C²AM: Contrastive learning of Class-agnostic Activation Map for Weakly Supervised Object Localization and Semantic Segmentation

Jinheng Xie, Jianfeng Xiang, Junliang Chen, Xianxu Hou, Xiaodong Zhao, Linlin Shen*
School of Computer Science & Software Engineering, Shenzhen University, China
Dept. of Computer Science, Wenzhou-Kean University, Wenzhou, China
Shenzhen Institute of Artificial Intelligence of Robotics of Society, Shenzhen, China
Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, China
xiejinheng2020@email.szu.edu.cn, hxianxu@gmail.com, llshen@szu.edu.cn

Abstract

While class activation map (CAM) generated by image classification network has been widely used for weakly supervised object localization (WSOL) and semantic segmentation (WSSS), such classifiers usually focus on discriminative object regions. In this paper, we propose Contrastive learning for Class-agnostic Activation Map (C²AM) generation only using unlabeled image data, without the involvement of image-level supervision. The core idea comes from the observation that i) semantic information of foreground objects usually differs from their backgrounds; ii) foreground objects with similar appearance or background with similar color/texture have similar representations in the feature space. We form the positive and negative pairs based on the above relations and force the network to disentangle foreground and background with a class-agnostic activation map using a novel contrastive loss. As the network is guided to discriminate cross-image foreground-background, the class-agnostic activation maps learned by our approach generate more complete object regions. We successfully extracted from C²AM class-agnostic object bounding boxes for object localization and background cues to refine CAM generated by classification network for semantic segmentation. Extensive experiments on CUB-200-2011, ImageNet-1K, and PASCAL VOC2012 datasets show that both WSOL and WSSS can benefit from the proposed C²AM. Code will be available at https://github.com/CVI-SZU/CCAM.

1. Introduction

Massive image data and manual annotations are usually required to train deep neural networks for many vision tasks, e.g., object detection and semantic segmentation. However, it is time-consuming and labor-intensive to obtain bounding box or pixel-level annotations. In recent years, weaker supervision, e.g., image-level label, has been introduced in weakly supervised object localization (WSOL) and semantic segmentation (WSSS), which aims to achieve localization or segmentation with only image-level supervision, i.e., without bounding box or pixel-level annotations. Most previous WSOL and WSSS methods rely on class activation map (CAM) to estimate location of the target object. With image-level supervision, the classifier tries to find discriminative regions of target objects. Thus, though image-level labels enable CAM to indicate the correct location of target objects, they also limit the focus of CAM
Figure 2. Difference between (a) the class activation map (CAM) and (b) class-agnostic activation map. CAM consists of $K$ (number of class) activation maps and $C^2$AM only predicts one class-agnostic activation map for an image, which directly indicates the foreground and background regions. Best viewed in color.

on sparse and discriminative object regions. As a result, it could be very difficult to precisely estimate complete object regions by directly applying CAM to WSOL and WSSS.

Different methods have been introduced to mitigate the above problem of CAM. However, most of them are trained with image-level supervision, which would potentially affect the completeness of CAM. By contrast, this paper proposes to generate class-agnostic activation maps based on a novel cross-image foreground-background contrastive learning, without the requirement of image-level supervision. In contrast with $K$ class activation maps (CAM) (including one target activation map), $C^2$AM only predicts one class-agnostic activation map to indicate the foreground and background regions in an image. The difference between CAM and $C^2$AM is illustrated in Figure 2. Compared with CAM, the class-agnostic activation maps learned from cross-image foreground-background contrast, without the involvement of any image-level supervision, produce more reliable foreground regions.

As shown in Figure 1, the semantic information of the foreground object differs from its background, which can be presented as the large distance between foreground and background representations in the feature space. This can be further extended to the cross-image cases where the semantic information of the foreground object from one image should also be largely different from that of background in other images. By contrast, the distance among foregrounds with similar appearance or backgrounds with similar color/texture shall be small. Based on the above observations, we propose the cross-image foreground-background contrastive loss to force the network to disentangle the foreground object and background in an image with a class-agnostic activation map. Specifically, as shown in Figure 3 and 4, the network first generates a class-agnostic activation map using an activation head, such that the image representation can be disentangled into the foreground and background representations, respectively. Then, the foreground and background representations form the negative pairs and the foreground-foreground or background-background representations form the positive pairs. As only foregrounds with similar appearance or backgrounds with similar color/texture have similar representations in the feature space, pairs of foreground-foreground or background-background sharing less similar semantics might affect the learning of network. To mitigate this problem, we design a feature similarity based rank weighting to automatically reduce the influence of those dissimilar positive pairs. As the activation head is randomly initialized, the initial class-agnostic activation maps are random as well at beginning. When contrastive loss is applied to pull close and push apart the representations of positive and negative pairs, the class-agnostic activation map gradually separates the regions of foreground object and background in the image.

