decision-based attack via sampling distribution reshaping. Pattern Recogn, 2022, 129: 108728
4 Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 1–9
5 Han J, Yao X, Cheng G, et al. P-CNN: part-based convolutional neural networks for fine-grained visual categorization. IEEE Trans Pattern Anal Mach Intell, 2022, 44: 579–590
6 Cheng G, Li R M, Lang C B, et al. Task-wise attention guided part complementary learning for few-shot image classification. Sci China Inf Sci, 2021, 64: 120104
7 Cheng G, Yang C, Yao X, et al. When deep learning meets metric learning: remote sensing image scene classification via learning discriminative CNNs. IEEE Trans Geosci Remote Sens, 2018, 56: 2811–2821
8 Wei X S, Xie C W, Wu J, et al. Mask-CNN: localizing parts and selecting descriptors for fine-grained bird species categorization. Pattern Recogn, 2018, 76: 704–714

9 Huang S, Xu Z, Tao D, et al. Part-stacked CNN for fine-grained visual categorization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 1173–1182
10 Lin T Y, RoyChowdhury A, Maji S. Bilinear CNN models for fine-grained visual recognition. In: Proceedings of the IEEE International Conference on Computer Vision, 2015. 1449–1457
11 You T, Zhang H, Zhang Y, et al. The influence of experienced guider on cooperative behavior in the Prisoner's dilemma game. Appl Math Comput, 2022, 426: 127093
12 Feng X, Han J, Yao X, et al. TCANet: triple context-aware network for weakly supervised object detection in remote sensing images. IEEE Trans Geosci Remote Sens, 2021, 59: 6946–6955
13 Everingham M, Eslami S M A, van Gool L, et al. The pascal visual object classes challenge: a retrospective. Int J Comput Vis, 2015, 111: 98–136
14 Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context. In: Proceedings of the European Conference on Computer Vision, 2014. 740–755
15 Wah C, Branson S, Welinder P, et al. The Caltech-UCSD Birds-200–2011 dataset. 2011. https://authors.library.caltech.edu/27452/
16 Guo H, Zheng K, Fan X, et al. Visual attention consistency under image transforms for multi-label image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 729–739
17 Guo J, Ma X, Sansom A, et al. SPANet: spatial pyramid attention network for enhanced image recognition. In: Proceedings of the IEEE International Conference on Multimedia and Expo, 2020. 1–6
18 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 7132–7141
19 Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks. In: Proceedings of Advances in Neural Information Processing Systems, 2015. 28: 2017–2025
20 Qin Z, Zhang P, Wu F, et al. FcaNet: frequency channel attention networks. 2020. ArXiv:2012.11879
21 Wang F, Jiang M, Qian C, et al. Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 3156–3164
22 Wang X, Girshick R, Gupta A, et al. Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 7794–7803
23 Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision, 2018. 3–19
24 Taniguchi F, Kudo M, Shimbo M. Estimation of class regions in feature space using rough set theory. In: Proceedings of the 1st International Conference on Conventional and Knowledge Based Intelligent Electronic Systems, 1997. 373–377
25 Ramirez I, Sprechmann P, Sapiro G. Classification and clustering via dictionary learning with structured incoherence and shared features. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010. 3501–3508
26 Yang M, Zhang L, Feng X, et al. Fisher discrimination dictionary learning for sparse representation. In: Proceedings of the International Conference on Computer Vision, 2011. 543–550
27 Zeiler M D, Fergus R. Visualizing and understanding convolutional networks. In: Proceedings of the European Conference on Computer Vision, 2014. 818–833
28 Prakash A, Storer J, Florencio D, et al. RePr: improved training of convolutional filters. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 10666–10675
29 Chu P, Bian X, Liu S, et al. Feature space augmentation for long-tailed data. In: Proceedings of the European Conference on Computer Vision, 2020. 694–710
30 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. ArXiv:1409.1556
31 Wei Y, Xia W, Lin M, et al. HCP: a flexible CNN framework for multi-label image classification. IEEE Trans Pattern Anal Mach Intell, 2016, 38: 1901–1907
32 Wen S, Liu W, Yang Y, et al. Multilabel image classification via feature/label co-projection. IEEE Trans Syst Man Cybern Syst, 2021, 51: 7250–7259
33 Wang M, Luo C, Hong R, et al. Beyond object proposals: random crop pooling for multi-label image recognition. IEEE Trans Image Process, 2016, 25: 5678–5688
34 Wei Y, Xia W, Huang J, et al. CNN: single-label to multi-label. 2014. ArXiv:1406.5726
35 Yang H, Zhou J T Y, Zhang Y, et al. Exploit bounding box annotations for multi-label object recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 280–288
36 Chen Z M, Wei X S, Wang P, et al. Multi-label image recognition with graph convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 5177–5186
37 Zhou F, Huang S, Xing Y. Deep semantic dictionary learning for multi-label image classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2020. 3572–3580
