TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 02/20 pp644–655
DOI: 10.26599/TST.2023.9010022
Volume 29, Number 3, June 2024

Key-Part Attention Retrieval for Robotic Object Recognition

Jierui Liu, Zhiqiang Cao*, and Yingbo Tang

Abstract: The ability to recognize novel objects with a few visual samples is critical in the robotic applications. Existing methods mainly concern the recognition of inter-category objects, however, the object recognition from different sub-classes within the same category remains challenging due to their similar appearances. In this paper, we propose a key-part attention retrieval solution to distinguish novel objects of different sub-classes according to a few samples without re-training. Especially, an object encoder, including convolutional neural network with attention and key-part aggregation, is designed to generate object attention map and extract the object-level embedding, where object attention map from the middle stage of the backbone is used to guide the key-part aggregation. Besides, to overcome the non-differentiability drawback of key-part attention, the object encoder is trained in a two-step scheme, and a more stable object-level embedding is obtained. On this basis, the potential objects are located from a scene image by mining connected domains of the attention map. By matching the embedding of each potential object and embeddings from support data, the recognition of the potential objects is achieved. The effectiveness of the proposed method is verified by experiments.

Key words: key-part attention; retrieval; robotic object recognition

1 Introduction

Nowadays the robotic system has been widely used in daily life, and its ability to recognize objects is of importance in practical applications[1–3]. Different from the normal object recognition that is affordable to collect a large scale of training data[4, 5], the robotic object recognition often faces scenes with novel objects, and in some cases it is intractable to collect enough object samples for network training. Moreover, some robot tasks need to recognize objects of different appearances within the same category. These challenges increase the difficulty of object recognition for robots.

With the rapid development of Convolutional Neural Network (CNN)[6, 7], fruitful outcomes emerge for robotic object recognition. A typical solution is to achieve novel object classification using centroid-based concept learning[8], and then localization technology[4, 5] is combined to mine the novel objects in complex scenes[9–11]. These existing methods mainly concern learning from several samples of novel objects and the networks need to be re-trained with annotated novel instances. Still, it is challenging to distinguish objects of different subclasses within the same category due to the subtle differences among subclasses (e.g., varieties of birds). To address these issues, image retrieval provides reference and it devotes to evaluate the similarity of images by matching the corresponding image-level representations (hand-crafted[12] or CNN descriptor[13]). With hand-crafted features, such as Scale-Invariant Feature Transform (SIFT) and Speeded-Up Robust Features (SURF), a large-scale

• Jierui Liu, Zhiqiang Cao, and Yingbo Tang are with State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China. E-mail: liujierui2019@ia.ac.cn; zhiqiang.cao@ia.ac.cn; tangyingbo2020@ia.ac.cn.
* To whom correspondence should be addressed.
Manuscript received: 2022-09-06; revised: 2023-01-09; accepted: 2023-03-21

visual codebook is usually built. Then an embedding with fixed length is calculated and image retrieval is fulfilled by embedding matching. A possible problem of this solution is its weak image-level representation. CNN-based solutions become mainstream with the advantage of powerful feature extraction. Based on the extracted CNN feature maps, the advanced aggregators are designed to form image-level representation or embedding by considering global information[13, 14] or local information[15–18]. Although image retrieval is able to recognize the unseen image without extra training process, it cannot provide localization information for robotic object recognition. Furthermore, the differentiation of similar objects of different subclasses still needs to be further explored.

In this paper, a robotic object recognition method based on key-part attention retrieval is proposed. The main contributions are two-fold. Firstly, a novel object encoder based on key-part attention is designed to offer the refined object-level embedding for retrieval. The proposed object encoder is composed of CNN with attention and key-part aggregation. Specifically, object attention map is generated based on the middle stage of CNN backbone, which better guides the key-part aggregation to mine and aggregate the key local information of the object. As a result, high-quality object-level embedding is obtained. Considering that key-part aggregation will lead to the non-differentiability problem during network optimization, we solve it through a two-step training scheme with the supervision of image-level annotation. Attributing to the mining of key-part information, the slight differentiation of different subclass objects is explored. Secondly, different from object recognition with re-training[9–11], a robotic object recognition framework based on retrieval with object encoder is proposed, which recognizes novel objects of different subclasses according to a few visual samples without re-training. Particularly, the module of CNN with attention is multiplexed to locate the potential objects from a scene image. Each potential object and visual sample are then encoded into a potential embedding and a support embedding through the object encoder, respectively. By using embedding matching, similar objects of different subclasses within the same category are effectively recognized. The proposed method is easy to extend by simply expanding the visual samples and the experimental results prove its effectiveness.

The paper is organized as follows. Section 2 describes the related work. Section 3 introduces the proposed method in detail. The experiments are given in Section 4, and Section 5 concludes the paper.

2 Related Work

This section discusses the related work from the following two aspects: object recognition and image retrieval.

2.1 Object recognition

Krizhevsky et al.[19] proposed a pioneering deep network AlexNet, which can recognize the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. Simonyan and Zisserman[20] made a significant improvement on AlexNet by increasing depth using an architecture with 3×3 convolution filters. The aforementioned methods aim to recognize object with a large-scale annotated training data, which is not suitable to recognize novel objects with a few visual samples. Denninger and Triebel[1] proposed an object recognition method based on the random forest classifier, which learns a sustainable representation for new samples from a previously unseen class. Ayub and Wagner[8] developed a centroid-based concept learning network for robotic recognition, where a set of centroids are generated using clustering algorithm with CNN feature extractor for every new object class. To obtain the recognition results of novel objects in complex scenes, Dehghan et al.[10] combined a localization network faster R-CNN[4] with a matching module that uses the features from ResNet50[21] pre-trained on ImageNet. Valipour et al.[11] used Densecap[5] as the localization unit and novel objects are introduced to the robot for network re-training in a form of active human-robot interaction. Turkoglu et al.[9] introduced localization based on point cloud segmentation into robotic object recognition, which takes RGB and depth images as inputs and returns object bounding boxes with class labels. These methods run well on the recognition of novel classes, however, the differentiation of objects within the same category is rarely concerned.

