Hide-CAM: Finding Multiple Discriminative Regions in Weakly Supervised Location

JIE XU1, (Member, IEEE), SHUWEI SHENG1, HAOLIANG WEI1, JINHONG GUO1, (Member, IEEE)

1 School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China

Corresponding author: JINHONG GUO (e-mail: guojinhong@uestc.edu.cn).

This work was supported by National Key Research and Development Program (Grant No. 2016YFB0800105), Sichuan Province Scientific and Technological Support Project (Grant Nos. 2018GZ0255, 2019YFG0191).

ABSTRACT Weakly supervised localization is a more challenging task due to the absence of an object’s annotation. Because the depth convolution feature can well represent the spatial information of the object, the position of the object can be located by the saliency study of the image. However, the most discriminative area tends to focus too much on the details of the object and lacks the perception of the object’s overall structure, whereas the information of the complementary object regions is the complement of the most discriminative area. The combination of these information types can fully express the global information of the object. Therefore, our method takes into account the entire area of the object rather than the most discriminative area. In this paper, the hide strategy is used to locate the most discriminative and the complementary object regions of the object. First, we use CAM to extract the most discriminative area. Next, we mask the most discriminative area and use CAM to extract complementary object regions in the masked image. Finally, the two areas are integrated to complete the task of location. Our method only needs the classification label of the image instead of a detailed object annotation. The operation is simple and convenient, and does not require training a complex model or additional annotation. Experiments show our method achieves good results in ILSVRC 2012 validation.

INDEX TERMS Weakly Supervised Location, Significant Region, Biologically Inspired Image, Global Information

I. INTRODUCTION

Weakly supervised localization (WSL) is to learn the position of an object based only on the label of the image. In recent years, deep learning has achieved significant improvements in various areas of computer vision. The most basic element of deep learning is still data. The larger the amount of data, the richer the information and the better the corresponding model. However, the acquisition of massively annotated data is a very time-consuming and laborious task, and it is often difficult to achieve in practical applications. Therefore, weakly supervised learning requires only the category label of the image and becomes an important method for solving data problems. Weakly supervised learning has been proved to have good application in various fields, such as object detection, semantic segmentation, and visual attribute positioning. Therefore, WSL is a challenging and practical task. Compared with supervised learning, it requires less detailed annotation; thus, it is possible to accomplish multiple large-scale image tasks with the help of a large number of weakly labelled visual data on the Web.

The current WSL method is mainly generated by mining discriminative image regions or features by analyzing the higher-level features generated by the deep network trained for image classification. Zhou et al. proposed a simple and efficient feature-based CAM method. Instead of the neural network max pooling layer, he used global average pooling (GAP) after the last layer of convolution to obtain larger size feature maps. To ensure more detailed features, he used the weights of the neurons in the fully connected layer corresponding to the classification results to optimize the feature map of the output of the last layer of the convolutional layer. Therefore, the positioning effect of
CAM is closely related to the accuracy of the classification. However, CAM only focuses on the most prominent areas of the object while ignoring the less important ones. The information in the most prominent areas is often not perfect and this less important information is a good complement and perfect for the most prominent areas. However, due to variations within a category or relying solely on a classification goal, these methods generally do not recognize the entire range of the object, but only locate the most discriminative region in the image.

To tackle such issues, Dahun Kim proposed a two-phase learning method to discover the entire portion of the target object by combining the activations of both phases. However, one main disadvantage of two-phase learning is that it cannot capture import information from two activations. Recently, Singh et al. proposed “Hide-and-Seek” by randomly hiding the patches of input images to force the network to look for other discriminative parts. However, randomly hiding patches without any high-level guidance is inefficient and cannot guarantee that networks always discover new object regions.

