Activation Modulation and Recalibration Scheme for Weakly Supervised Semantic Segmentation

Jie Qin\textsuperscript{1,2,3*}, Jie Wu\textsuperscript{2}, Xuefeng Xiao\textsuperscript{2}, Lujun Li\textsuperscript{3}, Xingang Wang\textsuperscript{3}

\textsuperscript{1} School of Artificial Intelligence, University of Chinese Academy of Sciences
\textsuperscript{2} ByteDance Inc \textsuperscript{3} Institute of Automation, Chinese Academy of Sciences
\{qinjie2019, lilujun2019, xingang.wang\}@ia.ac.cn, \{wujie.10, xiaoxuefeng.ailab\}@bytedance.com

Abstract

Image-level weakly supervised semantic segmentation (WSSS) is a fundamental yet challenging computer vision task facilitating scene understanding and automatic driving. Most existing methods resort to classification-based Class Activation Maps (CAMs) to play as the initial pseudo labels, which tend to focus on the discriminative image regions and lack customized characteristics for the segmentation task. To alleviate this issue, we propose a novel activation modulation and recalibration (AMR) scheme, which leverages a spotlight branch and a compensation branch to obtain weighted CAMs that can provide recalibration supervision and task-specific concepts. Specifically, an attention modulation module (AMM) is employed to rearrange the distribution of feature importance from the channel-spatial sequential perspective, which helps to explicitly model channel-wise interdependencies and spatial encodings to adaptively modulate segmentation-oriented activation responses. Furthermore, we introduce a cross pseudo supervision for dual branches, which can be regarded as a semantic similar regularization to mutually refine two branches. Extensive experiments show that AMR establishes a new state-of-the-art performance on the PASCAL VOC 2012 dataset, surpassing not only current methods trained with the image-level of supervision but also some methods relying on stronger supervision, such as saliency label. Experiments also reveal that our scheme is plug-and-play and can be incorporated with other approaches to boost their performance. Our code is available at: https://github.com/jieqin-ai/AMR.

Introduction

Semantic segmentation is a fundamental and crucial task due to extensive applications in the field of computer vision. It aims to perform a pixel-level prediction to cluster parts of an image together that belong to the same object class. Although with varying degrees of progress, most of its recent successes (Chen et al. 2017, 2018) are involved in a fully supervised setting. It is still arduous to acquire such granular pixel-level annotations that require a huge amount of manual effort. To alleviate such expensive and unwieldy annotations, many works tend to resort to weakly supervised manner (Wu et al. 2020, 2021), such as bounding boxes super-

\*This work was done when Jie Qin interned at Bytedance Inc.

Figure 1: Visualizations of CAMs in the AMR scheme. “Spotlight CAMs” focus more on the discriminative regions similar to conventional CAMs. “Compensation CAMs” help to dig out the important but easily ignored regions. The spotlight CAMs are recalibrated by the compensation CAMs and further obtain the “Weighted CAMs”, which contribute to providing more segmentation-oriented concepts.
quire more seed regions. Similarly, (Wei et al. 2017) presses the model to concentrate on the other regions by iteratively erasing the seeds of CAMs. However, these methods usually formulate the expanding process as a complex training stage, e.g., the iterative erasing manner is time-consuming and difficult to determine the best number of iterations. Furthermore, it heavily relies on the discriminative regions provided via the classification networks, which easily fails to take the minor important regions into account.

To better cope with the above issues, we propose a novel Activation Modulation and Recalibration scheme, termed AMR. The scheme leverages a spotlight branch and a compensation branch to provide complementary and task-oriented CAMs for WSSS. The spotlight branch denotes the fundamental classification network to produce CAMs, which usually highlight the discriminative and classification-specific regions, such as the head of horse and the window of the car (refer to Fig. 1). AMR alleviates the task gap issue of using classification-based CAMs to perform segmentation tasks in previous works, which contributes to providing more semantic segmentation-specific cues. Moreover, an attention modulation module (AMM) is employed to rearrange the distribution of activation importance from the channel-spatial sequential perspective, which contributes to modulating segmentation-oriented activation responses adaptively. The contributions of AMR can be summarized as follows:

