



## Clustering-inspired Channel Selection Method for Weakly Supervised Object Localization

Xiaofeng Wang<sup>a,\*\*</sup>, Zhe Liu<sup>a</sup>, Xiangru Qiao<sup>a</sup>, Zhiquan Li<sup>a</sup>, Sidong Wu<sup>a</sup>, Jiao Zhang<sup>a</sup>, Yonghuai Liu<sup>b</sup>, Zhan Li<sup>a,\*\*</sup>, Hongbo Guo<sup>c</sup>, Huaizhong Zhang<sup>b</sup>

<sup>a</sup>School of Information Science and Technology, North West University, Xi'an, 710127, China

<sup>b</sup>Department of Computer Science, Edge Hill University, Ormskirk, UK

<sup>c</sup>ZTE Corporation, Xi'an, China

### ABSTRACT

Weakly Supervised Object Localization (WSOL) aims to utilize the features learned by a classifier on the image-level labels to locate target objects. However, these existing channel selection methods for WSOL still cannot effectively select the important channels and remove the unimportant ones. To address this issue, we propose a Clustering-inspired Channel Selection method based on Class Activation Maps (CCS-CAM). Compared with the traditional methods, the advantage of CCS-CAM is that it is very simple yet effective for channel selection due to the K-means clustering based on Class Activation Maps. It can effectively ensure both object localization and classification accuracy. The effectiveness of the proposed CCS-CAM method has been demonstrated using multiple public datasets, with GT-Know Loc reaching 87.9% and 63.71% on the CUB200-2011 and ImageNet-1k, which is superior to the other state-of-the-art Methods.

© 2024 Elsevier Ltd. All rights reserved.

### 1. Introduction

Object localization [1] aims to learn and locate the objects in images, which has made great progress with the advent of convolutional neural network [2]. But, the traditional object localization needs not only the image-level labels but also accurate position-level labels or the pixel-level labels, which hinder its real application due to the high annotation complexity. On the contrary, Weakly Supervised Object Localization (WSOL) needs only image-level labels, and thus greatly reduces its annotation complexity. As a result, it receives intensive attentions [2, 3, 4, 5].

The most common method for WSOL is based on the Class Activation Mapping [1](CAM), which is fine-tuned on the pre-trained classification network by using the Global Average

Pooling layer (GAP) to replace the full connection layer in the Convolutional Neural Network (CNN). The previous work [2] has shown that CAM often fails to capture the entire target object area, which is typically smaller than the actual one. The reason is that CAM only learns the object itself, or the most discriminative or representative parts of the object, that is most relevant to classification. As illustrated in Fig.1, CAM tends to focus only on the bird head, as shown in Fig.1(b), while ignoring other less discriminative regions such as the bird body, which is well detected with our method CCS-CAM in Fig.1(c), leading to relatively small and inaccurate positioning areas.

Many methods have been proposed to solve the above mentioned problem. For example, [3, 4, 5] attempt to expand the class identification region, enabling these networks to learn other regions with a relatively weak correlation with the target object. However, expanding the identification region usually reduces the object classification performance. In addition, [6] and [7] propose two different methods to complete object localization using category-independent determination areas. These methods [6, 7] use a pretrained model to extract image features for positioning, without optimizing different categories of target objects. [6] shows poor performance on these images with complex backgrounds, while [7] requires additional data to help locate objects. Moreover, [8, 9] adopt feature addition to acquire

\*\*Corresponding author

*e-mail:* xfwang@nwu.edu.cn (Xiaofeng Wang),

201932128@stumail.nwu.edu.cn (Zhe Liu),

202032914@stumail.nwu.edu.cn (Xiangru Qiao),

lizhiquan@stumail.nwu.edu.cn (Zhiquan Li),

202133547@stumail.nwu.edu.cn (Sidong Wu),

202233476@stumail.nwu.edu.cn (Jiao Zhang), Liuyo@edgehill.ac.uk

(Yonghuai Liu), lizhan@nwu.edu.cn (Zhan Li), 10053934@zte.com.cn

(Hongbo Guo), Huaizhong.Zhang@edgehill.ac.uk (Huaizhong Zhang)

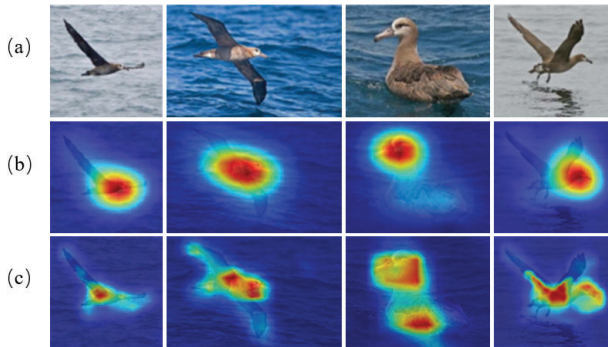


Fig. 1. Comparison of methods for randomly selected images in CUB200-2011. (a) original images; (b) original CAM; (c) our proposed method CCS-CAM.

the image depth descriptor, which was used to locate the region of the object through the activation values. However, due to the lack of category information, the deep descriptor often contains a lot of noise.