In this paper, we propose Hide-CAM for weakly supervised location, which can combine the most salient and complementary object regions of the image to locate the entire range of the object. In the first step, we use the CAM method to extract the most salient areas. In the second step, we cover up the most discriminative area and use CAM to extract complementary object regions in the remaining area. Finally, we integrate the two areas to get the entire range of objects. The algorithm is very convenient: it does not need to train the location network and only needs the corresponding classification network of the dataset, and it can be applied to any network architecture with a little modification.

II. RELATED WORK

The supervised CNN network has achieved good results in object detection, segmentation, and attribute positioning; however, it requires significant labor to label the dataset.

There has been a lot of work recently focused on weakly supervised localization algorithms based on neural networks. The previous WSL methods found high-confidence regions from abundant positive samples of images with positive annotations by applying multiple instance learning or other algorithms. That is, the object proposal method is first applied to decompose the image into object proposals and then it is used for iterative execution of proposal selections and classifiers (soft proposals) through latent variable learning methods (e.g., multi-instance learning, MIL). The premise of this method is that the candidate area containing the target can be obtained through object proposal methods, e.g., Selective Search (SS) and Edge Boxes (EB). The more candidate areas, the better the location effect. This is a way to pick out the areas obtained by the exhaustive method. The effect of this method is better, but the efficiency is inadequate. Song et al. proposed a graph-based method to initialize the object locations by solving a submodular cover problem. WSDDN, which significantly improves object detection by implementing proposal selection to improve learning and classifier learning. ProNet uses a cascade of two networks: the first generates bounding boxes and the second classifies them.

The requirement for the CAM to modify the structure of the original model led to the need to retrain the model, which greatly limited its use scenario. The general idea of Grad-CAM is the same as that of CAM: they all assign weights to feature maps. The difference is that CAM adopts the weight of the fully connected layer, whereas Grad-CAM uses the gradient global average to calculate the weight. At the same time, the ReLU layer was added and only those pixels that have a positive influence on the category are concerned.

Recent work has modified the CNN architecture designed for image classification so that convolutional layer learning learns to locate objects while performing image classification. Other network architectures are designed for weakly supervised object detection. Dilated residual networks obtain larger features by expanding the receptive fields of the neural network for object localization.

There have also been many applications recently based on masking image patches. For object location, trained a classification network and then located the regions. Because these approaches mask the image patches only during testing and not during training, they pay too much attention to highly discriminative object parts. Two-phase learning can discover more discriminative features so as to obtain the entire parts of the object. However, in the first phase, it applies a simple threshold to activation that cannot capture the most discriminative region in the feature map. Hide-and-Seek randomly hides the patches of input images to force the network to look for other discriminative parts. Because random hiding is an inefficient method, the network cannot always discover new discriminative regions.

Our proposed method is fundamentally different from the previous approaches. We do not focus on picking the most discriminative regions, but rather on finding more comprehensive features of objects. We enhance CAM by masking the most discriminative region. In addition, our approach relies on no part annotations.

III. METHOD

In this section, we describe the Hide-CAM algorithm for object localization in an image. Our algorithm includes using CAM to extract the most discriminative area, masking the most discriminative area, using CAM to
extract the complementary object regions belonging to the same objects or categories, and finally combining the two parts. Our algorithm is shown in Figure 1.

A. First-step CAM

In this subsection, our goal is to obtain the most discriminating area in the image. Therefore, CNNs were trained for classification tasks. CAM is used to obtain the first significant region and heatmap of the image. However, there are still several small noise parts that are activated in a complex context. Therefore, we collect the maximum connected component of the heatmap to eliminate interference caused by the noise component. Then, we convert the heatmap to a binary image and find the largest connected area.