2.2 Image retrieval

Image retrieval endeavors to find out similar images from a large database corresponding to a query image, which is beneficial to support robotic object classification with fine-grained representation. Hand-crafted features are used as the image representation,

where the feature descriptors are usually expected to be scale-invariant and robust to the viewpoints or illuminations, e.g., SIFT[22] and SURF[23]. After a large-scale visual codebook is built by using hierarchical or approximate cluster, an image is represented as a vector termed as embedding with fixed length, and image retrieval is achieved by embedding matching. Khan et al.[24] proposed a novel texture descriptor termed as directional magnitude local hexadecimal patterns for image retrieval. In Refs. [25, 26], texture, color, and shape features are combined together on the basis of modified local binary patterns. Such combination of features increases the feature discriminability, which is beneficial to improve retrieval performance.

In recent years, CNN-based image retrieval trained on a large-scale dataset achieves significant progress, where embeddings that are aggregated by CNN features play an important role. Azizpour et al.[13] applied max-pooling layer to obtain embeddings from the feature maps of last convolution layer. Yandex and Lempitsky[15] proposed an aggregation strategy termed as Sum-Pooling of Convolutional (SPoC) features to build compact global image descriptor. Tolias et al.[27] introduced Regional Maximum Activations of Convolutions (RMAC) to compute embeddings at different scales. Husain and Bober[28] further improved RMAC to control the aggregation process by explicitly employing regions discrimination, which is measured by their respective kullback-leibler divergence values. Radenovic et al.[14] designed Generalized-Mean pooling and image descriptor (GeM). Wei et al.[16] proposed a strategy of Selective Convolutional Descriptor Aggregation (SCDA), which discards the noisy background and keeps useful features for fine-grained retrieval. In the aforementioned methods, metric learning[29, 30] is employed to train the CNNs. Thus the resulting embeddings have the following attribute: distances between images within the same class are expected to be much smaller than those from different classes, which is favorable for embedding matching. However, affected by similar object appearance, viewpoints, and scales, existing image retrieval still remains challenges, especially for robotic object recognition with limited samples. In this paper, key-part information of objects is effectively mined to deal with challenges.

3 Methodology

This paper endeavors to recognize novel objects with a few examples and the proposed robotic recognition method is presented in Fig. 1, where the object encoder is the core and it includes two modules: CNN with attention and key-part aggregation. The former generates object attention map and CNN features, and the latter acquires the key-part regions from this

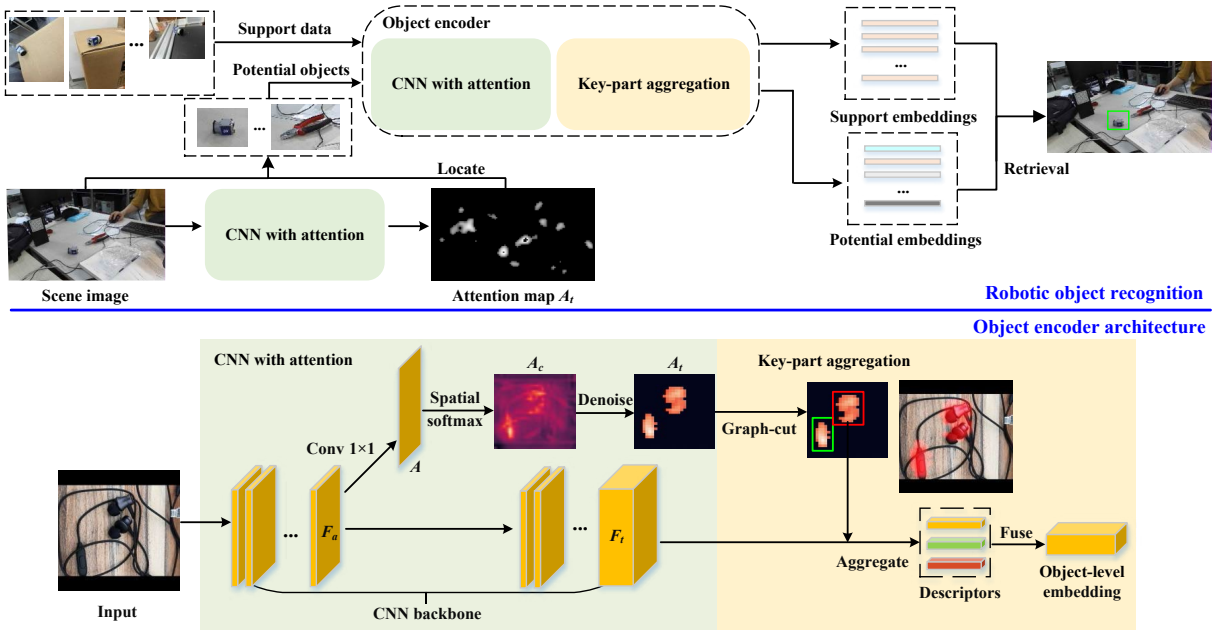

Fig. 1 Proposed robotic object recognition method with the detailed object encoder architecture. F_a and F_t are feature maps outputted by the middle and last stages of the CNN backbone, respectively. A, A_c, and A_t refer to object attention maps.

attention map, and then obtains the expected embeddings by feature aggregation. The whole object encoder is trained on a few images per class with image-level annotations.

For each novel object, a few examples are collected from different viewpoints and distances, which are sent to the object encoder in turn for support embeddings. To mine the novel object from a scene image, we firstly borrow the module of CNN with attention to obtain a corresponding attention map. On this map, the potential objects are located by extracting the connected domains. Then potential embeddings are obtained by feeding the potential objects cropped from the scene image into the object encoder. By matching each potential embedding with the support embeddings, the sub-class of each potential object consistent with the support data is judged and determined.

3.1 Object encoder

Attention mechanism plays an important role in the field of pattern recognition. Shu et al.[31] fully aggregated discriminative information through modal-wise expansion-squeeze-excitation attention and channel-wise expansion-squeeze-excitation attention. In this way, the recognition performance of elderly activities is significantly improved. In Ref. [32], a novel Skeleton-joint Co-attention Recurrent Neural Networks (SC-RNN) is proposed to predict human motion with the spatial coherence among joints and the temporal evolution among skeletons. Attention is also proved to be effective for person re-identification[33]. The successful applications of attention inspire us to enhance the feature discriminability of our object encoder with the help of attention.