[8, 9] have demonstrated the efficiency of channel selection by reducing the unimportant or redundant channels. Adopting the similar idea, we propose a clustering-inspired method (CCS-CAM) for Weakly Supervised Object Localization (WSOL), which utilizes the channel redundancy mechanism and the features of channels to locate objects. Our method employs standard CNNs for object classification, and identifies the most discriminative parts of an object by removing the redundant channels based on the last convolutional layer of the backbone. Fig.1 shows that our CCS-CAM can capture more object regions than CAM.

To sum up, our contributions are:

- We present the WSOL network, which is an extension of image classification network. Making full use of the features learned by the classification network, WSOL can well locate the objects in images.
- We design a Clustering-inspired Channel Selection (CCS) module, which is used to filter the unimportant channels and select the important ones. The advantage is that the K-means clustering is effective and simple to find object regions and correctly localize them.
- Experimental results on the public datasets CUB200-2011 [10], ImageNet-1K [11], FGVC-Aircraft [12] and Stanford Cars [13] show that our method can deal with classification and object location well.

The remainders of this paper are organised as follows. Section 2 reviews relevant works, Section 3 details the proposed CCS-CAM, and Section 4 presents the experimental results. Finally, Section 5 concludes the paper and indicates the future work.

## 2. Related work

The most relevant works include weakly supervised object localization and channel redundancy.

### 2.1. Weakly-supervised object localization (WSOL)

WSOL is a very challenging task that requires both accurate image classification and object location. The most representative work of WSOL is CAM [1], which performs object localization by aggregating the last convolutional layer of a network trained for image classification. The shortcoming of CAM is that the areas of the location are usually too small compared to the actual ones. To address this problem, HaS and CutMix [3] randomly remove parts of the input images, which can be regarded as a data augmentation method to make the network learn other regions that are weakly correlated with the target object. [14] proposes a Gradient-based Refined Class Activation Map (GRCAM), which is a training-free weakly supervised localization algorithm and doesn't increase huge training resources. [15] combines the Generative Adversarial Network (GAN) with weak supervision object localization, using the method of Gradient-weighted CAM (Grad-CAM).

To enhance the localization region, [16] proposes the Region-based Dropout with Attention Prior (RDAP) algorithm, to improve the WSOL accuracy. Considering the power of self-supervised learning, [17] proposes an unsupervised object localization method using self-supervised learning. Following this idea, [18] introduces a Discriminative Pseudo-label samplings (DiPs) to locate weakly supervised targets, where only image-class labels are available. Given multiple attention graphs, DiPs relies on a pre-trained classifier to identify the most discriminative region in each attention graph. [19] proposes a graph-based method that uses the self-supervised transformer features to localize objects. [20] proposes a Strengthen Learning Tolerance approach (SLT-Net) to implement object localization in a localization and classification separation framework. [2] adopts a foreground consistency loss to strengthen the activation of the target area and weaken the response of the background area.

In addition to the above-mentioned CAM-based methods, some methods relies on deep descriptors of the network. The most representative works are Selective Convolutional Descriptor Aggregation (SCDA) [6] and Deep Descriptor Transforming (DDT) [7]. SCDA [6] introduces deep descriptors to guide the object localization, which accumulates all the features in the last layer, and takes the activation intensity as a localization basis. However, due to the lack of category constraint, SCDA [6] tends to be influenced by background noise in complex images. DDT [7] identifies object regions based on deep descriptors, enabling it to locate objects on a batch of images in the same category. It is worth mentioning that neither of these methods fine-tuned the pretrained classification model, so the classification performance will not be affected. However, for the generality of the network, SCDA [6] is not suitable for real scene images, while DDT [7] requires the network to generate sample image-constrained localization regions. Recently, Shieh [21] proposes a positive-weighting feature enhancement method for WSOL, which is a plug-in module and can correctly utilize the existing information of the positive weights.

### 2.2. Channel Redundancy

With the success of deep learning, many powerful models [22, 23] have been proposed. However, the number of channels

has also grown with the accuracy. Traditional convolutional networks, such as VGG [22] and ResNet [24], rely on local features of channels and map them to the fully-connected layer. These models typically require hundreds or even thousands of channels.

To reduce the number of parameters and the training time, a recent trend is to remove the redundant channels. DenseNet [23] directly connects all layers with only 32 channels to ensure information transfer between different layers. To reduce the complexity of ResNet [24], ResNeXt [25] adopts a parallel stacked block to reduce the number of network parameters. IGCNet [8] uses continuous cross-group convolution to reduce channel redundancy, ultimately increasing the performance of the network. GhostNet [9] applies fixed convolutions to reduce the number of parameters and redundant channels. SP-Conv [26] separates features into representative and uncertain two groups and processes them differently, which can reduce the number of parameters and training time.

In this paper, we propose a Clustering-inspired Channel Selection-based Class Activation Mapping (CCS-CAM) method, which generates activation maps using selected channels. It can effectively reduce redundant channels for object localization while maintaining the performance of image classification.