Extensive experiments conducted on the tasks of WSOL and WSSS show that the proposed $C^2$AM can replace or refine CAM for better performance. Specifically in WSOL, we follow [47] to divide WSOL into two tasks: class-agnostic object localization and object classification, and thus the class-agnostic activation maps can be used to extract class-agnostic object bounding boxes for localization. Background cues can also be extracted from the class-agnostic activation maps to refine the initial CAM such that false activation of background can be efficiently reduced to generate more reliable object regions. The refined CAM can thus greatly enhance the following segmentation performance in WSSS.

Collectively, the main contributions of this paper can be summarized as:

- We propose cross-image foreground-background contrastive learning to generate class-agnostic activation maps with unlabeled image data, which doesn’t involve any image-level supervision and produces much more reliable foreground object regions than CAM.
- The class-agnostic activation maps can be used to extract class-agnostic bounding boxes for accurate object localization and serve as background cues to improve the quality of initial CAM.
- Extensive experiments show that both WSOL and WSSS tasks can benefit from our method. We provide an alternative solution to replace or refine CAM for improvements of weakly supervised learning. Besides, it might be further used in a lot of vision tasks to detect foreground regions without manual annotations.

2. Related Works

Class activation mapping (CAM) [50] produces a spatial class activation map that indicates the object regions responsible for the prediction of object category. However, the classifier tends to perform recognition based on the most discriminative regions of the object [12, 30], which affects the precision of CAM based localization or segmentation.
Figure 3. The overall network architecture of the proposed method. The encoder network $h(\cdot)$ maps image $X_i$ to the feature map $Z_i$. In disentangler, the activation head $\varphi(\cdot)$ produces the class-agnostic activation map $P_i$. Suppose $P_i$ activates the foreground regions and the background activation map can be derived as $(1 - P_i)$. Based on the foreground and background activation maps, $Z_i$ can be disentangled into the foreground and background feature representations, i.e., $v_i^f$ and $v_i^b$. In evaluation, only the trained $h(\cdot)$ and $\varphi(\cdot)$ are used for generating the class-agnostic activation map $P_i$. Flatt.: matrix flattening; Trans.: matrix transpose; $\odot$: matrix multiplication.

**Weakly supervised object localization (WSOL).** Most of existing WSOL works consider the image-level labels in the training dataset [28, 35]. To solve the above problem, various methods [6, 24, 31, 42–44, 48, 49] have been proposed. Baek et al. [2] introduce class-agnostic activation mapping, which can effectively generate heat map from a self-supervised model. Xie et al. [42] propose a low-level feature based two-stage learning framework to refine coarse activation maps. Differently, Zhang et al. [47] divide WSOL into two tasks: 1) class-agnostic object localization and 2) object classification. This allows a network to complete localization or classification separately. They use DDT [37] to extract class-agnostic object bounding boxes as pseudo labels to train a localization model and use state-of-the-art networks for classification.

**Weakly supervised semantic segmentation (WSSS).** Most WSSS works follow a three-stage learning process: initial CAM generation, pseudo masks generation, and segmentation model training. Hou et al. [13] propose two self-erasing strategies to focus attention only on the reliable regions and generate complete initial CAM. Chang et al. [3] propose to mine more object parts by investigating object sub-categories. Jungbeom et al. [19] propose an anti-adversarial manner to discover more regions of the target object in the activation map. Ahn and Kwak [1] propose pixel-level semantic affinity, which helps to generate fine-grained pseudo masks. Works [15, 20, 23, 32, 38] additionally use background cues from a fully supervised saliency detector to get better pseudo masks. In our work, the class-agnostic activation maps generated by $C^2$AM are used to extract background cues to refine the initial CAM, which increases the potential of initial CAM for the next stage.