38 Li Q, Qiao M, Bian W, et al. Conditional graphical Lasso for multi-label image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 2977–2986
39 Wang J, Yang Y, Mao J, et al. CNN-RNN: a unified framework for multi-label image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 2285–2294
40 Chen T, Wang Z, Li G, et al. Recurrent attentional reinforcement learning for multi-label image recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2018. 6730–6737
41 Wang Z, Chen T, Li G, et al. Multi-label image recognition by recurrently discovering attentional regions. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. 464–472
42 Chen T, Xu M, Hui X, et al. Learning semantic-specific graph representation for multi-label image recognition. In: Proceedings of the IEEE International Conference on Computer Vision, 2019. 522–531
43 Dai J, Li Y, He K, et al. R-FCN: object detection via region-based fully convolutional networks. In: Proceedings of Advances in Neural Information Processing Systems, 2016. 379–387
44 Peng Y, He X, Zhao J. Object-part attention model for fine-grained image classification. IEEE Trans Image Process, 2018, 27: 1487–1500
45 Zhang H, Xu T, Elhoseiny M, et al. SPDA-CNN: unifying semantic part detection and abstraction for fine-grained recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 1143–1152
46 Lin T Y, Maji S. Improved bilinear pooling with CNNs. 2017. ArXiv:1707.06772
47 Fu J, Zheng H, Mei T. Look closer to see better: recurrent attention convolutional neural network for fine-grained image

recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 4438–4446
48 Zheng H, Fu J, Mei T, et al. Learning multi-attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. 5209–5217
49 Sun M, Yuan Y, Zhou F, et al. Multi-attention multi-class constraint for fine-grained image recognition. In: Proceedings of the European Conference on Computer Vision, 2018. 805–821
50 Hu T, Qi H, Huang Q, et al. See better before looking closer: weakly supervised data augmentation network for fine-grained visual classification. 2019. ArXiv:1901.09891
51 Jetley S, Lord N A, Lee N, et al. Learn to pay attention. 2018. ArXiv:1804.02391
52 Ji R, Wen L, Zhang L, et al. Attention convolutional binary neural tree for fine-grained visual categorization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020. 10468–10477
53 Mnih V, Heess N, Graves A, et al. Recurrent models of visual attention. In: Proceedings of Advances in Neural Information Processing Systems, 2014. 2204–2212
54 Papadopoulos D P, Clarke A D, Keller F, et al. Training object class detectors from eye tracking data. In: Proceedings of the European Conference on Computer Vision, 2014. 361–376
55 Xiao T, Xu Y, Yang K, et al. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 842–850
56 Ye J, He J, Peng X, et al. Attention-driven dynamic graph convolutional network for multi-label image recognition. In: Proceedings of the European Conference on Computer Vision, 2020. 649–665
57 Arandjelovic R, Gronat P, Torii A, et al. NetVLAD: CNN architecture for weakly supervised place recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 5297–5307
58 Zhang H, Dana K, Shi J, et al. Context encoding for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 7151–7160
59 Mairal J, Bach F, Ponce J, et al. Online dictionary learning for sparse coding. In: Proceedings of the 26th Annual International Conference on Machine Learning, 2009. 689–696
60 Zhang J, Wu Q, Shen C, et al. Multilabel image classification with regional latent semantic dependencies. IEEE Trans Multimedia, 2018, 20: 2801–2813
61 He S, Xu C, Guo T, et al. Reinforced multi-label image classification by exploring curriculum. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2018. 3183–3190
62 Chen S F, Chen Y C, Yeh C K, et al. Order-free RNN with visual attention for multi-label classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2018. 6714–6721
63 Lee C W, Fang W, Yeh C K, et al. Multi-label zero-shot learning with structured knowledge graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 1576–1585
64 Zhu F, Li H, Ouyang W, et al. Learning spatial regularization with image-level supervisions for multi-label image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 5513–5522
65 Wang Q, Wu B, Zhu P, et al. ECA-Net: efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020. 5177–5186
66 Zhang N, Donahue J, Girshick R, et al. Part-based R-CNNs for fine-grained category detection. In: Proceedings of the European Conference on Computer Vision, 2014. 834–849
67 Zhang X, Xiong H, Zhou W, et al. Picking deep filter responses for fine-grained image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 1134–1142
68 Yao H, Zhang S, Yan C, et al. AutoBD: automated bi-level description for scalable fine-grained visual categorization. IEEE Trans Image Process, 2018, 27: 10–23
69 Huang S, Wang X, Tao D. SnapMix: semantically proportional mixing for augmenting fine-grained data. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2020. 1628–1636
70 Wang Q, Xie J, Zuo W, et al. Deep CNNs meet global covariance pooling: better representation and generalization. IEEE Trans Pattern Anal Mach Intell, 2020, 43: 2582–2597