### 3. The proposed CCS-CAM method

We propose a CCS-CAM method for WSOL, which contains two components, Classification Network (CN) and Clustering-inspired Channel Selection module (CCS module). CCS-CAM adds a module to check the availability of channels before the fully connected layer, with the aim of clearing out channels that fail to locate the target area during training. As shown in Fig.2, CN network is a fully convolutional network for image classification followed by the Global Average Pooling (GAP) layer, Fully Connected (FC) layer, and Softmax Function (SF). The CCS module adopts a clustering-based method to locate the objects in images. The Clustering-inspired channel selection module is located in Fig.2 (b).

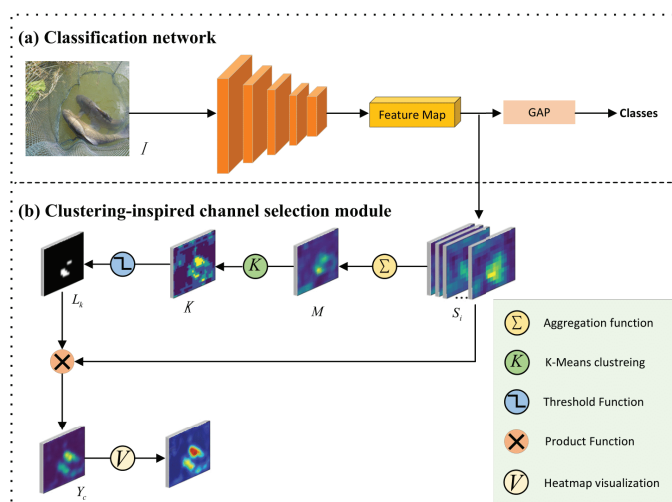


Fig. 2. The network of CCS-CAM.

We first revisit CAM method [1] as the basis of CCS-CAM, and then the channel redundancy analysis and the details of the CCS module are described.

#### 3.1. Revisit of CAM

CCS-CAM is based on well-known CAM [1], which is a classical WSOL method. CAM also contains the above mentioned Classification Network (CN) in Fig.2.

Assume input image  $I$  is an image from class  $c$ , the task of WSOL is to learn the location heat map of the image  $I$ . In Fig.2, the CN network first obtains the last convolutional feature map.

$$S \in R^{N \times H \times W} \quad (1)$$

where  $N$ ,  $H$  and  $W$  represent the number of channels, the width and height of the feature map.  $S$  is feeded into GAP to generate feature  $G$ , which is further processed by Fully Connected layers (FC) to obtain weight matrix  $WM$ . The final classification score  $M_c$  of class  $c$  is defined as

$$M_c = \sum_{n=1}^N WM_n^c \times G_c \quad (2)$$

where  $G_c$  is the output of GAP for class  $c$ ,  $WM_n^c$  is the element of the matrix  $WM$  at the  $n$ -th row and the  $c$ -th column.

The location heat map  $Y$  of  $I$  is computed by the last convolutional feature map  $S$  and the weight matrix of FC layer  $W$ , which is denoted as

$$Y_c = \sum_{n=1}^N W_n^c \times S_n \quad (3)$$

where  $S_n$  is the  $n$ -th channel in feature map  $S$ . Finally, a rectangular box for object location is produced to frame the pixels above a given threshold.

#### 3.2. Analysis of Channel Redundancy

By visualizing the last convolutional layers, it can be easily found that many channels are redundant as they correspond to the same regions of interest. Fig.3 shows the class activations for the randomly selected four channels, where the activation intensity is progressively increased from blue to red. Fig.3(b), (c), (d) and (e) show four randomly selected feature channels. The red parts are the strongest activation intensity areas of the channels. It can be seen that not all channels can respond to the object itself [22].

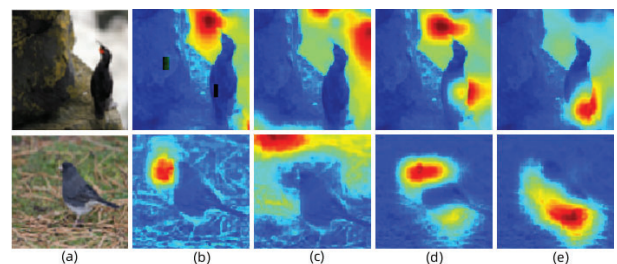


Fig. 3. Heat map of some channels in VGG over the images in (a), (b)(c)(d)(e) are picked randomly from all the channels.

Channel redundancy exists in probably all the CNNs, which is determined by the characteristics of the channels themselves. In addition, all these channels that contribute to object classification [23] may mislocate object location due to channel redundancy. As a result, we need a method to filter out the channels that are not responsible for object location and improve the performance of WSOL.

### 3.3. CCS module

From the above analysis, we can see that some channels are not available in the identification classification, that is, redundant channels. Therefore, these channels can be removed in the forward propagation.