A general object encoder extracts features using a CNN and aggregates the features for embeddings. The architecture of CNN backbone can be flexibly selected, such as ResNet101, ResNet50[21], and DenseNet169[34]. The CNN backbone is built with five cascaded CNN stages. For ResNet101 or ResNet50 backbone, these CNN stages are consistent with those in Ref. [21], and the details of DenseNet169 backbone are given in Table 1, where the input and output sizes, structure of each CNN stage are provided. The terms "Convolution", "Pooling", "Dense block", and "Transition layer" are directly from Ref. [34]. To further improve embeddings, an attention mechanism and key-part aggregation are introduced in this paper. As shown in Fig. 1, the feature map outputted from the middle stage of CNN is concerned due to that the feature maps from the lower stages of CNN are usually specific while those from higher stages are abstract. The chosen feature map F_a is then convolved by a 1×1 convolution to get an initial object attention map A. Taking the spatial relevance of each element in A into account, a spatial softmax operation is applied on A for object attention map A_c, which enables all the elements in A to be trained jointly,

$$A_{c(i,j)} = \frac{\alpha e^{A_{i,j}}}{\sum_{h=1}^{H}\sum_{w=1}^{W} e^{A_{h,w}}} \tag{1}$$

where $A_{i,j}$ and $A_{c(i,j)}$ denote the i-th row and j-th column of A and A_c, respectively, H and W are height and width of A. The scalar α is a learnable scale factor, which avoids vanishing gradient caused by a small $A_{c(i,j)}$. For object attention map A_c, it inevitably contains noise. Hence, a 3×3 mean blur denoising filter is applied followed by a thresholding operation, where the elements in A_c smaller than 0.75 times of the maximum value are set to be 0. After that, we get the final object attention map A_t.

It is worth mentioning that object attention map A_t can be regarded as a possibility map, where its each element records a value proportional to the probability that the element belongs to a key part. On this basis, key-part aggregation is designed. There are several means, such as graph-cut and k-means, to mine the key parts. In this paper, the graph-cut is chosen to better maintain the completeness of the key parts. The locations of non-zero elements in A_t constitute a set

Table 1 Details of DenseNet169 backbone[34].

Stage	Input size (pixel × pixel × channel)	Output size (pixel × pixel × channel)	Structure
1	512×512×3	256×256×64	Convolution
2	256×256×64	128×128×256	Pooling + Dense block 1
3	128×128×256	64×64×512	Transition layer 1 + Dense block 2
4	64×64×512	32×32×1280	Transition layer 2 + Dense block 3
5	32×32×1280	16×16×1664	Transition layer 3 + Dense block 4

$L:\{(x_i,y_i)\}_{i=1}^{l}$, where l denotes the number of non-zero elements and (x_i,y_i) refers to the coordinate of i-th element. Each element in set L is regarded as a node, and a graph G is built, where the edges are established by connecting each node to its nearest k nodes. The weight of the edge that connects $\text{node}_i\,(x_i,y_i)$ and $\text{node}_j\,(x_j,y_j)$ is given by $1/\text{dist}\,(\text{node}_i,\text{node}_j)$, where $\text{dist}\,(\text{node}_i,\text{node}_j)$ is their Euclidean distance. Afterwards, an adjacent matrix $W\in\mathbf{R}^{l\times l}$ is obtained according to G, where $W_{i,j}>0$ denotes $\text{node}_i\,(x_i,y_i)$ and $\text{node}_j\,(x_j,y_j)$ are connected directly. On this basis, the problem of mining the key parts in A_t is formulated as minimizing graph-cut problem. For convenient description, a dichotomy scheme is given as follows:

$$\begin{aligned}&\text{cut}\,(L_1,\,L_2)=\arg\min_{L_1,\,L_2}\sum_{i\in L_1,\,j\in L_2}W_{i,\,j},\\ \text{s.t.,}\quad &L_1\neq\varnothing,\,L_2\neq\varnothing,\,L_1\cap L_2=\varnothing,\\ &\text{and}\;L_1\cup L_2=L\end{aligned}\tag{2}$$

where the subsets L_1 and L_2 stand for two possible key-part regions. Unlike graph cuts stereo matching[35] that considers mini-cut as max-flow with an iterative solution, we solve Eq. (2) with more efficient spectral clustering[36]. If number of elements in a subset is too small, it means that this subset is meaningless, and a preferable processing is to replace it with all pixels of A_t. We regard each subset as a key-part region, which is represented by a bounding box bbox (t,l,b,r), where t, l, b, and r are the top, left, bottom, and right of the bounding box, respectively.

For each key-part region, it is aggregated on the feature map F_t outputted by CNN backbone to generate a region descriptor. Besides each key-part region, the global information is also aggregated on F_t to generate a global descriptor. We label the global and region descriptors as $g\,(F_t)$, $g\,(r_1)$, and $g\,(r_2)$, where $g(\cdot)$ denotes the GeM aggregator[14], r_1 and r_2 describe the cropped feature maps in F_t corresponding to key-part regions. Then, an object-level embedding is formed by fusing these three descriptors, which is given as follows:

$$v=\alpha_0\,g\,(F_t)+\alpha_1\,g\,(r_1)+\alpha_2\,g\,(r_2)\tag{3}$$

where α_0, α_1, and α_2 are learnable parameters to balance different descriptors.