For a given image \( I \) of size \( W \times H \times 3 \), we extract the features \( f : W' \times H' \times C \) of the last convolutional layer, the weights of the fully connected layer \( N \times C \times W' \times H' \), and the output \( N \), where \( N \) is the number of dataset categories. We then select the weight corresponding to the classification result and calculate the CAM characteristic map:

\[
M_1 = \sum_{i} w^i f_i . \tag{1}
\]

From Eq. (1), we obtain the most discriminative heatmap of size \( W' \times H' \). For \( M_1 \), we obtain the most discriminative region by selecting its maximum connected region, scale it to the original image size by interpolation, and then obtain the mask map \( M_R (M_R \in \mathbb{R}^{W\times H}) \) of the most discriminative region by a standardization operation. The size of each point in the mask map is in the range 0–1, which represents the relevance of each spatial location point to the category.

B. Hide salient region

As can be seen from Figure 2, some significant areas are the most important parts of the image, such as the head of a bird. These are ideal situations. The most striking thing behind these areas is the bird’s claws, abdomen, etc. These areas also play an important role in the recognition of birds. After hiding these images, these parts will be missing from the image. From the image, the salient regions in the new masked image will fall in the less significant regions, which are an important complement to the most discriminative regions. Because many of the most discriminative areas are too focused on the most prominent parts, the selected area is too small. Therefore, extracting less significant regions can make up for this deficiency.

We obtain a new mask map \( M_s \) by obscuring the mask map to cover the most discriminative area. \( M_s \) ignores the most discriminative area. We multiply \( M_s \) by the original image to mask the most discriminative area in the original image and obtain a new input image \( M_R^i \):

\[
M_R^i = l_i \times M_s, i \in \{0,1,2\} \tag{2}
\]

where \( M_R^i \) represents the \( i \)-th channel of the image.

C. Second-step CAM

In this subsection, we provide the CNNs with the most prominent area of coverage to obtain a second CAM heatmap, which indicates the next significant area in the image.

For a hidden image, we calculate the CAM heatmap of the image:

\[
M_2 = \sum_{i} w^i f_i . \tag{3}
\]

We fixed the weight of the FC layer to eliminate the impact of false prediction. The classification result of images that lack the most discriminative region probably lead to false output of the FC layer, which will cause the weight of the corresponding FC layer to be incorrect. Thus, we adopt the weights of the FC layer in first-step CAM to ensure the heatmap concentrates on the next salient region.

D. Merge two heatmaps

Through the above steps, we have obtained \( M_1 \) that reflects the most discriminative area and \( M_2 \) that reflects the complementary object regions. We combine these two heatmaps and normalize them to obtain the heatmap that finally reflects the overall significant area:

\[
M = \text{Normalize}(\beta M_1 + \gamma M_2) \tag{4}
\]

where \( \beta \) and \( \gamma \) are factors that trade off the relative importance between the two heatmaps. By selecting the circumscribed rectangle of the \( M \) largest connected area,
the final predicted bounding box of the object can be obtained.

IV. EXPERIMENT

In this section, we evaluate the Hide-CAM positioning capabilities. We first introduce the experimental datasets and evaluation criteria, and then perform quantitative and qualitative evaluations of Hide-CAM for object localization in images localization. We also perform ablative studies to compare the different design choices of our algorithm.

A. Setup

Dataset and evaluation indicators: Because the data for the classification and localization tasks in ILSVRC 2017 remain unchanged from ILSVRC 2012, we use ILSVRC 2012[30] to evaluate object localization accuracy. We use the positioning of the bounding box to measure the performance of the algorithm. The metric calculates the percentage of the images whose bounding boxes have over 50% IoU with the ground-truth.

Implementation details: to compare with CAM, we use the modified VGGnet-GAP, AlexNet-GAP, and GoogLeNet-GAP. For AlexNet, we removed the layer after conv5 (i.e., pool5 to prob), resulting in a mapping resolution of 13x13. For VGG, we removed the layer after conv5-3 (i.e., pool5 to prob), resulting in a mapping resolution of 14x14. For GoogLeNet[31], we removed the layers after inception4e (i.e., pool4 to prob), resulting in a mapping resolution of 14x14. For each of these networks, we have added a convolutional layer of 3x3, a stride of 1, pad 1 with 1,024 units, followed by a GAP layer[32], and a softmax layer.