**Contrastive learning.** The idea of contrastive learning is to pull close the samples from the positive pair and to push apart the samples from the negative pair [4, 7, 10, 11, 29]. Based on object class labels, the positive pair is created with the samples from the same class, while the samples with different class labels form the negative pair [7, 10, 17, 29]. Unsupervised contrastive learning can be divided into two categories, i.e., instance-wise contrastive learning [4, 11, 39, 46] and clustering based contrastive learning [21, 41, 45]. In instance-wise contrastive learning, two samples augmented from the same instance form the positive pair, while samples augmented from different instances [4, 11] become the negative pair. For clustering based contrastive learning, the clustering algorithm is applied to generate the pseudo label for training samples, and then the supervised contrastive learning is applied [9].

3. Methodology

3.1. Architecture

The overall network architecture of $C^2$AM is shown in Figure 3. Given a batch of $n$ images $X_{1:n} = \{X_i\}_{i=1}^n$, the encoder $h(\cdot)$ maps $X$ to high-level feature maps $Z_{1:n} = \{Z_i\}_{i=1}^n$, in which $Z_i \in \mathbb{R}^{C \times H \times W}$. $C$ and $H \times W$ denote the channel number and spatial dimension, respectively. The popular network, e.g., ResNet [12] or VGG [30], is used as the encoder $h(\cdot)$. Supervised or unsupervised pretraining, e.g., moco [11] and detco [40], on ImageNet-1K [28] can be adopted as initialization of $h(\cdot)$. Based on the extracted feature maps $Z$, the disentangler employs an activation head $\varphi(\cdot)$ to produce the class-agnostic activation maps $P_{1:n} = \{P_i\}_{i=1}^n$, in which $P_i \in \mathbb{R}^{1 \times H \times W}$. Specifically, $\varphi(\cdot)$ is a $3 \times 3$ convolution with a batch normalization layer. Suppose $P_i$ activates foreground regions, the background activation map of the $i$-th sample can be formulated as $(1 - P_i)$. The foreground and background activation maps can finally disentangle the feature map $Z_{1:n}$ into the foreground and background feature representations, i.e., $v_{1:n}^f$ and $v_{1:n}^b$, respectively. For the $i$-th sample, $v_i^f$ and
\( v^b_i \) can be derived as:
\[
    v^f_i = P_i \otimes Z_i^T, \quad v^b_i = (1 - P_i) \otimes Z_i^T. \tag{1}
\]

Here, \( P_i \) and \( Z_i \) are flattened, i.e., \( P_i \in \mathbb{R}^{1 \times HW} \) and \( Z_i \in \mathbb{R}^{C \times HW} \). \( v^f_i \in \mathbb{R}^{1 \times C} \) and \( v^b_i \in \mathbb{R}^{1 \times C}. \otimes \) and \( \top \) indicate the matrix multiplication and transpose, respectively.

### 3.2. Foreground-background Contrast

Considering the fact that there is no label information in the training process, C3AM is proposed based on the idea of cross-image foreground-background contrast, which locates the foreground object regions by only utilizing the semantic information among foreground and background representations. As aforementioned, given an image, its foreground and background representations contain different semantic information and thus should have a large distance in the feature space. This observation is also true for cross-image cases. The distance between their foreground and background representations should also be large.

Based on these analyses, we propose to apply the contrastive learning to push apart foreground-background representations. Given \( n \) samples \( X_{1:n} \), the disentangler separates them into \( n \) foreground and \( n \) background representations, i.e., \( v^f_{1:n} \) and \( v^b_{1:n} \). The foreground-background representation pair, i.e., \((v^f_i, v^b_j)\), is treated as the negative pair. The negative contrastive loss is designed as:
When contrastive loss is applied to pull close and push apart the representations in positive and negative pairs, the class-agnostic activation map gradually separates the regions of foreground object and background in the image.

The contrastive loss is formulated as the summation of $\mathcal{L}_{POS}$ and $\mathcal{L}_{NEG}$:

$$
\mathcal{L} = \mathcal{L}_{POS} + \mathcal{L}_{NEG}.
$$

How to determine foreground regions. As the same contrastive loss is applied to foreground-background and background-background positive pairs, it can not guarantee that foreground or background regions are activated in $P_i$. To solve this problem, we set a threshold to binarize the class-agnostic activation maps and detect the largest contour to determine the object regions.