The aforementioned graph-cut operation maintains the completeness of key parts, but it also brings in non-differentiability problem of key-part aggregation. Therefore, the gradient backpropagated from the loss function cannot be used to optimize the parameters of conv 1×1 during the generation of the initial object attention map A. Moreover, the effectiveness of the key-part aggregation also relies on whether object attention map can well capture the key-part information. This means that key-part aggregation and attention cannot be jointly trained. Herein, we design a two-step training scheme, where the weights of the CNN stages before the feature map F_a are always frozen for accelerating the training, as shown in Fig. 2. The first training step endeavors to obtain the weights of conv 1×1, whereas the second one trains the remaining weights of object encoder. In Fig. 2, the data streams shared in both two training steps are labeled in black, and the data streams only for the first training step and second training step are labeled in blue and red, respectively. For the first training step, the feature map F_a is multiplied with object attention map A_c to get the resulting feature map F_c, which is fed into the remaining CNN stages for the feature map F_t. F_t is further aggregated by Gem pooling and an embedding for an input image is acquired. With such a network, conv 1×1 is well trained and it keeps frozen after the first training step. This provides a stable object attention map for training of the second step. During

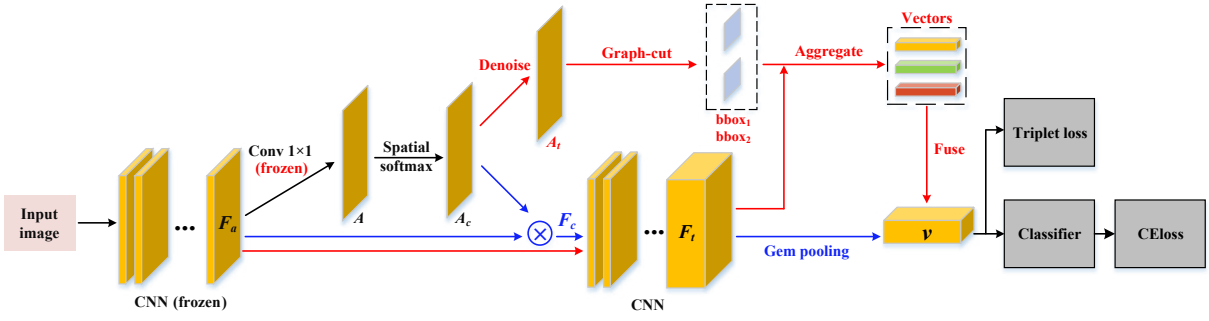

Fig. 2 Two-step training of the object encoder. Data streams shared in both two training steps are labeled in black, and only for the first and second training steps are labeled in blue and red, respectively.

the training process, both triplet loss and CEloss are applied[30].

3.2 Robotic object recognition

The object recognition pipeline is built on the basis of the trained object encoder. Take a novel sub-class into account, we get the support data with its several examples. For a scene image I_q, potential objects are captured by employing the module of CNN with attention on I_q. Then support data and cropped potential object images are sent into the object encoder to obtain support embeddings $\{E_a\}_{a=0}^{n-1}$ and potential embeddings $\{Q_b\}_{b=0}^{m-1}$, respectively, where n and m are the numbers of the support examples and potential objects, respectively. A similarity matrix S is calculated and its element is given as follows:

$$S_{b,a} = f(Q_b, E_a),$$
$$a = 0, 1, \ldots, n-1, b = 0, 1, \ldots, m-1 \quad (4)$$

where $S_{b,a}$ is the element of b-th row and a-th column of the matrix S; $f(\cdot)$ is a similarity function representing the relevance between Q_b and E_a,

$$f(Q_b, E_a) = \sqrt{\sum_{r=0}^{C-1} (Q_{b,r} E_{a,r})} \quad (5)$$

where $Q_{b,r}$ and $E_{a,r}$ mean the r-th elements of Q_b and E_a, respectively, and C is the dimension of an embedding. With the similarity matrix S, the probability p_b of the b-th potential object matching a specific novel object is assigned by

$$p_b = \max_{a=0,1,\ldots,n-1} S_{b,a} \quad (6)$$

Then, according to the matching probabilities $\{p_b\}_{b=0}^{m-1}$ of all potential objects, the object recognition is achieved. A potential object with the highest matching probability is classified into this novel sub-class (see Algorithm 1). The function CNNAttention $(\cdot)$ is used to generate an attention map. Mining $(\cdot)$ refers to the process of mining the attention map using the region connectivity analysis for potential objects. Crop $(\cdot)$ and ObjectEncoder $(\cdot)$ represent the image cropping of a potential object and image-to-embedding encoding function, respectively.

Algorithm 1 Robotic object recognition

Input: support data $\{d_a\}_{a=0}^{n-1}$ of a novel sub-class and a scene image I_q

Output: boxes $\{\text{box}_b\}_{b=0}^{m-1}$ and matching probabilities $\{p_b\}_{b=0}^{m-1}$

A_t = CNNAttention (I_q);
$\{\text{box}_b\}_{b=0}^{m-1}$ = Mining (A_t);
for $b \in [0, m-1]$ **do**
  po_b = Crop (I_q, box_b);
  Q_b = ObjectEncoder (po_b);
end for
for $a \in [0, n-1]$ **do**
  E_a = ObjectEncoder (d_a);
end for
Intilize $S_{b,a} = 0$, $b = 0, 1, \ldots, m-1$, $a = 0, 1, \ldots, n-1$;
for $b \in [0, m-1]$ **do**
  for $a \in [0, n-1]$ **do**
    Calculate $S_{b,a}$ using Eq. (4);
  end for
  Calculate p_b using Eq. (6);
  if $p_b \geqslant 0.5$ **then**
    The b-th potential object belongs to the novel sub-class;
  end if
end for
return

4 Experiment

In this section, the key-part aggregated retrieval performance of the proposed object encoder is evaluated on three datasets: CUB200-2011[37], Stanford-Cars[38], and INRIA Holidays[39]. Furthermore, the proposed method is also testified on an active three-camera platform.

4.1 Evaluation on public datasets

The object encoder is used to generate embedding for retrieval, which is verified on three public datasets:

- CUB200-2011[37] with 5934 train images, 600 validation images, and 5794 test images in 200 sub-classes of the category bird.
- Stanford-Cars[38] with 16 185 images in 196 sub-classes of the category car.
- INRIA Holidays[39] with 1491 images of different landmarks.

Note that INRIA Holidays dataset consists of 500 queries and 991 gallery images without training images. We have to resort to a subset of Google Landmarks Dataset v2[40] to train our network. Also, CUB200-2011 and Stanford-Cars are datasets for fine-grained classification about bird and car, respectively, and we regard each test image as query and all remaining images as gallery. In our method, the backbone is firstly initialized with weights pre-trained on ImageNet, and then it is further trained on the

dataset. For different datasets, the CNN backbone is trained to achieve better performance. Notice that we reproduce other advanced methods and compare them with the proposed method on the same training setting. All methods run on a server with i9-10400 CPU and NVIDIA RTX 3090 GPU. We evaluate the performance of methods with two metrics: mean Average Precision (mAP) and top-1 mean Precision (top-1 mP), where Average Precision (AP) refers to the average of precision values at the ranks of all true positives among the retrieved items, and then mAP is obtained by computing the mean of AP over a large set of queries[41].