- To the best of our knowledge, we offer the first attempt to explore a plug-and-play compensation branch to provide complementary supervision and task-specific CAMs in WSSS. The compensation branch can dig out the essential regions for segmentation (such as the legs of the horse and the chassis of the car in Fig. 1), which is very critical to break through the bottleneck of classification-based CAMs for applying in the segmentation task. The compensation CAMs assist in generating the segmentation-oriented CAMs by recalibrating the spotlight CAMs. Additionally, we introduce a cross pseudo supervision to optimize the output CAMs from dual branches, which can be viewed as the semantic similar regularization to avoid the compensation CAMs concentrating on the background and force it close to spotlight CAMs.
- We design an attention modulation module (AMM), which encourages the activation maps to pay equal attention to the whole target objects by performing feature modulation in the channel and spatial dimensions sequentially. A modulation function is leveraged to rearrange the distribution of activation features, which attempts to emphasize minor features and penalize the saliency features that have been captured by the spotlight branch. The channel-spatial sequential manner contributes to explicitly modelling channel-wise interdependencies and spatial encodings within local receptive fields at each layer to adaptively modulate segmentation-oriented features responses.
- Our approach achieves 68.8% and 69.1% in terms of mIoU on validation and test set, which establishes a new state-of-the-art performance in WSSS on the PAS-CAL VOC2012 dataset (Everingham et al. 2015). Extensive experiments show that AMR surpasses not only current methods trained with the image-level supervision but also some methods relying on stronger supervision, such as saliency label. Experiments also reveal that our scheme is plug-and-play and can be incorporated with other approaches to boost their performance.

### Related Work

**Weakly Supervised Semantic Segmentation**

With the refined research of semantic segmentation, on the one hand, AutoML (Li et al. 2021; Ren et al. 2021; Li et al. 2020, 2019; Xuefeng Xiao and Lianwen Jin 2017; Xia and Ding 2020) based technologies are employed to improve the segmentation quality. On the other hand, training with lightweight annotation cost is widely explored, image-level WSSS has been extensively studied in recent years. Existing advanced methods usually rely on the seed area of Class Activation Maps (CAMs) (Zhou et al. 2016) generated by the classification networks. Most of these efforts can be classified in two aspects: generating high-quality CAM seeds and refining the pseudo labels. On the one hand, some approaches directly expand the response regions of CAMs because the original activation maps only highlight the discriminative regions of the images. (Wei et al. 2018) uses dilated convolution with different dilate rates to increase the target regions. SEAM (Wang et al. 2020b) captures different regions from transformed images via equivariance regularization in classification networks. On the other hand, some works focus on refining the pseudo labels based on the initial CAMs. SEC (Kolesnikov and Lampert 2016) explores three principles to refine the seeds, i.e., seed, expansion, and constraining. AffinityNet (Ahn and Kwak 2016) learns the relations of pixels and propagates the similar semantic pixels by a random walk algorithm. In addition, several methods (Yao et al. 2021; Lee et al. 2019) take the CAMs as foreground cues and saliency maps (Zhang et al. 2019) as background cues. (Yao et al. 2021) introduces a graph-based global reasoning unit to discover the objects in the non-salient regions. However, these approaches are formulated in an iterative and random manner, which may lose essential information. To alleviate this issue, we propose an activation modulation and recalibration scheme to generate high-quality CAMs.

### Attention Mechanism

The attention mechanism (Wu, Hu, and Yang 2019; Wu, Hu, and Wu 2018) has been widely used in segmentation networks to build the global context relation of images. Non-local (Wang et al. 2018a) is the first to take account of the correlation between each spatial point in the feature maps. Then, asymmetric (Zhu et al. 2019) proposes an asymmetric non-local network to strengthen the connection of non-local networks. SE (Hu, Shen, and Sun 2018) learns the importance of channel features by computing the interactions between channels. Following this work, (Wang et al. 2020a) uses a channel-based convolution to learn the interactions. CBAM (Woo et al. 2018) exploits the spatial-wise and the channel-wise attention to highlight the important cues in the...
channel and spatial dimension. (Cao et al. 2019) incorporates long-range dependencies to the fundamental attention module. In this paper, we introduce an attention modulation module to enhance the minor but essential features for the segmentation task.