3.4. Weakly Supervised Object Localization

We follow the route of PSOL [47] to adapt our C$^2$AM to WSOL. Specifically, WSOL is divided into two tasks: class-agnostic object localization and object classification. PSOL uses DDT [37] to generate class-agnostic object bounding boxes on the training set. Given a group of images of the same category, DDT obtains those category-consistent regions to extract class-agnostic object bounding boxes. By contrast, the proposed C$^2$AM directly learns class-agnostic activation maps from the whole dataset without any manual annotations. We set a threshold to binarize the class-agnostic activation maps and then extract class-agnostic object bounding boxes as pseudo labels (we present a fair comparison between DDT and C$^2$AM in the next section). The localization model is trained with these pseudo labels for object bounding box predictions. The popular network, e.g., EfficientNet [33], is adopted for object classification.

3.5. Weakly Supervised Semantic Segmentation

We first use CAM-based methods to generate initial CAM for each image and then apply C$^2$AM to refine it. Specifically, we use the background activation maps ($1 - P$) as pseudo labels to further train a model to predict the background regions, i.e., background cues, in the image. As shown in Figure 5, we concatenate the predicted background cues with the initial CAM and perform the argmax process along the channel dimension to refine initial CAM. This helps to reduce the false activation of background and activate more foreground regions in initial CAM. More details of the pseudo training and refinement procedure are included in supplementary materials. We just use a simple way to demonstrate the effectiveness of our C$^2$AM for refinement of initial CAM and leave more works in the future.

4. Experiments

4.1. Experimental Setup

Dataset. CUB-200-2011 [35] is a fine-grained classification dataset, which consists of 200 species of birds with 5,994 images for training and 5,794 images for test. ImageNet-1K [28] is a large scale visual recognition dataset with 1,000 classes, which contains 1,281,197 training images and 50,000 validation images. PASCAL VOC2012 [8] is a popular semantic segmentation dataset with 20 object categories. It consists of 1,464 images for training, 1,449 images for validation, and 1,456 images for test.

Evaluation Metrics. In WSOL, we follow [50] to use Top-1 localization accuracy (Top-1 Loc), Top-5 localization accuracy (Top-5 Loc), and GT-known localization accuracy (GT-known Loc) for evaluation. For GT-known Loc, a localization is correct when the predicted bounding box overlaps 50% with one of the ground truth bounding boxes belonging to the same class. For Top-1 Loc, a prediction is correct when the Top-1 classification result and GT-known Loc are both correct. For Top-5 Loc, a prediction is correct when one of the Top-5 classification results and GT-known Loc are both correct. The MaxBoxAccV2 proposed in [5] is also adopted, which averages the performance across different bounding boxes ratios, i.e., 30%, 50%, and 70%. In WSSS, the Intersection over Union (IoU) and mean Intersection over Union (mIoU) is adopted as the evaluation metric.

The implementation details are provided in the supplementary materials.

5. Results and Analysis

We first present a visual comparison between CAM and the proposed C$^2$AM to validate that more complete and correct foreground regions can be predicted in the class-agnostic activation maps generated by C$^2$AM. Then, we apply C$^2$AM to WSOL and show that CAM can be replaced by the class-agnostic activation maps for more accurate object localization. Besides, we also apply C$^2$AM to
Figure 6. Visual comparison between CAM and the class-agnostic activation maps generated by C2AM. Best viewed in color.