During the test, we performed 10-oversampling cuts for each image, taking the map (top left, bottom left, top right, bottom right, and center) and their horizontal flips. The CNNs feature maps and the predicted classification results are extracted separately for ten pictures. Finally, the ten feature maps are merged to obtain the final heatmap.

B. Object localization quantitative results

<table>
<thead>
<tr>
<th>Method</th>
<th>Top-1 Location Error</th>
<th>Top-1Cls error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hide-CAM in AlexNet-GAP</td>
<td>61.64%</td>
<td>43.38%</td>
</tr>
<tr>
<td>AlexNet-GAP(CAM)[31]</td>
<td>63.75%</td>
<td>44.9%</td>
</tr>
<tr>
<td>AlexNet-HaS-Mixed[32]</td>
<td>62.35%</td>
<td>41.32%</td>
</tr>
<tr>
<td>Hide-CAM in GoogLeNet-GAP</td>
<td>54.61%</td>
<td>34.5%</td>
</tr>
<tr>
<td>GoogLeNet-GAP(CAM)[31]</td>
<td>56.40%</td>
<td>35.0%</td>
</tr>
<tr>
<td>GoogLeNet-HaS-32[32]</td>
<td>54.79%</td>
<td>29.30%</td>
</tr>
<tr>
<td>Hide-CAM in VGGnet-GAP</td>
<td>52.67%</td>
<td>32.40%</td>
</tr>
<tr>
<td>VGGnet-GAP(CAM)[31]</td>
<td>57.20%</td>
<td>33.4%</td>
</tr>
<tr>
<td>Grad-CAM on VGG16[31]</td>
<td>56.51%</td>
<td>30.8%</td>
</tr>
</tbody>
</table>

We first make a quantitative analysis of the results of our algorithm on ILSVRC 2012 validation. Table 1 shows our location results.

In this experiment, we use AlexNet-GAP, VGGnet-GAP, and GoogLeNet-GAP to extract convolutional features and evaluate the classification and localization error on the ILSVRC 2012 validation. The experimental results are shown in Table 1. Our method is only for the reduction of the location error, thus, there is no significant improvement in the recognition accuracy. As shown in Table 1, compared with CAM, the error of our method has been reduced to different degrees in different network structures. The location error rate on AlexNet-GAP was reduced by 2.11%, by 4.6% on VGG-GAP, and by 0.9% on GoogLeNet-GAP. There are big differences in the improvement of different basic network architectures in the table. The improvement of our methods on AlexNet and VGG is larger than that of GoogLeNet. This result proves that our method pays more attention to the whole part of the object than CAM, thus improving the accuracy of location.

![Fig. 2](image_url) Results of Hide-CAM. (a) and (b) are examples of localization from GoogleNet-GAP. The ground-truth boxes are in blue, bounding boxes of CAM are in green, and the predicted bounding boxes of Hide-CAM are in red.
Fig. 3 Heatmaps of Hide-CAM. The first line shows the most discriminative area of the object. The second line shows the complementary object regions. The third line shows the integral feature of the object.

Compared with the current excellent location method, we have increased by 3.21% on VGG-GAP compared with Grad-CAM. Hide-and-Seek randomly masks some patches of the picture during training. This is a method similar to data augmentation. Therefore, the recognition accuracy of AlexNet and GoogLeNet is significantly improved and the accuracy of positioning is correspondingly improved. Therefore, Hide-and-Seek has high location accuracy. GoogLeNet-HaS-32 represents the infrastructure GoogLeNet, which divides the picture into 32×32 small meshes during training and randomly masks. AlexNet-HaS-Mixed represents the infrastructure AlexNet. During the training process, the pictures are randomly divided into $N \times N, N = 1,16,32,44,56$, where $N = 1$ means normal training of the neural network not randomly masked. Because we are purposefully covering up the areas in the picture, eliminating the training process, only offline feature extraction and integration. At the same time, our approach has certain advantages and the error of location on AlexNet is reduced by 0.71%.