Methodology

In this section, we first briefly introduce the conventional method for CAMs generation. Then we illustrate the activation modulation and recalibration scheme (AMR). The motivation and details of the proposed AMM is introduced in the next section. Finally, the modulation function and training loss functions are illustrated.

Preliminary

Class Activation Maps (CAMs) (Zhou et al. 2016) denote the response regions of specific classes for the input images $I \in \mathbb{R}^{3 \times H \times W}$. A multi-label classification network is employed for encoding the features of all classes, which can be leveraged to extract the feature maps $F(I) \in \mathbb{R}^{C \times H \times W}$ before the last classification layer to obtain CAMs. $C$ indicates the channel numbers of features maps. Then we simply perform matrix multiplication on $F(I)$ to generate CAMs:

$$M(I) = w^T_N F(I),$$

where $M(I) \in \mathbb{R}^{N \times H \times W}$ is the obtained CAMs. $w^T_N$ is the weight of the last fully-connected layer for $N$ classes.

However, such CAMs are classification-oriented and ignore the task-specific of semantic segmentation. Namely, the network is optimized via classification-based loss, which resorts to some discriminative regions of the full objects to accomplish the classification task. It will sacrifice the performance of weakly supervised semantic segmentation, which needs to obtain the holistic bound of the whole object. To address this issue, we propose the Activation Modulation and Recalibration (AMR) scheme to recalibrate initial CAMs to be more task-specific.

Activation Modulation and Recalibration Scheme

We illustrate the activation modulation and recalibration (AMR) scheme in Fig. 2. The AMR consists of the spotlight branch and the compensation branch. The spotlight branch is similar to the previous methods (Wei et al. 2017; Jiang et al. 2019; Lee, Kim, and Yoon 2021), which employs the classification loss to optimize itself and generate the spotlight CAMs $M_S$. Because the spotlight branch frequently activates the informative features during the training procedure, the obtained CAMs mainly highlight the discriminative regions of target objects.

The compensation branch is craftily designed to play as
auxiliary supervision for the spotlight CAMs. It alleviates the task gap issue of using classification-based CAMs to perform segmentation tasks in previous work, which contributes to providing more semantic segmentation-special cues. The compensation branch can be regarded as a plug-and-play component, which can dig out the essential regions for segmentation that are easily ignored by the spotlight branch. The obtained compensation CAMs $M_C$ helps to recalibrate the spotlight CAMs $M_S$ to generate the final weighted CAMs $M_W$, which is illustrated as:

$$M_W(I) = \xi M_S(I) + (1 - \xi) M_C(I), \quad (2)$$

where $\xi$ denotes the recalibration coefficient.

**Attention Modulation Module**

The attention modulation module (AMM) is proposed to assist the compensation branch to extract more regions essential for semantic segmentation tasks. As shown in Fig 2, AMM consists of channel attention modulation and spatial attention modulation. We firstly feed features $F(I)$ to the channel AMM. The channel interdependencies are explicitly modeled by the average pooling and the convolutional layer, which reflect the sensitivity to informative features. Inspired by (Jiang et al. 2019), the most sensitive features correspond to the discriminative regions, the minor features denote the important but easily ignored regions, and the insipid features may indicate the background concepts. Therefore, we exploit a modulation function to enhance the minor features and restrain the most and least sensitive features. The above operations can be denoted as:

$$A_c = \mathcal{G}(H(P_s(F(I))))), \quad (3)$$

where $A_c$ is the channel attention map. We denote $P_s$ as the spatial average pooling function and $H$ as the convolution layer. Then the modulation function $\mathcal{G}$ is leveraged to reassign the distribution of features to highlight the minor features in the channel dimension.