(1) Ablation studies

As mentioned above, the proposed object encoder termed as KANet is built on a CNN backbone. We consider DenseNet169[34], ResNet50, and ResNet101[21] separately as the backbone of KANet in ablation experiments. We firstly determine which CNN stage to output the feature map F_a required by the attention block. Table 2 provides comparison of using feature map from different CNN stages for attention on the basis of DenseNet169 backbone. It is seen that CNN stage4 performs better, which indicates the reasonability of our method.

Besides, KANet and its two variants are concerned according to whether attention block, key-part aggregation (k-means), and key-part aggregation (graph-cut) are adopted, where the scheme of key-part aggregation (k-means) refers to that the key-part regions are obtained by k-means instead of graph-cut. The ablation results are shown in Table 3. Meanwhile, three types of backbones are considered. One can see that the combination of attention block and key-part aggregation stably improves retrieval performance for different backbones. Moreover, our adopted key-part region based on graph-cut is slightly better than the k-means solution.

Table 2 Comparison of using feature map from different CNN stages for attention.

CNN stage for attention	CUB200-2011	
	Top-1 mP	mAP
None	0.800 93	0.713 95
CNN Stage 3	0.785 00	0.704 09
CNN Stage 4	**0.825 00**	**0.732 91**
CNN Stage 5	0.812 50	0.729 44

(2) Comparison with existing methods

The comparison of different aggregators with the DenseNet169 backbone on three public datasets in terms of mAP and top-1 mP is shown in Table 4. The methods include:

- **SPoC**[15]**:** sum-pooling of convolutional features;
- **Mac**[13]**:** max-pooling of convolutional features;
- **RMAC**[27]**:** regional maximum activations of convolutional features;
- **GeM**[14]**:** generalized-mean pooling of convolutional features;
- **SCDA**[16]**:** selective convolutional descriptor aggregation of convolutional features;
- **Non-Local**[42] **+ GeM:** GeM with non-local attention module;
- **RGA**[33] **+ GeM:** GeM with RGA attention module.

All data are from our reproduction and it is observed that our proposed KANet achieves 5 best results over 6 trials. Figure 3 presents visualization of object attention maps and key-part regions for six images from three datasets, where the first to fourth columns of each sub-figure represent the rescaled image, the initial object attention map A, the final object attention map A_t, and key-part regions, respectively. The results indicate that KANet can effectively extract the key-part regions with better completeness.

Table 3 Comparison of the proposed object encoder KANet with different variants on CUB200-2011 in terms of mAP.

Object encoder	Attention block	Key-part aggregation (k-means)	Key-part aggregation (graph-cut)	mAP
KANet-I with DenseNet169	√	×	×	0.716 54
KANet-II with DenseNet169	√	√	×	0.724 13
KANet with DenseNet169	√	×	√	**0.732 91**
KANet-I with ResNet101	√	×	×	0.699 38
KANet-II with ResNet101	√	√	×	**0.733 16**
KANet with ResNet101	√	×	√	0.732 76
KANet-I with ResNet50	√	×	×	0.683 54
KANet-II with ResNet50	√	√	×	0.707 87
KANet with ResNet50	√	×	√	**0.716 44**

Table 4 Comparison of different aggregators with DenseNet169 backbone on three datasets in terms of top-1 mP and mAP.

Aggregator	CUB200-2011		Stanford-Cars		INRIA Holidays	
	top-1 mP	mAP	top-1 mP	mAP	top-1 mP	mAP
SPoC[15]	0.778 91	0.682 59	0.916 18	0.835 95	0.898 00	0.866 89
Mac[13]	0.798 41	0.706 27	0.916 43	0.849 21	0.908 00	0.883 53
RMAC[27]	0.798 58	0.719 89	0.917 80	0.864 14	0.922 00	0.905 49
GeM[14]	0.800 93	0.713 95	0.918 12	0.863 19	0.928 00	0.907 06
SCDA[16]	0.822 93	0.729 99	0.914 14	0.863 26	0.922 00	0.901 88
Non-Local[42] + GeM	0.811 71	0.719 58	0.909 79	0.850 04	0.918 00	0.906 20
RGA[33] + GeM	0.814 64	0.722 99	0.915 88	0.852 67	0.936 00	**0.920 81**
Ours	**0.825 00**	**0.732 91**	**0.923 47**	**0.873 79**	**0.938 00**	0.915 97