Table 1. Comparison of the performance between the proposed method and the state-of-the-art methods on CUB-200-2011 test set and ImageNet-1K validation set. Loc Back. denotes the localization backbone. Cls Back. denotes the backbone for classification. † and †† indicate the $h(\cdot)$ is initialized by the supervised and unsupervised (moco) pretraining, respectively. ∗: our re-implementation results.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Top-1 Loc</td>
<td>Top-5 Loc</td>
</tr>
<tr>
<td>CAM CVPR'16</td>
<td>VGG-GAP</td>
<td></td>
<td>36.13</td>
<td>-</td>
</tr>
<tr>
<td>ADL CVPR'19</td>
<td>VGG-GAP</td>
<td></td>
<td>52.36</td>
<td>-</td>
</tr>
<tr>
<td>GCNet ICCV'20 [49]</td>
<td>GoogLeNet</td>
<td></td>
<td>65.99</td>
<td>68.34</td>
</tr>
<tr>
<td>SPA CVPR'21</td>
<td>VGG16</td>
<td></td>
<td>60.27</td>
<td>72.5</td>
</tr>
<tr>
<td>FAM ICCV'21</td>
<td>VGG16</td>
<td></td>
<td>69.26</td>
<td>-</td>
</tr>
<tr>
<td>ORNet ICCV'21 [42]</td>
<td>VGG16</td>
<td></td>
<td>67.74</td>
<td>80.77</td>
</tr>
<tr>
<td>C2AM CVPR'16 [50]</td>
<td>DenseNet161</td>
<td></td>
<td>29.81</td>
<td>39.85</td>
</tr>
<tr>
<td>PSOL CVPR'20 [47]</td>
<td>ResNet50</td>
<td></td>
<td>70.68</td>
<td>86.64</td>
</tr>
<tr>
<td>PSOL CVPR'20 [47]</td>
<td>DenseNet161</td>
<td>EfficientNet-B7</td>
<td>80.89*</td>
<td>89.97*</td>
</tr>
<tr>
<td>C�AM(Our)†</td>
<td>DenseNet161</td>
<td>EfficientNet-B7</td>
<td>83.28</td>
<td>92.74</td>
</tr>
<tr>
<td>C�AM(Our)††</td>
<td>DenseNet161</td>
<td>EfficientNet-B7</td>
<td>81.76</td>
<td>91.11</td>
</tr>
</tbody>
</table>

WSSS and show that the class-agnostic activation maps can serve as background cues to reduce false-activation of background in CAM and segment more reliable object regions. Details are presented in the following sections.

Table 2. The GT-Known Loc of DDT and our C�AM on the CUB-200-2011 and ImageNet-1K. Init. denotes the parameters initialization. sup-pre.: the parameters initialized from supervised pretraining. moco [11] and detco [40]: the unsupervised pretraining. ∗: the result is obtained with our re-implementation.

<table>
<thead>
<tr>
<th>Method</th>
<th>Bac.</th>
<th>Init.</th>
<th>CUB-200-2011</th>
<th>ImageNet-1K</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Top-1 Loc</td>
<td>Top-5 Loc</td>
</tr>
<tr>
<td>DDT</td>
<td>VGG16</td>
<td>sup-pre.</td>
<td>84.55</td>
<td>61.41</td>
</tr>
<tr>
<td>C�AM</td>
<td>VGG16</td>
<td>sup-pre.</td>
<td>75.34</td>
<td>63.41</td>
</tr>
<tr>
<td>DDT</td>
<td>ResNet50</td>
<td>sup-pre.</td>
<td>72.39</td>
<td>59.92</td>
</tr>
<tr>
<td>C�AM</td>
<td>ResNet50</td>
<td>sup-pre.</td>
<td>88.99</td>
<td>65.89</td>
</tr>
<tr>
<td>DDT†</td>
<td>ResNet50</td>
<td>moco</td>
<td>34.95</td>
<td>40.52</td>
</tr>
<tr>
<td>C�AM</td>
<td>ResNet50</td>
<td>moco</td>
<td>89.90</td>
<td>66.51</td>
</tr>
<tr>
<td>DDT†</td>
<td>ResNet50</td>
<td>detco</td>
<td>35.86</td>
<td>41.23</td>
</tr>
<tr>
<td>C�AM</td>
<td>ResNet50</td>
<td>detco</td>
<td>88.21</td>
<td>65.48</td>
</tr>
</tbody>
</table>

Table 3. Evaluation results in terms of MaxBoxAccV2.

<table>
<thead>
<tr>
<th>Methods</th>
<th>ImageNet</th>
<th>CUB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best WSOL</td>
<td>60.6</td>
<td>63.9</td>
</tr>
<tr>
<td>C�AM(Our)†</td>
<td>66.3</td>
<td>65.8</td>
</tr>
<tr>
<td></td>
<td>81.4</td>
<td>82.4</td>
</tr>
</tbody>
</table>

5.1. Visual Comparison with CAM

Figure 6 presents the visual comparison between CAM and the proposed C�AM on ImageNet-1K and PASCAL VOC2012 datasets. The second and fifth columns present the visualization of CAM. It can be seen that with image-level supervision, CAM can correctly indicate the location of the object but usually focus only on the most discriminative regions of the target object, e.g., the head of bird. Besides, CAM also activates class-related regions, e.g., the railroad, and ignores the complete body of the train. These issues degrade the performance of CAM in WSOL and WSSS. In contrast, the class-agnostic activation maps gen-
Table 4. Semantic segmentation performance (IoU(\%)) comparison for 11 categories between the initial CAM and CAM refined by using the proposed C2AM. The initial CAM is generated using PSA, SC-CAM, SEAM, PuzzleCAM, and AdvCAM, respectively. †† indicates that unsupervised pretraining (moco) is adopted to initialize the backbone network $h(\cdot)$ of C2AM.