Then we conduct an element-wise multiplication between the channel attention maps and input feature maps to generate the redistributed features, which is defined as:

$$F_c(I) = \tilde{A}_c \odot F(I), \quad (4)$$

where $\tilde{A}_c$ denotes the channel attention maps which are expanded to the dimensions of feature maps. $F_c(I)$ represents the output feature maps.

To further model inter-spatial relationship in the spatial dimension, we also introduce a spatial AMM to cascade after the channel AMM. Specifically, we first employ a channel average pooling $P_c$ on $F_c(I)$ in channel dimension and then apply a convolution operation $H$ to them. The output feature maps illustrate the importance of the features in the spatial dimensions. Then we perform a modulation function on the output feature maps to increase the minor activations.

The implementation process can formulate as:

$$A_s = \mathcal{G}(H(P_c(F_c(I))))), \quad (5)$$

where $A_s$ is the spatial attention map. The high activation values in $A_s$ reflect the easily ignored regions. Then we make an element-wise multiplication between the spatial attention maps and the feature maps:

$$F_s(I) = \tilde{A}_s \odot F_c(I), \quad (6)$$

where $\tilde{A}_s$ denotes the spatial attention maps that are expanded to the dimensions of feature maps.

**Modulation Function**

In AMM, we employ the modulation function to redistribute the activation values of feature maps:

$$V_A = \mathcal{G}(V_{A_f}), \quad (7)$$

where $\mathcal{G}$ represents the gaussian function, which maps all activation values into a gaussian distribution. The parameters of “mean” and “std” are calculated by the values of $V_{A_f}$:

$$\mu = \frac{1}{M} \sum_{i=1}^{M} (V_{A_f}^{i}), \quad \sigma = \sqrt{\frac{1}{M} \sum_{i=1}^{M} (V_{A_f}^{i} - \mu)^2} \quad (8)$$

where $\mu$ and $\sigma$ are the mean and standard deviation of activation maps. We follow the setting of $\mu$ and $\sigma$ to project the activation values in $\mathcal{G}$.

We visualize the distribution of activations before and after the modulation in Fig. 3. We observe that the gaussian projection greatly suppresses the most and the least important activations. And it emphasizes the minor activations to extract the easily-ignored regions directly, which is crucial for the segmentation task. In addition, we also explore directly set the thresholds to change the importance distribution. But it is difficult to determine an uniform threshold for all images. The experimental results of different modulation functions are summarized in Tab. 4.

**Loss Function**

In the training procedure, we employ a global average pooling operation and a fully-connected layer to obtain the prediction $Y$, which represents the class probability for all categories. Finally, we leverage the multi-label soft margin loss.
where $L$ indicates the supervision of the spotlight branch. And the total classification loss can be illustrated as:

$$
L_{cls} = \frac{1}{N} \sum_{i=1}^{N} (\hat{Y}_i \log(\frac{1}{1 + e^{-Y_i}}) + (1-\hat{Y}_i) \log(\frac{e^{-Y_i}}{1 + e^{-Y_i}})),
$$

where $N$ denotes the number of classes and $\hat{Y}_i$ denote the label of the category $i$. We provide two classification losses to supervise two classification heads in the AMR. The $L_{cls}$ is supervised for the compensation branch. In short, the total classification loss can be illustrated as:

$$
L_{cls} = \frac{1}{2} (L_{cls}^s + L_{cls}^t).
$$

To make full use of complementary CAMs from the counterpart branch, we employ a cross pseudo supervision on the spotlight CAMs and the compensation CAMs. It can be viewed as a semantic similar regularization for each branch:

$$
L_{cps} = \| M_S - M_C \|_1 ,
$$

where $L_{cps}$ not only regularizes the compensation branch but also pulls the discriminative regions and easily ignored regions close to each other. Therefore, we can obtain two complementary regions as seeds to recalibrate the initial CAMs. To sum up, the proposed AMR is optimized with the final loss function $L_{all}$:

$$
L_{all} = L_{cls} + L_{cps}.
$$

**Experiment**

**Datasets and Evaluation Metric**

We evaluate our approach on the PASCAL VOC2012 dataset (Everingham et al. 2015). It contains 20 foreground objects classes and one background class. Following the common methods (Wei et al. 2017; Wang et al. 2020b), we use 10,582 images for training, 1,449 images for validation, and 1,456 ones for testing. During the whole training process, we only adopt the image-level class labels for supervision. Each image may contain multi-class labels. To evaluate the performance of experiments, we calculate the mean intersection over union (mIoU) of all classes.