The idea of our method is to select specific channels from the pre-trained classification model and then accumulate them to generate the final object localization result. Assuming  $S_n$  is the  $n$ -th channel in feature map  $S$ , our approach can be expressed as

$$Y = \sum_{n=1}^N L_n \times S_n \quad (4)$$

where  $L_n$  is a coefficient with a value of 0 or 1 which indicates whether the channel  $S_n$  is selected or not.  $L_n$  is calculated as shown in the CCS module in Fig.2, which is obtained by the K-means clustering.

The reason for using the k-means algorithm here is because the past work [22] uses the form of heat maps to directly present the values of regions, but the region values can be divided into several categories, and K-means is an effective unsupervised algorithm for image region division.

Our method is to select the maximum value  $\delta_n$  of all the segmentation regions in each channel and compare it with the average value  $\bar{\delta}$  of all the channels. If the maximum value  $\delta_n$  is greater than the average value  $\bar{\delta}$ , the channel is selected; otherwise, it is considered a redundant one. Specifically, we first conduct K-means clustering on the feature map generated by the sum of all the channels. Then, we calculate the ratio of the sum of feature values in each segmentation region of the  $i$ -th channel to the size of the segmentation region, and define it as the region density. We choose the maximum region density as the  $\delta_n$  of the  $i$ -th channel. And  $\bar{\delta}$  is the ratio of the sum of all the channel feature values to the size of the feature map.

Assume that  $f$  is the trained classification network of Fig.2 and  $S_n$  is the  $n$ -th channel of the last convolutional feature map. We first sum these channels and obtain

$$M = \sum_{n=1}^N S_n \quad (5)$$

Subsequently, we calculate  $L_n^c$  by using the  $M$ . The idea is to use the K-Means clustering method to divide  $M$  into  $K$  different feature regions and filter out channels. And then we compute the mean value of the  $K$  different feature regions on each channel. If its value is bigger than the feature mean of  $M$ , we assign the value of  $L_n^c$  to 1.

Assume the clustering matrix obtained by the K-Means clustering method on  $M$  be  $K$ , whose  $K$  different segmentation regions of  $M$  are  $\{R_i\}_{i=1}^K$ . We calculate the mean values of the  $K$

different segmentation regions on the channel  $n$  of  $S$  separately and select the biggest of them as the  $\delta_n$

$$\delta_n = \frac{\sum S_n^{(x,y)}}{H \times W} \quad (6)$$

where  $S_n^{(x,y)}$  represents the value of the  $n$ -th channel  $S_n$  in the  $S$ .

We define the mean of the feature map on  $M$  as

$$\bar{\delta} = \frac{\sum M^{(x,y)}}{H \times W} \quad (7)$$

where  $M^{(x,y)}$  are the feature values of the  $(x, y)$  positions in the  $M$ .

For image  $I$  whether to choose the  $n$ -th channel for the object location depends on the below ratio, which is defined as

$$L_n = \begin{cases} 1 & \text{if } \bar{\delta} < \delta_n \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

The core of CCS-CAM is to filter out these channels that are not relevant to object location to improve its performance.

### 3.4. Relationship with other WSOL methods.

In complex scenarios, such as the image-net test, because the convolutional neural network already has good classification ability, the image is downsampled into some regions with objects of interest and background regions, see Fig.8, and the noise background is cleared.

Almost all of our work is unsupervised exception the classification part. Compared to other existing WSOL methods, our network is lightweight and can easily be plugged into other classification networks.

Our method CCS-CAM can be seen as a variant of the CAM. CCS-CAM could be implemented by only replacing the object location part of the CAM. CCS-CAM actively selects channels that are helpful for object location and achieves better performance. In addition, some studies [6, 7] have shown that pre-trained models have significant potential to be exploited for object location, and our CCS-CAM takes full advantage of this. Furthermore, our method uses a very simple clustering method to select channels that are suitable for the object location task due to its effectiveness and conciseness.

## 4. Experimental results

In this section, we demonstrate the effectiveness of our CCS-CAM on various publicly available datasets.

### 4.1. Dataset

We adopt ImageNet-1k [11], ILSVRC dataset [27] as the dataset for training a classifier model, which is further used for object location. The datasets used for evaluation include CUB 200-2011 [10], FGVC-aircraft [12], Stanford cars [28], and ImageNet-1k [11], where the first three belong to fine-grained image dataset and the last to coarse-grained image dataset.

The ImageNet-1k [11] is a massive dataset containing 10 million images from 1,000 classes. We chose 1.2 million images

for training, 50,000 images for testing, and 50,000 images for evaluation. The CUB 200-2011 [10] consists of 11,788 images of 200 categories, with 5,994 images for training and 5,794 for testing. The FGVC-aircraft [12] includes 10,000 images categorized into 100 classes, divided into three equally-sized datasets for training, validation, and testing. Stanford Cars [28] contains 196 classes and 16,185 images, which is also divided equally.

#### 4.2. Experimental setting

We evaluate the proposed CCS-CAM network using VGG and ResNet as the backbone which is pre-trained on the ImageNet-1K dataset [11].