C. Object localization qualitative results

We visualize the results and heatmap. In Figure 2, we present a schematic diagram of the results of Hide-CAM. In the figure, the blue border represents ground-truth, the green border is derived from CAM, and our method is represented by a red border.

In Figure 2(a), CAM focuses on the saliency area of the image, such as the husky’s or cat’s head. However, because it pays too much attention to the most discriminative area, it lacks some less significant areas of the object and cannot capture the whole object. As shown in Figure 2(a), CAM lacks attention to the body parts of dogs or cats, resulting in a certain gap between the location results of CAM and ground-truth. Our method masks the most prominent area to obtain complementary object regions, such as the husky’s and cat’s legs. The combination of the most prominent area and the less significant area makes our method closer to ground-truth.

From Figure 2(b), we show another advantage of saliency-based localization based on CAM and masking algorithms: exclude interference from complex backgrounds. When there are multiple identical objects in the image, the CAM is easily confused, resulting in the location area containing too much complex content to highlight the salient region of the image. For example, in Figure 2(b), when the target is a degree cap, CAM will count the arm with the degree cap (which is obviously not what we want). When the target is the cello, CAM will include the person. Although these parts are more prominent in the image, there is not much connection with the image category and our method can avoid the interference of complex background and focus on the significant area related to the image category.

Fig. 4 Some failure examples: ground-truth boxes are in red, CAM boxes are in blue, and the predicted bounding boxes from the class activation map are in green.
In Figure 3, we visualize some heatmaps in our Hide-CAM. In this figure, the region with warmer color is more important than that with cool color. We can see that the CAM heatmaps only focus on the most prominent areas of the image, such as animals’ heads as shown in the first line. After hiding these regions, the network can explore complementary object regions such as an animal’s leg or abdomen as shown in the second line. We then integrate the prominent area and complementary object regions to obtain the final overall area of the object. Our algorithm is more comprehensive and can obtain more information about the target objects, which provides more possibilities for our future operations.

![Heatmap](image)

**Fig. 5** Effect of super-parameter on experimental results. (a) Relation between θ and location error. (b) Relation between σ and location error.

At the same time, through the analysis of the WSL results, we find that because we emphasize the overall characteristics of the object, we have better results in locating some pictures that need to integrate features of different regions, such as animals. However, there are still some images that need to be localized with more detailed features, as shown in Figure 4. In these images, the effect of our method is not ideal. Because our method is based on feature maps, the size of the feature map is small, making it difficult to be sensitive to small objects. In addition, high-level feature maps highlight areas with high-level semantics and lack many details.

### D. Further Analysis

**Effect of super-parameter on experimental results:** In this experiment, two super-parameters are important. First, in order to generate the most discriminative area in first-step CAM, we set an appropriate threshold θ to segment the heatmap and then take the bounding box that covers the largest connected component in the segmentation map. Second, in order to obtain the final bounding box from the heatmap, which includes the most discriminative and complementary features, we set a threshold σ to divide the heatmap. These two super-parameters are important for the final location accuracy.

Therefore, we observe the impact on the results by setting different thresholds. For the value of θ, according to the percentage of the maximum value of the heatmap, we vary θ to be 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, or 0.9. In this experiment, VGG is used as the basis network. The result is shown in Figure 5(a) where we set θ = 0.8. It can be seen from the graph that different values of θ have a key role in the error rate of location. The larger the value of θ, the smaller the most discriminative area selected. The selection of the most discriminative areas affects the location of the complementary object regions. If the selection of the most discriminative areas is too large, then it will easily cover the less significant areas, resulting in the wrong background as a less significant area and producing a negative impact on the final location. If the selection of the most discriminative area is too small, then the less significant area is likely to coincide with the most discriminative area, making it difficult to locate the entire area of the object.