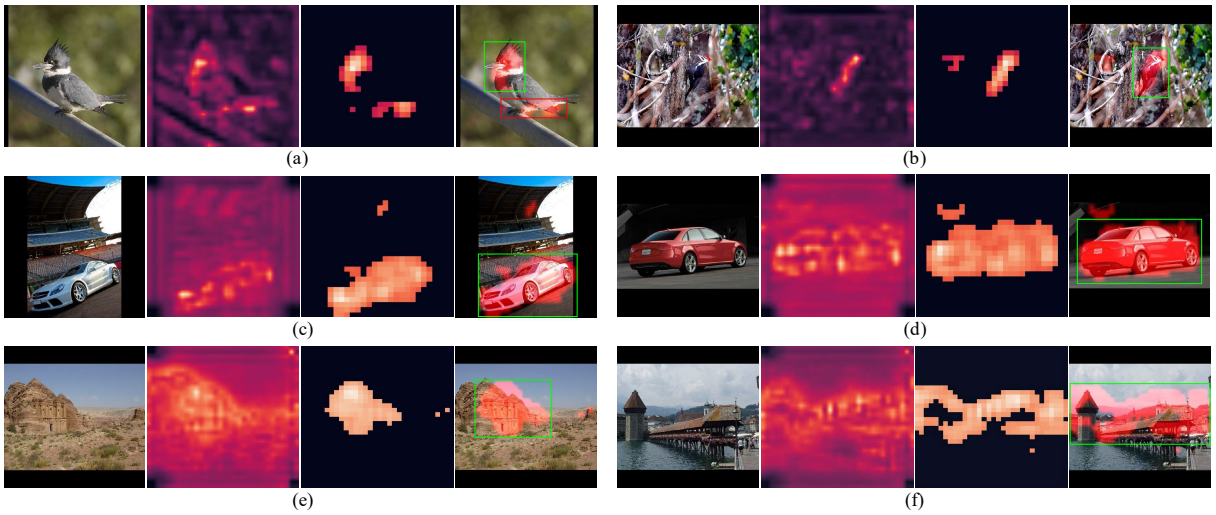

Fig. 3 Visualizations of object attention maps and key-part regions for six images on three datasets. (a) and (b) CUB200-2011[37], (c) and (d) Standford-Cars[38], and (e) and (f) INRIA Holidays[39]. The first to fourth columns of each sub-figure refer to the rescaled image, the initial object attention map, the final object attention map, and key-part regions, respectively.

4.2 Experiments on active three-camera platform

In this section, experiments are executed on our active three-camera platform with an RGBD camera (Cam_1) and two RGB cameras (Cam_2 and Cam_3), as illustrated in Fig. 4. Cam_1 can pitch and deflect with the whole platform driven by motors. Cam_2 and Cam_3 can rotate around their respective rotation axis driven by motors. ResNet50[21] is used as the backbone in KANet. To better train the object encoder, we collect training data about 3000 classes from the internet with average 15 images per class.

Distinguishing different sub-class objects. In this experiment, we consider three different cups and three different bottles, where each object corresponds to a sub-class. Each sub-class is captured in three different viewpoints, as shown in Fig. 5. Therefore, we have 18 images altogether.

Figure 6 presents the matching probabilities between each sub-class in Fig. 5 and the support data from other viewpoints, where the horizon axis corresponds to the sub-class and the vertical axis refers to the support data. Figure 6(a) gives the 1-shot results where the

Fig. 4 Active three-camera platform.

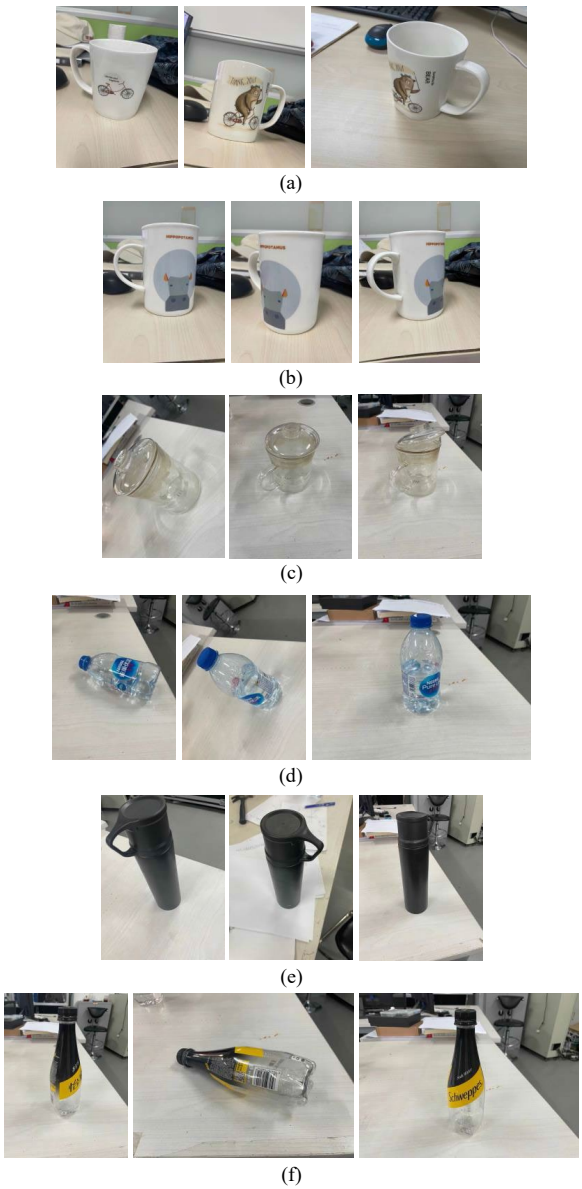

Fig. 5 Six sub-classes with three cups and three bottles. (a) cup1, (b) cup2, (c) cup3, (d) bottle1, (e) bottle2, and (f) bottle3.

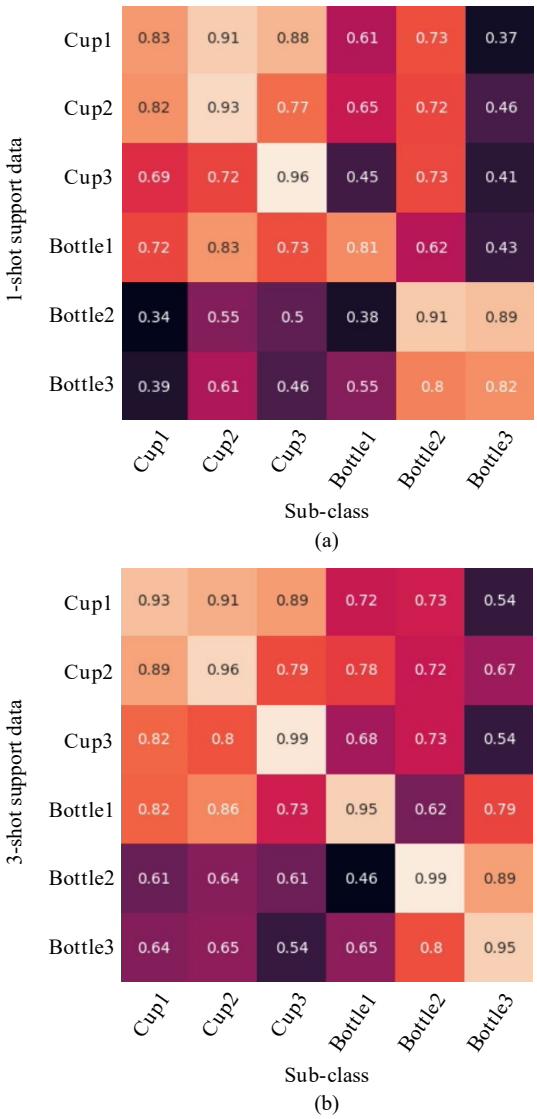

Fig. 6 Matching probabilities between sub-classes in Fig. 5 and the support data from other viewpoints, where the horizon axis corresponds to the sub-class and the vertical axis refers to the support data. (a) 1-shot results and (b) 3-shot results.

support data are composed of one image per sub-class and only minimum matching probability per sub-class is provided. Figure 6b demonstrates the 3-shot results and each of which is calculated by maximizing the minimum matching probability per sub-class among 3-shot support data. Compared to Fig. 6a with 1-shot, the results of Fig. 6b with 3-shot indicate that the matching probabilities within the same object (see diagonal elements) increase rapidly, which means that the performance of distinguishing the unique objects can be significantly improved.