All images used in CCS-CAM network are resized to  $256 \times 256$  pixels and then randomly cropped to  $224 \times 224$  pixels. The feature maps of the final convolutional layer are aggregated to form the feature map M for the CCS module.

We apply K-means clustering method on feature map M with  $K = 3$  in 50 iterations to generate L. The reason using  $K=3$  is that the regions can be divided into three categories, the regions where objects do not exist, the regions where objects exist with low probability, and the regions where objects exist with high probability. Here we only focus on high-probability regions.

We employ transfer learning to fine-tune the pre-trained model on ImageNet-1K dataset. We train CCS-CAM network with 60 epochs, learning rate of  $5e-5$  and batch size of 128.

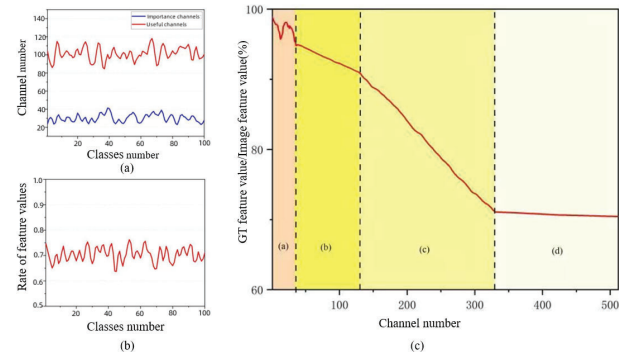
#### 4.3. Evaluation metrics

We adopt Top-1 Loc, Top-5 Loc, and GT-Know Loc (Zhou et al., 2016) as evaluation metrics. Top-1 Loc computed the ratio of images that are correctly classified and the intersection ratio between predicted and actual annotation boxes is greater than 0.5. Top-5 Loc selects the top 5 ratio similar to Top-1 Loc. GT-Know Loc considers the intersection ratio of the real annotation box and the predicted one to be larger than 0.5, regardless of whether the predicted classification is correct or not.

#### 4.4. Channel analysis of CCS-CAM

To analyze the channel selection performance of our CCS-CAM, we carry out three experiments using the VGG16 network, whose results are shown in Fig.4. Fig.4(a) is the comparison of the number of important channels and the number of useful channels. The important channels are the ones selected by our CCS-CAM, while the useful channels are those channels whose object regions overlap with the Ground Truth (GT) box. The GT box refers to the real annotation box of the target object in an image. Fig.4(b) is the ratio of the sum of feature map values in selected important channels to the sum of feature map values in the aggregation map M. Fig.4(c) shows the ratio of the feature information of the GT box to the aggregation M as the number of channels increases.

As Fig.4 (a) shows, the number of useful channels obtained by VGG16 is far greater than the number of important channels selected by our CCS-CAM. Fig.4 (b) shows that the amount of information contained in important channels is about 70% of the aggregation map M, indicating that most of the necessary information for classification in channels are selected by our



**Fig. 4.** The analysis of channel redundancy characteristics. (a) shows the number of important and useful channels, (b) shows the proportion of information contained in important channels in aggregation map  $M$ , (c) shows the results of random 100 class aggregation channels in ImageNet-1k.

CCS-CAM. From Fig.4 (a) and Fig.4 (b), it can be seen that our network CCS-CAM can significantly reduce the model parameters in object localization, while still retaining the classification performance of the network.

To observe that the channels we select contain most of the information in the GT box, we calculate the ratio of the information in the GT box to the M as the ordinate in Fig.4(c). Fig.4(c) shows the results of 100 random class aggregation channels in ImageNet-1k with VGG as the backbone. Starting from 0, we continuously add new channels to the channels that our algorithm has already selected. From Fig.4(c), it can be seen that as the number of channels increases, the ratio of the information in the GT box to the aggregation M is initially fluctuated slightly, and then gradually decreases. We use different colors to indicate that when the number of channels reaches a certain number, the positioning area will be rapidly reduced, there are about 100 or so channels that play a great role in positioning, and when the channels reach more than 300, the channels basically do not have an effect on positioning. These redundant channels have a relatively small impact on algorithm performance, thus eliminating the redundant channels cannot affect the performance of our network.

**Table 1.** The performance of CAM and CCS-CAM for object location on CUB-200-2011.

| Dataset      | Method         | GT-Know Loc  |
|--------------|----------------|--------------|
| CUB-200-2011 | CAM            | 51.09        |
|              | <b>CCS-CAM</b> | <b>87.90</b> |