<table>
<thead>
<tr>
<th>Method</th>
<th>bgk</th>
<th>aero</th>
<th>bike</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>bus</th>
<th>car</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSA [1]</td>
<td>79.0</td>
<td>41.0</td>
<td>27.5</td>
<td>42.0</td>
<td>34.4</td>
<td>44.5</td>
<td>44.5</td>
<td>63.3</td>
<td>53.3</td>
<td>43.2</td>
<td>38.9</td>
</tr>
<tr>
<td>+C2AM††</td>
<td>85.0</td>
<td>75.4</td>
<td>45.1</td>
<td>79.5</td>
<td>52.6</td>
<td>64.8</td>
<td>78.0</td>
<td>75.2</td>
<td>70.4</td>
<td>80.8</td>
<td>75.8</td>
</tr>
<tr>
<td>SC-CAM [4]</td>
<td>76.8</td>
<td>42.1</td>
<td>29.2</td>
<td>44.5</td>
<td>37.0</td>
<td>56.1</td>
<td>69.9</td>
<td>58.2</td>
<td>59.8</td>
<td>72.8</td>
<td>52.5</td>
</tr>
<tr>
<td>+C2AM††</td>
<td>85.0</td>
<td>74.8</td>
<td>46.6</td>
<td>80.5</td>
<td>52.8</td>
<td>68.5</td>
<td>79.0</td>
<td>74.6</td>
<td>84.7</td>
<td>81.3</td>
<td>76.3</td>
</tr>
<tr>
<td>SEAM [15]</td>
<td>82.5</td>
<td>51.0</td>
<td>35.4</td>
<td>57.4</td>
<td>31.6</td>
<td>50.0</td>
<td>57.9</td>
<td>63.2</td>
<td>62.4</td>
<td>25.3</td>
<td>63.0</td>
</tr>
<tr>
<td>+C2AM††</td>
<td>87.1</td>
<td>71.9</td>
<td>44.2</td>
<td>77.9</td>
<td>45.6</td>
<td>61.9</td>
<td>75.5</td>
<td>70.4</td>
<td>82.5</td>
<td>22.5</td>
<td>78.0</td>
</tr>
<tr>
<td>PuzzleCAM [16]</td>
<td>75.8</td>
<td>43.2</td>
<td>32.3</td>
<td>56.7</td>
<td>23.1</td>
<td>51.7</td>
<td>67.6</td>
<td>61.4</td>
<td>76.7</td>
<td>16.7</td>
<td>60.4</td>
</tr>
<tr>
<td>+C2AM††</td>
<td>87.8</td>
<td>71.5</td>
<td>48.4</td>
<td>78.0</td>
<td>44.6</td>
<td>66.7</td>
<td>77.3</td>
<td>74.3</td>
<td>85.7</td>
<td>21.3</td>
<td>80.2</td>
</tr>
<tr>
<td>AdvCAM [19]</td>
<td>81.5</td>
<td>50.6</td>
<td>33.5</td>
<td>57.5</td>
<td>37.0</td>
<td>53.3</td>
<td>64.8</td>
<td>64.7</td>
<td>35.0</td>
<td>68.4</td>
<td>79.5</td>
</tr>
<tr>
<td>+C2AM††</td>
<td>87.5</td>
<td>72.9</td>
<td>46.4</td>
<td>78.9</td>
<td>50.7</td>
<td>60.5</td>
<td>77.8</td>
<td>71.2</td>
<td>84.5</td>
<td>26.8</td>
<td>79.5</td>
</tr>
</tbody>
</table>

Table 5. Semantic segmentation performance (mIoU(\%)) comparison among different methods. dCRF: dense CRF [18]. FSSD: background cues from fully supervised saliency detector [22]. ††: moco is adopted to initialize the backbone network $h(\cdot)$ of C2AM.