Take the images in Fig. 5 as the support data, the processed results for a given image are shown in Fig. 7, where circle box instead of rectangle bounding box is used. It is seen that all objects related to the support data are recognized correctly (see green circular boxes), even the object bottle3 is placed upside down. This verifies the effectiveness of our method. Besides, the efficiency is discussed, where the processing speed varies with different objects number in a scene image. With input resolution of 1280 pixel × 720 pixel and nine potential objects, the inference speed is about 11 Frames Per Second (FPS), which indicates that the proposed method can run in real time.

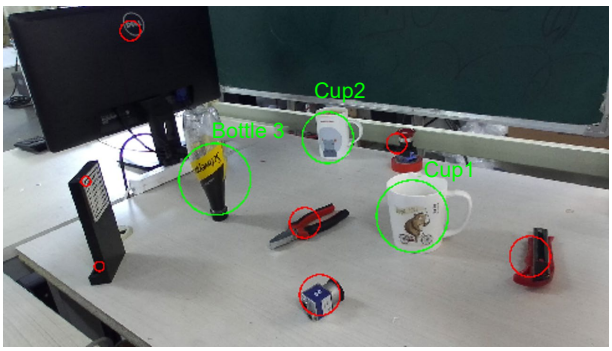

Fig. 7 Object recognition results, where the potential objects are marked in red circular boxes.

Object recognition and gazing experiments. We consider recognition of novel object plier. Meanwhile, each camera in the active three-camera platform is required to gaze at the latest novel object by keeping the object in the center of its visual field.

Initially, the plier is only regarded as a potential object, as shown in Fig. 8a. After 7 images of the object plier are provided as support data (see the first row of Fig. 8b), the proposed method recognizes this plier correctly and all cameras adjust their respective postures to gaze at this plier, which are illustrated in the second row of Fig. 8b.

5 Conclusion

This paper proposes a robotic object recognition method based on key-part attention retrieval to know novel objects through a few samples. An object encoder is designed to generate object attention map and extract the object-level embedding, where object attention map from the middle stage of the CNN backbone is used to better guide the key-part aggregation. Besides, a two-step training scheme is developed for the object encoder to output a more stable embedding. After each potential object is mined with the attention map of scene image, its corresponding embedding is matched with support embeddings, and then object recognition is achieved. The experiment results indicate that the proposed method can effectively recognize different subclass objects in the same category with good adaptability to viewpoints and poses. For distant object, it shall lead to tiny pixel area in scene image, which makes it difficult to be recognized. In the future, we shall incorporate multi-scale attention into the proposed method for

(a)

(b)

Fig. 8 Recognition and gazing of plier with 7-shot support data. (a) Potential objects (labeled in red circular boxes) in Cam_1 as well as the initial viewpoints of Cam_2 and Cam_3, (b) 7 images of the novel object plier (the first row) and its recognition result marked in a green circular box (the second row), and the red circular boxes still refer to potential objects, where each camera keeps this object in the center of its visual field.

better recognition of tiny object.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Nos. 62073322 and 61973302), the CIE-Tencent Robotics X Rhino-Bird Focused Research Program (No. 2022-07), and the Beijing Natural Science Foundation (No. 2022MQ05).

References

[1] M. Denninger and R. Triebel, Persistent anytime learning of objects from unseen classes, in *Proc. 2018 IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, Madrid, Spain, 2018, pp. 4075–4082.

[2] B. Xiong and X. Ding, A generic object detection using a single query image without training, *Tsinghua Science and Technology*, vol. 17, no. 2, pp. 194–201, 2012.

[3] L. Sun, J. Ma, and L. Jing, Object counting using a refinement network, *Tsinghua Science and Technology*, vol. 27, no. 5, pp. 869–879, 2022.

[4] S. Ren, K. He, R. Girshick, and J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.

[5] J. Johnson, A. Karpathy, and F. F. Li, DenseCap: Fully convolutional localization networks for dense captioning, in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 4565–4574.

[6] R. Xin, J. Zhang, and Y. Shao, Complex network classification with convolutional neural network, *Tsinghua Science and Technology*, vol. 25, no. 4, pp. 447–457, 2020.

[7] M. H. Haghighat and J. Li, Intrusion detection system using voting-based neural network, *Tsinghua Science and Technology*, vol. 26, no. 4, pp. 484–495, 2021.

[8] A. Ayub and A. R. Wagner, Tell me what this is: Few-shot incremental object learning by a robot, in *Proc. 2020 IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, Las Vegas, NV, USA, 2020, pp. 8344–8350.

[9] M. O. Turkoglu, F. B. Ter Haar, and N. van der Stap, Incremental learning-based adaptive object recognition for mobile robots, in *Proc. 2018 IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, Madrid, Spain, 2018, pp. 6263–6268.

[10] M. Dehghan, Z. Zhang, M. Siam, J. Jin, L. Petrich, and M. Jagersand, Online object and task learning via human robot interaction, in *Proc. 2019 Int. Conf. Robotics and Automation*, Montreal, Canada, 2019, pp. 2132–2138.

[11] S. Valipour, C. Perez, and M. Jagersand, Incremental learning for robot perception through HRI, in *Proc. 2017 IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, Vancouver, Canada, 2017, pp. 2772–2777.

[12] Y. Shinagawa, Homotopic image pseudo-invariants for openset object recognition and image retrieval, *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. 30, no. 11, pp. 1891–1901, 2008.