#### 4.5. Ablation study

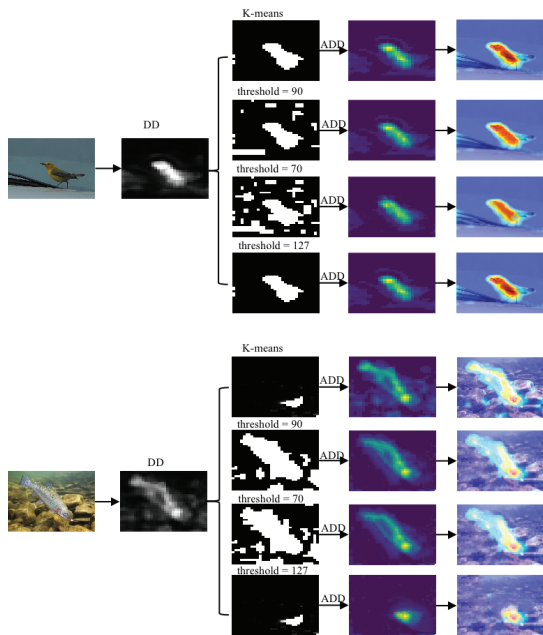
The ablation study is conducted using the ImageNet-1k. Fig.5 illustrates the high response regions in different images using various thresholds on the aggregation M. When analyzing the fish image from the ImageNet-1k or the bird image from the CUB200-2011, it is obvious that threshold value 90 is more appropriate, but the K-means is the best. Therefore, dynamic segmentation using K-means is more suitable than using a fixed or average threshold alone.

We carry on an experiment on CUB200-2011. As shown in Table 1, CCS-CAM outperforms CAM [1] on CUB200-2011, with more than 36% increase in GT-Know Loc respectively.

**Table 2. The performance of the proposed and state-of-the-art methods over different datasets using different backbone networks.**

| Model                | Backbone | CUB 200-2011  |               |                  | ImageNet-1k   |               |                  |
|----------------------|----------|---------------|---------------|------------------|---------------|---------------|------------------|
|                      |          | Top-1 Loc (%) | Top-5 Loc (%) | GT-Known Loc (%) | Top-1 Loc (%) | Top-5 Loc (%) | GT-Known Loc (%) |
| CAM [1]              | VGG      | 44.15         | -             | 57.96            | 42.80         | 54.86         | 59.00            |
| CutMix [3]           | VGG      | 52.53         | -             | -                | 43.45         | -             | -                |
| I <sup>2</sup> C [4] | VGG      | 55.99         | 68.40         | -                | 47.41         | 58.51         | <b>63.90</b>     |
| Ci-CAM [5]           | VGG      | 58.39         | 70.54         | 75.68            | 48.71         | 58.76         | <b>62.36</b>     |
| SLT-Net [20]         | VGG      | 67.8          | -             | -                | -             | -             | -                |
| PDAP [16]            | VGG      | 59.35         | -             | -                | 46.52         | -             | -                |
| GRCAM [14]           | VGG      | 56.01         | 66.44         | -                | -             | -             | -                |
| <b>CCS-CAM</b>       | VGG      | <b>69.21</b>  | <b>80.16</b>  | <b>87.9</b>      | <b>49.26</b>  | <b>59.13</b>  | 62.15            |
| CAM [1]              | ResNet   | 42.72         | -             | -                | 46.19         | -             | -                |
| CutMix [3]           | ResNet   | -             | -             | -                | 47.25         | -             | -                |
| GRCAM [14]           | ResNet   | 60.53         | 72.08         | -                | -             | -             | -                |
| TokenCut [19]        | ViT-S/16 | -             | -             | -                | 52.3          | -             | -                |
| Dips [18]            | DeiT-s   | 79.2          | 92.2          | -                | -             | -             | -                |
| PDAP [16]            | ResNet   | 63.65         | -             | -                | 50.33         | -             | -                |
| RCAM+PFE [21]        | ResNet   | 63.93         | -             | 79.77            | -             | -             | -                |
| <b>CCS-CAM</b>       | ResNet   | <b>71.45</b>  | <b>83.69</b>  | <b>86.37</b>     | <b>52.32</b>  | <b>63.68</b>  | <b>63.71</b>     |

This reason is that the clustering in CCS-CAM is more efficient and the channel selection method of it can improve the performance.



**Fig. 5. Heatmap region extraction using different methods: K-means versus fixed threshold. DD is the features of aggregation  $M$ , and ADD is the features of  $X_c$ .**

#### 4.6. Comparative study

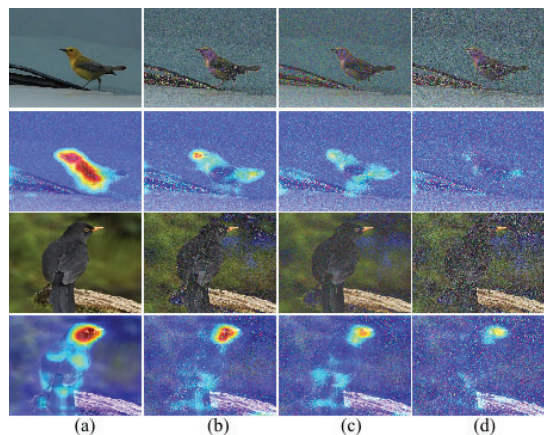
We compare CCS-CAM with other WSOL methods quantitatively and qualitatively. Because CCS-CAM simply throws away all the channels that aren't able to help object localization, and other models just try to retrain those models, we get better results.

##### Quantitative Evaluation

The quantitative experiment is conducted on CUB 200-2011 and ImageNet-1k by using VGG and ResNet as the backbone.