<table>
<thead>
<tr>
<th>Method</th>
<th>CAMs</th>
<th>+dCRF</th>
<th>+FSSD</th>
<th>+C2AM††</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSA CVPR'18 [1]</td>
<td>48.0</td>
<td>-</td>
<td>62.0</td>
<td>65.5 (+17.5)</td>
</tr>
<tr>
<td>SC-CAM CVPR'20 [3]</td>
<td>50.9</td>
<td>55.3</td>
<td>62.1</td>
<td>66.0 (+15.1)</td>
</tr>
<tr>
<td>SEAM CVPR'20 [30]</td>
<td>55.4</td>
<td>56.8</td>
<td>61.5</td>
<td>63.9 (+8.5)</td>
</tr>
<tr>
<td>PuzzleCAM ICPR'21 [16]</td>
<td>51.5</td>
<td>-</td>
<td>62.1</td>
<td>65.5 (+14.0)</td>
</tr>
<tr>
<td>AdvCAM CVPR'21 [19]</td>
<td>55.6</td>
<td>62.1</td>
<td>62.3</td>
<td>65.4 (+9.8)</td>
</tr>
</tbody>
</table>

optimized by C2AM effectively mitigate the above issues. As shown in the left of Figure 6, C2AM can completely separate the foreground objects out of background regions, in which the whole body of bird is activated. Besides, as shown in the right of Figure 6, C2AM successfully discriminate the railroad from the train and thus the complete region of train can be activated.

5.2. Results of WSOL

Class-agnostic object bounding boxes. Table 2 compares the quality of class-agnostic bounding boxes generated by C2AM and DDT [37]. When supervised pretraining on ImageNet-1K are adopted as initialization of $h(\cdot)$, class-agnostic bounding boxes generated by C2AM have a higher GT-known Loc than DDT (except the backbone of VGG16). This indicates that our method has a good capability to generate higher quality class-agnostic bounding boxes than DDT. When considering unsupervised pretraining, e.g., moco [11] and detco [40], the GT-known Loc of DDT on two benchmarks has a large drop. By contrast, our C2AM maintains high performance, which is even better than that of DDT using supervised pretraining.

Results on CUB-200-2011. Table 1 compare the performance of C2AM and state-of-the-art methods on CUB-200-2011. Our method achieves Top-1 Loc of 83.28%, Top-5 Loc of 92.74%, and GT-known Loc of 94.46%. When compared with SPA and ORNet, our C2AM surpasses them by a large margin in terms of Top-1 Loc, Top-5 Loc, and GT-known Loc. Compared with PSOL, our C2AM with supervised pretraining surpasses it by 2.39%, 2.77%, and 2.68%, in terms of Top-1 Loc, Top-5 Loc, and GT-known Loc, respectively. When $h(\cdot)$ is initialized by moco, C2AM also achieves better performance than PSOL. We also provide the results of MaxBoxAccV2 metric in Table 3, where the result of Best WSOL is directly taken from [5]. One can observe that our C2AM achieves significantly better performance for different networks and datasets.

Results on ImageNet-1K. Table 1 also compares our method with other recently introduced weakly supervised object localization methods on ImageNet-1K validation set. Our method achieves Top-1 Loc of 59.56%, Top-5 Loc of 67.05%, and GT-known Loc of 68.53%, which achieves the state-of-the-art performance. When using the same localization (DenseNet161 [14]) and classification backbone (EfficientNet-B7 [33]), the performance of C2AM on the validation set significantly surpasses PSOL by 1.56%, 2.03%, and 2.25% in terms of Top-1 Loc, Top-5 Loc, and GT-known Loc, respectively.

5.3. Results of WSSS

Results on PASCAL VOC2012. As this section is designed to demonstrate the effectiveness of C2AM to refine initial CAM and the refined CAM can be further used in other stages of WSSS, the IoU and mIoU are reported on the training set of PASCAL VOC2012. Table 4 presents improvement of each category for existing CAM-based methods by using C2AM. It can be seen that the segmentation performances of many CAM-based methods for most categories, including PSA [1], SC-CAM [3], SEAM [36], PuzzleCAM [16], and AdvCAM [19], are largely improved by using C2AM. For example, there is a mean of 14.6%, 31.3%, 20.1% IoU improvements of these five methods for bike, bird, and cow, respectively.