[13] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, From generic to specific deep representations for visual recognition, in *Proc. 2015 IEEE Conf. Computer Vision and Pattern Recognition Workshops*, Boston, MA, USA, 2015, pp. 36–45.

[14] F. Radenovic, G. Tolias, and O. Chum, Fine-tuning CNN image retrieval with no human annotation, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1655–1668, 2019.

[15] A. B. Yandex and V. Lempitsky, Aggregating local deep features for image retrieval, in *Proc. 2015 IEEE Int. Conf. Computer Vision*, Santiago, Chile, 2015, pp. 1269–1277.

[16] X. S. Wei, J. H. Luo, J. Wu, and Z. H. Zhou, Selective convolutional descriptor aggregation for fine-grained image retrieval, *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2868–2881, 2017.

[17] L. Ma, X. Li, Y. Shi, J. Wu, and Y. Zhang, Correlation filtering-based hashing for fine-grained image retrieval, *IEEE Signal Process. Lett.*, vol. 27, pp. 2129–2133, 2020.

[18] Y. Li, Y. Xu, J. Wang, Z. Miao and Y. Zhang, MS-RMAC: Multiscale regional maximum activation of convolutions for image retrieval, *IEEE Signal Process. Lett*., vol. 24, no. 5, pp. 609–613, 2017.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ImageNet classification with deep convolutional neural networks, *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[20] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv: 1409.1556, 2015.

[21] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 770–778.

[22] D. G. Lowe, Distinctive image features from scale-invariant keypoints, *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[23] H. Bay, T. Tuytelaars, and L. Van Gool, SURF: Speeded up robust features, in *Proc. 9th European Conf. Computer Vision*, Graz, Austria, 2006, pp. 404–417.

[24] A. Khan, A. Javed, M. T. Mahmood, M. H. A. Khan, and I. H. Lee, Directional magnitude local hexadecimal patterns: A novel texture feature descriptor for content-based image retrieval, *IEEE Access*, vol. 9, pp. 135608–135629, 2021.

[25] F. Tajeripour, M. Saberi, and S. Fekri-Ershad, Developing a novel approach for content based image retrieval using modified local binary patterns and morphological transform, *Int. Arab J. Inf. Technol.*, vol. 12, no. 6, pp. 574–581, 2015.

[26] N. Kayhan and S. Fekri-Ershad, Content based image retrieval based on weighted fusion of texture and color features derived from modified local binary patterns and local neighborhood difference patterns, *Multimed. Tools Appl.*, vol. 80, no. 21, pp. 32763–32790, 2021.

[27] G. Tolias, R. Sicre, and H. Jégou, Particular object retrieval with integral max-pooling of CNN activations, in *Proc. 4th Int. Conf. Learning Representations*, San Juan, Puerto Rico, doi: 10.48550/arXiv.1511.05879.

[28] S. S. Husain and M. Bober, REMAP: Multi-layer entropy-guided pooling of dense CNN features for image retrieval,

IEEE Trans. Image Process., vol. 28, no. 10, pp. 5201–5213, 2019.

[29] X. He, Y. Zhou, Z. Zhou, S. Bai, and X. Bai, Triplet-center loss for multi-view 3D object retrieval, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 1945–1954.

[30] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, A strong baseline and batch normalization neck for deep person re-identification, *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2597–2609, 2020.

[31] X. Shu, J. Yang, R. Yan, and Y. Song, Expansion-squeeze-excitation fusion network for elderly activity recognition, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5281–5292, 2022.

[32] X. Shu, L. Zhang, G. J. Qi, W. Liu, and J. Tang, Spatiotemporal co-attention recurrent neural networks for human-skeleton motion prediction, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3300–3315, 2022.

[33] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, Relation-aware global attention for person re-identification, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 3183–3192.

[34] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, Densely connected convolutional networks, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 2261–2269.

[35] V. Kolmogorov, P. Monasse, and P. Tan, Kolmogorov and Zabih's graph cuts stereo matching algorithm, *Image Process. Line*, vol. 4, pp. 220–251, 2014.

[36] A. Y. Ng, M. I. Jordan, and Y. Weiss, On spectral clustering: Analysis and an algorithm, in *Proc. 14th Int. Conf. Neural Information Processing Systems*, Vancouver, Canada, 2011, pp. 849–856.

[37] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, *The Caltech-UCSD Birds-200–2011 Dataset*. Pasadena, CA, USA: California Institute of Technology, 2011.

[38] J. Krause, M. Stark, J. Deng, and F. F. Li, 3D object representations for fine-grained categorization, in *Proc. 2013 IEEE Int. Conf. Computer Vision Workshops*, Sydney, Australia, 2013, pp. 554–561.

[39] H. Jegou, M. Douze, and C. Schmid, Improving bag-of-features for large scale image search, *Int. J. Comput. Vis.*, vol. 87, no. 3, pp. 316–336, 2010.

[40] T. Weyand, A. Araujo, B. Cao, and J. Sim, Google landmarks dataset v2-A large-scale benchmark for instance-level recognition and retrieval, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 2572–2581.

[41] The INRIA Holidays dataset, https://lear.inrialpes.fr/~jegou/data.php', 2022.

[42] X. Wang, R. Girshick, A. Gupta, and K. He, Non-local neural networks, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 7794–7803.

Jierui Liu received the BEng degree from South China University of Technology, China in 2019. He is currently a PhD candidate in control theory and control engineering at Institute of Automation, Chinese Academy of Sciences, Beijing, China. His current research interests include visual measurement and service robot.

Zhiqiang Cao received the BEng degree in industrial automation and the MEng degree in control theory and control engineering from Shandong University of Technology, China in 1996 and 1999, respectively. In 2002, he received the PhD degree in control theory and control engineering from Institute of Automation, Chinese Academy of Sciences, Beijing, China. He is currently a professor at Institute of Automation, Chinese Academy of Sciences. His research interests include service robots and intelligent robot.

Yingbo Tang received the BEng degree from North China Electric Power University, China in 2020. She is currently a PhD candidate in control theory and control engineering at Institute of Automation, Chinese Academy of Sciences, Beijing, China. Her research interests include object detection and robotic grasping.