The compared WSOL methods include state-of-the-art unsupervised, co-supervised, and weakly supervised methods. As shown in Table 2, CCS-CAM achieves almost optimal performance on three evaluation indicators. CCS-CAM outperforms the other WSOL methods, achieving 87.90%, 63.71% accuracy on CUB 200-2011, and ImageNet-1k in GT-Known Loc respectively. To verify the performance of our algorithm on fine-grained data sets, we have also conducted experiments specifically on this. As shown in Table 3, our method used ResNet still outperforms other WSOL methods.

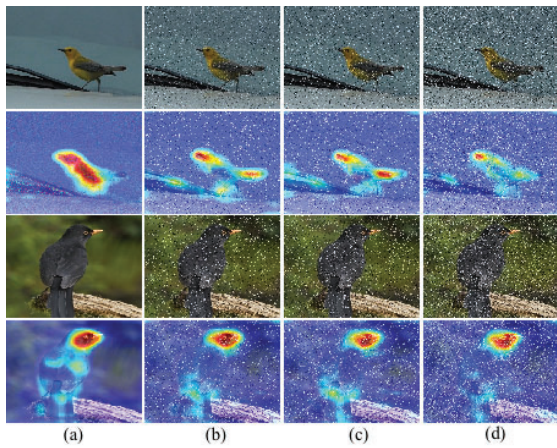
To verify the effect of noise on the CCS-CAM method, we add noise to the images and see whether the objects can still be localized (Fig.6 and Fig.7) and get the result that Gaussian noise and Salt-pepper noise added to the images don't affect the localization effect of our CCS-CAM method.



**Fig. 6. Image (first and third line) and Heatmap (second and fourth line) of Bird with Gauss noise. (a) Gauss\_sigma=0 (b) Gauss\_sigma=0.1 (c) Gauss\_sigma=0.12 (d) Gauss\_sigma=0.15.**

**Table 3. The performance of state-of-the-art methods for object localization over three fine-grained datasets**

| Dataset       | Method         | GT-Know Loc  |
|---------------|----------------|--------------|
| FGVC-Aircraft | SCDA [11]      | 94.91        |
|               | DDT [7]        | 92.53        |
|               | MO [29]        | 94.94        |
|               | PsyNet [30]    | 95.59        |
|               | Self [17]      | 96.72        |
|               | <b>CCS-CAM</b> | <b>98.43</b> |
| Stanford Cars | SCDA [6]       | 90.96        |
|               | DDT [7]        | 71.33        |
|               | MO [29]        | 92.51        |
|               | PsyNet [30]    | 96.61        |
|               | Self [17]      | 97.73        |
|               | <b>CCS-CAM</b> | <b>98.12</b> |



**Fig. 7. Image (first and third line) and Heatmap (second and fourth line) of Bird with Salt\_pepper noise. (a) Salt\_pepper=0 (b) Salt\_pepper=0.05 (c) Salt\_pepper=0.07 (d) Salt\_pepper=0.1.**

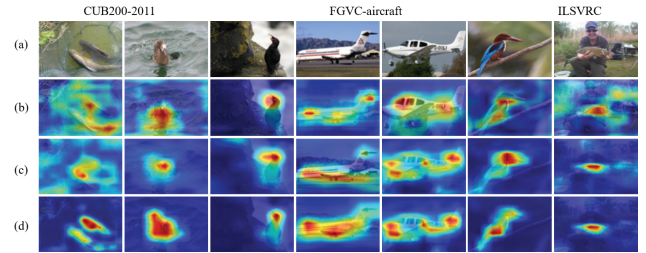
### Qualitative Evaluation

To visually and qualitatively evaluate the performance of different methods, heatmaps generated by various methods over randomly selected images from the CUB 200-2011, FGVC-aircraft, and ILSVRC2016 are presented in Fig.8. The original images are shown in Fig.8(a), while Fig.8(b), (c) and (d) show the heatmaps obtained using CAM, the aggregation map  $M$ , and the proposed method CCS-CAM, respectively. The results demonstrate that the proposed method captures less discriminative parts than CAM, and eliminates noisy regions while highlighting integral objects, such as fish, birds, and airplanes.

Compared with the latest methods such as Dips, CCS-CAM is an unsupervised attention module, which is slightly lacking in accuracy, but has no training cost and training time.

### 5. Conclusion and Future Work

In this paper, we propose a novel method CCS-CAM to tackle the challenging WSOL task. The advantage of our method is the effectiveness of the K-means clustering algorithm and channel redundancy determination mechanism. Our method clusters the feature map generated by the sum of all the channels using the K-means algorithm, and further compares the maximum region density of each channel with the



**Fig. 8. The heatmaps generated using different methods over randomly selected images in the CUB200-2011, FGVC-aircraft and ILSVRC2016, respectively. (a) original images; (b) renderings generated using CAM [1]; (c) unsupervised heatmap generated using the aggregation map  $M$ ; (d) The proposed CCS-CAM method.**

average of all the channels for useful channel selection. By selecting the useful channels in CNN network, CCS-CAM can effectively locate objects in images and classify them into different categories. The experiments on datasets CUB 200-2011, FGVC-aircraft, Stanford cars, and ImageNet-1k show the efficacy of CCS-CAM. The limitation of our algorithm is that  $K$  in K-means clustering needs to be manually selected, and the selected useful channels need to be further integrated for object localization. In the future work, we will further investigate how to choose more suitable automatic clustering algorithm and to construct more expressive features on these selected useful channels for object classification and localization.