**Implementation Details**

We employ ResNet50 (He et al. 2016) as the backbone of AMR. We train the network for 8 epochs with a batch size of 16. The initial learning rate is set to 0.01 with a momentum of 0.9. We leverage the stochastic gradient descent algorithm for network optimization with a 0.0001 weight decay. We also take some typical data augmentations on the training images such as random scaling and horizontal flipping. Following the works (Ahn and Kwak 2018; Ahn, Cho, and Kwak 2019), we exploit the random walk algorithm on the obtained CAMs to refine the pseudo labels. After obtained the final pseudo labels for segmentation, we train the DeepLab-v2 (Chen et al. 2017) with the backbone of ResNet101 (He et al. 2016), which is pre-trained on the ImageNet (Russakovsky et al. 2015).

**Comparison with State-of-the-art Methods**

**Comparison on semantic segmentation task.** We conduct the experiments on the DeepLab v2 (Chen et al. 2017) with the obtained pseudo labels of the training set. We report the results on the PASCAL VOC2012 validation and test set, which are shown in Tab. 1. On the one hand, AMR significantly outperforms the image-level weakly supervised method and establishes a new state-of-the-art performance. AMR achieves 68.8% of mIoU on the validation set and 69.1% on the test set, which outperforms the CONTA (Zhang et al. 2020b) with 1.4% respectively. On the other hand, AMR even achieves better or comparable results than some algorithms with more granular supervision cues. For instance, AMR surpasses the CIAN (Fan et al. 2020a) with 0.5% on validation and 0.6% on the test set, which uses the extra saliency supervision. This is an inspiring result as it reveals that our method can get impressive results via learning from massive and cheap annotations, which is of great benefit to practical application.

**Comparison on CAM and pseudo labels.** The proposed scheme aims to provide segmentation-specific CAMs to improve the quality of the pseudo labels. In order to verify the effectiveness of our method in generating CAMs and pseudo labels, we summarize the results of the CAMs and the pseudo-labels of the PASCAL VOC2012 training set with several competitive methods (see Tab. 2). It reveals that the AMR achieves the mIoU of 56.8% and 69.7% in terms of CAM and pseudo labels, respectively. Our method surpasses the advanced method SEAM (Wang et al. 2020b) by 1.4% in CAM and outperforms the CONTA (Zhang et al. 2020b) by 1.8% in pseudo labels. Note that SEAM (Wang et al. 2020b) uses Wide ResNet58 (Wu, Shen, and Van Den Hengel 2019) as the backbone, which achieves superior perfor-