For the value of σ, according to the range of RGB (0–255), σ ∈ (10–70) is taken, respectively. This experiment uses VGG as the foundation network. The result is shown in Figure 5(b) where we set σ = 30. It can be seen from the graph that σ of different values plays a key role in the error rate of location. A point on the feature map whose pixel value is less than σ will be considered as the background. A pixel value greater than σ is considered to be a valuable area. The larger the value of σ, the smaller the connected region that is formed and the more the positioned region tends to be a more detailed portion than the entire portion of the object. The smaller the value of σ, the more background and the pixel points that are not too much associated with the image semantics are also included, which also has a certain negative impact on the location of the object. Therefore, a suitable value of σ is the key to object localization. It can be seen from the graph that when σ is in the appropriate interval (30–50), the error rate of location is low.

**Relationship between identification and location of the image:** Because the correct rate of image recognition is closely related to the accuracy of location, if the classification of the prediction is wrong, then the location will be considered wrong. To describe the relationship between the two in detail, in this experiment, we assume that the classification of all images is correct and only
focus on whether the selection of the saliency region is correct. It can be seen from Table 2 that when the neural network knows the correct category of the picture, the error rate of the location is greatly reduced: on GoogLeNet-GAP, the location error rate is reduced by 19%; on VGGnet-GAP, the location error rate is reduced by 17%; and on AlexNet-GAP, the location error rate is reduced by 23%. In Hide-and-Seek, ground-truth was also used as the predictive label. In contrast, our results are 2% to 3% higher in GoogLeNet and AlexNet. It shows that the effect of positioning has been significantly improved after eliminating the influence of recognition accuracy. From the perspective of the selection of significant regions, our method performed better on the ILSVRC2012 dataset. Moreover, our method can be applied to a variety of network structures, is easily combined with other network structures, and has great flexibility.

<table>
<thead>
<tr>
<th>Method</th>
<th>Top-1 Location Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hide-CAM in AlexNet-GAP</td>
<td>39.43%</td>
</tr>
<tr>
<td>Hide-CAM in VGGnet-GAP</td>
<td>36.15%</td>
</tr>
<tr>
<td>Hide-CAM in GoogLeNet-GAP</td>
<td>36.28%</td>
</tr>
<tr>
<td>AlexNet-HaS-32</td>
<td>41.25%</td>
</tr>
<tr>
<td>GoogLeNet-HaS-32</td>
<td>39.71%</td>
</tr>
</tbody>
</table>

From the above experiments, the accuracy of image classification largely affects the accuracy of object location. First, from the evaluation criteria, the misclassified picture will not be calculated by IoU, which is directly considered as a location error. This directly affects the accuracy of object location. Second, from the perspective of neural network interpretability, our method needs to extract the characteristics of the convolution layer of the last layer and the weight of the fully connected layer of the prediction category. If the classification is wrong, then the extracted key features and target categories will also be different. Large difference can cause the CAM heatmap to not reflect the semantic information of the target category well, which makes the target of the location biased. Moreover, the performance of GoogLeNet-GAP is better than that of AlexNet-GAP and the results of VGGnet-GAP can also find the impact of image classification accuracy on location accuracy.

### E. FINE-GRAINED RECOGNITION

Fine-grained image recognition requires the discovery of multiple salient regions with identifying features. The above weakly supervised object localization method can be used to find a plurality of different levels of salient regions: Hide-CAM finally locates the region of the object as the overall feature of the object; the most discriminative region serves as a detail feature of the object. The complementary object regions in this chapter serve as another detail feature of the object. In this paper, fine-grained image recognition is realized by combining multiple features rich in different information. During the test, we averaged the combined results of the four CNNs to combine the output of the four CNNs. The specific combination scheme is shown in Figure 6.