### 6. Acknowledgement

This work is supported in part by the General Projects of Shaanxi Provincial Key R&D Plan - Industrial Field ( Number: 2024GX-YBXM-555, 2021GY-171) and the National Key R&D Plan Project (Number: 2020YFC1523300) and The National Natural Science Fund (Number: 61602380).

### References

- [1] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.
- [2] Minsong Ki, Youngjung Uh, Wonyoung Lee, and Hyeran Byun. In-sample contrastive learning and consistent attention for weakly supervised object localization. In *Proceedings of the Asian Conference on Computer Vision*, 2020.
- [3] Sangdoon Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Jun-suk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.
- [4] Xiaolin Zhang, Yunchao Wei, and Yi Yang. Inter-image communication for weakly supervised localization. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pages 271–287. Springer, 2020.
- [5] Feifei Shao, Yawei Luo, Li Zhang, Lu Ye, Siliang Tang, Yi Yang, and Jun Xiao. Improving weakly supervised object localization via causal intervention. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3321–3329, 2021.
- [6] Xiu-Shen Wei, Jian-Hao Luo, Jianxin Wu, and Zhi-Hua Zhou. Selective convolutional descriptor aggregation for fine-grained image retrieval. *IEEE transactions on image processing*, 26(6):2868–2881, 2017.
- [7] Xiu-Shen Wei, Chen-Lin Zhang, Yao Li, Chen-Wei Xie, Jianxin Wu, Chunhua Shen, and Zhi-Hua Zhou. Deep descriptor transforming for image co-localization. 2017b.

- [8] Ting Zhang, Guo-Jun Qi, Bin Xiao, and Jingdong Wang. Interleaved group convolutions. In *Proceedings of the IEEE international conference on computer vision*, pages 4373–4382, 2017.
- [9] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1580–1589, 2020.
- [10] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.
- [11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [12] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.
- [13] Bo Zhao, Xiao Wu, Jiashi Feng, Qiang Peng, and Shuicheng Yan. Diversified visual attention networks for fine-grained object classification. *IEEE Transactions on Multimedia*, 19(6):1245–1256, 2017.
- [14] Wenjun Hui, Chuangchuang Tan, Guanghua Gu, and Yao Zhao. Gradient-based refined class activation map for weakly supervised object localization. *Pattern Recognition*, 128:108664, 2022.
- [15] Minsoo Park, Jinyeong Bak, Seunghee Park, et al. Advanced wildfire detection using generative adversarial network-based augmented datasets and weakly supervised object localization. *International Journal of Applied Earth Observation and Geoinformation*, 114:103052, 2022.
- [16] Junsuk Choe, Dongyoon Han, Sangdoon Yun, Jung-Woo Ha, Seong Joon Oh, and Hyunjung Shim. Region-based dropout with attention prior for weakly supervised object localization. *Pattern Recognition*, 116:107949, 2021.
- [17] Yukun Su, Guosheng Lin, Yun Hao, Yiwen Cao, Wenjun Wang, and Qingyao Wu. Self-supervised object localization with joint graph partition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2289–2297, 2022.
- [18] Shakeeb Murtaza, Soufiane Belharbi, Marco Pedersoli, Aydin Sarraf, and Eric Granger. Dips: Discriminative pseudo-label sampling with self-supervised transformers for weakly supervised object localization. *Image and Vision Computing*, 140:104838, 2023.
- [19] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14543–14553, 2022.
- [20] Guangyu Guo, Junwei Han, Fang Wan, and Dingwen Zhang. Strengthen learning tolerance for weakly supervised object localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7403–7412, 2021.
- [21] Jeng-Lun Shieh, Sheng-Feng Yu, and Shanq-Jang Ruan. Positive-weighting feature enhancement for weakly supervised object localization. *Pattern Recognition Letters*, 170:56–63, 2023.
- [22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Computer Science*, 2014.
- [23] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [25] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.
- [26] Qiulin Zhang, Zhuqing Jiang, Qishuo Lu, Jia'nan Han, Zhengxin Zeng, Shang-Hua Gao, and Aidong Men. Split to be slim: An overlooked redundancy in vanilla convolution. 2020.
- [27] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.
- [28] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.
- [29] Runsheng Zhang, Yaping Huang, Mengyang Pu, Jian Zhang, Qingji Guan, Qi Zou, and Haibin Ling. Mining objects: Fully unsupervised object discovery and localization from a single image. *IEEE Trans. Image Processing*, 2020c.
- [30] Kyungjune Baek, Minhyun Lee, and Hyunjung Shim. Psynet: Self-supervised approach to object localization using point symmetric transformation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10451–10459, 2020.