To compare our C2AM with other CAM refinement approaches, Table 5 presents the segmentation performance of initial CAMs generated by literature works like PSA, SC-CAM, and PuzzleCAM etc., and the improvements after they are refined by dense CRF (+dCRF) [18], and background cues generated by a fully supervised saliency detector (+FSSD) [22], and our C2AM (+C2AM). The results of dense CRF are reported in their papers and the refinement process presented in Section 3.5 is used for both +FSSD and +C2AM. It can be clearly seen that the quality of initial CAM has been greatly improved by using our C2AM. In particular, the mIoU of initial CAM generated by PSA [1] and SC-CAM [3] have been improved by 17.5% and 15.1%, respectively. In particular, the increase of mIoU caused by background cues generated from C2AM is even higher than that generated by the fully supervised saliency detector.

We also provide a visualization of CAM refinement in Figure 7. It can be observed in the first column that, a lot of object regions, e.g., the body of the white cow and the feet...
of the brown horse, are not activated in the initial CAM. Besides, backgrounds, e.g., the grass and ground, are usually falsely activated. To address these issues, we use the class-agnostic activation maps to extract background cues (as shown in the second column) to efficiently reduce the false activation of background and predict more reliable object regions. The results shown in the third column highly overlaps with the ground truth (in the last column).

### 5.4. Sensitivity Analysis

We conduct experiments on ImageNet-1K to examine the hyper-parameter sensitivity of C²AM. The GT-known Loc is adopted as a performance evaluation.

**Initialization of Activation Head.** A $3 \times 3$ convolution with a batch normalization layer is adopted as the activation head $\varphi(\cdot)$. The backbone network $h(\cdot)$ can be initialized with supervised or unsupervised pretraining (e.g., moco, detco). How about the activation head? Does the initialization of activation head affect the performance of C²AM? We perform experiments with four different parameters initialization of activation head, i.e., kaiming_normal, kaiming_uniform, normal, and uniform initialization. The weights and bias in batch normalization layer are initialized with 1 and 0. Table 6 compares the performance of these four initialization choices. We can see that the performances are quite stable among these initialization methods, which shows that the activation head of C²AM is not sensitive to the parameters initialization.

**Hyper-parameter $\alpha.$** Figure 8 presents the GT-known Loc of C²AM with different setting of $\alpha$ in Eq. 5. The results are reported on ImageNet-1K validation set. When $\alpha$ equals 0, every $w_{i,j}$ is equal to 1, and every positive pair including those foreground with different appearance and background with significant different color/texture are also pulled together. This violates the assumption that only representations with similar appearance, color or texture can be pulled close. Thus, the GT-known Loc is the lowest when $\alpha$ is set to 0. When $\alpha$ is larger than 0, different positive pairs are assigned with different weights based on the feature similarity. As seen, the Gt-Known Loc has been largely improved when $\alpha$ is larger than 0, demonstrating that the proposed rank weighting works well to automatically reduce the influence of those dissimilar positive pairs. When $\alpha$ varies from 0.1 to 0.9, the number of positive pairs involved for contrastive learning is gradually decreasing. The stable performances show that our approach is not sensitive to the number of positive pairs (i.e., the setting of $\alpha$).

### 6. Conclusions and Discussions

We propose cross-image foreground-background contrast for class-agnostic activation maps generation using unlabeled image data. Class-agnostic activation map determines more reliable foreground regions, which can be used to replace or refine CAM for better WSOL and WSSS performance. Extensive experiments show that both WSOL and WSSS can benefit from our approach.

As the proposed C²AM works well in the unlabeled training data, we believe that it can be further used to efficiently detect the foreground regions without manual annotations in a lot of vision tasks, e.g., saliency detection and skin lesion segmentation, etc.

### Acknowledgments

The work was supported by the National Natural Science Foundation of China under grant 91959108.
References


[15] Peng-Tao Jiang, Qibin Hou, Yang Cao, Ming-Ming Cheng, Yunchao Wei, and Hong-Kai Xiong. Integral object mining via online attention accumulation. In ICCV, pages 2070–2079, 2019. 3


[38] Tong Wu, Junshi Huang, Guangyu Gao, Xiaoming Wei, Xiaolin Wei, Xuan Luo, and Chi Harold Liu. Embedded discriminative attention mechanism for weakly supervised semantic segmentation. In CVPR, pages 16765–16774, 2021. 3