In Figure 6, we use four different levels of features to achieve fine-grained image recognition. One is the image with background information, which is rich in a variety of contextual information and plays a role in the recognition of objects. The overall area of the object removes the background information, eliminating the interference of complex backgrounds or other objects on the recognition, and contributing to the overall information of the object. Hide-CAM provides a basis for the identification of fine-grained images. The first step is to extract the most discriminative area, which is often the representative part of the object, such as dogs and birds. Then we extract the next discriminative area, represented by part 2, which is generally the object with semantic information different from the most discriminative area, such as the dog’s leg or bird’s wings. The wings can effectively complement the details of the object and assist in the identification of fine-grained images.
We extract these four levels of images and unify the image size to 256×256 by bilinear interpolation and using random clipping to the size of 224×224. Then we send the image to the CNN network and finally weigh the results of the four classification networks. For fair comparison, we test it on the CUB-200-2011 dataset (a typical Fine-grained image recognition dataset), using ResNet50 as the backbone network. We use random crop, color jitter, and random flip as image augmentation methods. The classification results are shown in Table 3.

It can be found from the table that the direct input of the original image can achieve an accuracy of 82.04% on the CUB-200-2011 dataset. Feeding the image without background into the network can achieve 82.89%, which is 0.85% higher than the original image. It is indicated that the object area excludes the interference of background information and plays an important role in image recognition. When we use the original image and the image containing only the object as input to the network, the accuracy of the classification increases, reaching 83.44%, which is 1.4% higher than the original image. When the most discriminative region is added, the classification performance is increased by 0.92%, reflecting the importance of the salient region to fine-grained images. When the complementary object region is added, the classification performance is improved by 0.42%, indicating that the less significant area played a good auxiliary role in fine-grained image recognition. At the same time, it can be seen from the table that a plurality of detail areas rich in identification features (parts 1 and 2) are improved by 1.34% in fine-grained image recognition, which fully embodies the rationality of the method of the project.

<table>
<thead>
<tr>
<th>TABLE 3</th>
<th>FINE-GRAINED CLASSIFICATION RESULTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>Accuracy</td>
</tr>
<tr>
<td>Two-level Attention[33]</td>
<td>77.90%</td>
</tr>
<tr>
<td>FCAN[34]</td>
<td>82.00%</td>
</tr>
<tr>
<td>FOAF[35]</td>
<td>84.63%</td>
</tr>
<tr>
<td>PDFR[36]</td>
<td>84.50%</td>
</tr>
<tr>
<td>Image</td>
<td>82.04%</td>
</tr>
<tr>
<td>Object</td>
<td>82.89%</td>
</tr>
<tr>
<td>Image + Object</td>
<td>83.44%</td>
</tr>
<tr>
<td>Image + Object + Part 1</td>
<td>84.36%</td>
</tr>
<tr>
<td>Image + Object + Part 1 + Part 2</td>
<td><strong>84.78%</strong></td>
</tr>
</tbody>
</table>

In some existing methods, the method of this paper is also advantageous. Compared with some methods using detection or targets, this paper’s results are 7% higher than Two-level Attention. Compared with FCAN, FOAF, and PDFR are increased by 2.78%, 0.15%, and 0.23%, respectively. This also proves the rationality of the algorithm.

V. CONCLUSION

In this paper, we proposed Hide-CAM for weakly supervised location. We can discover the most discriminative region by first-step CAM and the complementary regions by hiding the most discriminative region. By combining the two regions, our Hide-CAM can find multiple discriminative regions to obtain the integral features.

Our method obtains comparable results: 61.64% location error for AlexNet-GAP, 51.61% location error for GoogLeNet-GAP, and 53.30% location error for VGGNet-GAP on ILSVRC 2012 validation. Our studies demonstrate that the most discriminative region and the complementary region is vital for weakly supervised location.

REFERENCES


This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2939267, IEEE Access