<table>
<thead>
<tr>
<th>Methods</th>
<th>Sup.</th>
<th>Val</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>AffinityNet (Ahn and Kwak 2018)</td>
<td>I</td>
<td>61.7</td>
<td>63.7</td>
</tr>
<tr>
<td>IRNet (Ahn, Cho, and Kwak 2019)</td>
<td>I</td>
<td>63.5</td>
<td>64.8</td>
</tr>
<tr>
<td>CLAN (Fan et al. 2020b)</td>
<td>I</td>
<td>64.3</td>
<td>65.3</td>
</tr>
<tr>
<td>SSDD (Shimoda and Yanai 2019)</td>
<td>I</td>
<td>64.9</td>
<td>65.5</td>
</tr>
<tr>
<td>OAA+ (Jiang et al. 2019)</td>
<td>I</td>
<td>65.2</td>
<td>66.9</td>
</tr>
<tr>
<td>SEAM (Wang et al. 2020b)</td>
<td>I</td>
<td>65.1</td>
<td>65.9</td>
</tr>
<tr>
<td>Chang et al. (Chang et al. 2020)</td>
<td>I</td>
<td>66.1</td>
<td>66.5</td>
</tr>
<tr>
<td>Zhang et al. (Zhang et al. 2020a)</td>
<td>I</td>
<td>66.3</td>
<td>66.5</td>
</tr>
<tr>
<td>Chen et al. (Chen et al. 2020)</td>
<td>I</td>
<td>65.7</td>
<td>66.7</td>
</tr>
<tr>
<td>CONTA (Zhang et al. 2020b)</td>
<td>I</td>
<td>66.1</td>
<td>66.7</td>
</tr>
<tr>
<td>DRS (Kim, Han, and Kim 2021)</td>
<td>I</td>
<td>66.8</td>
<td>67.4</td>
</tr>
<tr>
<td>AdvCAM (Lee, Kim, and Yoon 2021)</td>
<td>I</td>
<td>68.1</td>
<td>68.0</td>
</tr>
<tr>
<td>MCOF (Wang et al. 2018b)</td>
<td>I + S</td>
<td>60.3</td>
<td>61.2</td>
</tr>
<tr>
<td>SeeNet (Hou et al. 2018)</td>
<td>I + S</td>
<td>63.1</td>
<td>62.8</td>
</tr>
<tr>
<td>DSRG (Huang et al. 2018)</td>
<td>I + S</td>
<td>61.4</td>
<td>63.2</td>
</tr>
<tr>
<td>FickleNet (Lee et al. 2019)</td>
<td>I + S</td>
<td>64.9</td>
<td>65.3</td>
</tr>
<tr>
<td>MCIS (Sun et al. 2020)</td>
<td>I + S</td>
<td>66.2</td>
<td>66.9</td>
</tr>
<tr>
<td>ICD (Fan et al. 2020a)</td>
<td>I + S</td>
<td>67.8</td>
<td>68.0</td>
</tr>
<tr>
<td>Yao et al. (Yao et al. 2021)</td>
<td>I + S</td>
<td>68.3</td>
<td>68.5</td>
</tr>
<tr>
<td>AMR (Ours)</td>
<td>I</td>
<td>68.8</td>
<td>69.1</td>
</tr>
</tbody>
</table>

Table 1: Comparison with the state-of-the-art methods on PASCAL VOC2012 val and test set. All results are evaluated in mIoU(%). I represents the image-level label and S indicates the saliency label.

We conduct the experiments on the DeepLab v2 (Chen et al. 2017) with the obtained pseudo labels of the training set. We report the results on the PASCAL VOC2012 validation and test set, which are shown in Tab. 1. On the one hand, AMR significantly outperforms the image-level weakly supervised method and establishes a new state-of-the-art performance. AMR achieves 68.8% of mIoU on the validation set and 69.1% on the test set, which outperforms the CONTA (Zhang et al. 2020b) with 1.4% respectively. On the other hand, AMR even achieves better or comparable results than some algorithms with more granular supervision cues. For instance, AMR surpasses the CIAN (Fan et al. 2020a) with 0.5% on validation and 0.6% on the test set, which uses the extra saliency supervision. This is an inspiring result as it reveals that our method can get impressive results via learning from massive and cheap annotations, which is of great benefit to practical application.

**Comparison on CAM and pseudo labels.** The proposed scheme aims to provide segmentation-specific CAMs to improve the quality of the pseudo labels. In order to verify the effectiveness of our method in generating CAMs and pseudo labels, we summarize the results of the CAMs and the pseudo-labels of the PASCAL VOC2012 training set with several competitive methods (see Tab. 2). It reveals that the AMR achieves the mIoU of 56.8% and 69.7% in terms of CAM and pseudo labels, respectively. Our method surpasses the advanced method SEAM (Wang et al. 2020b) by 1.4% in CAM and outperforms the CONTA (Zhang et al. 2020b) by 1.8% in pseudo labels. Note that SEAM (Wang et al. 2020b) uses Wide ResNet58 (Wu, Shen, and Van Den Hengel 2019) as the backbone, which achieves superior perfor-
Figure 4: Visualization of the generated CAMs by our method on the VOC2012 train set. (a) Input images. (b) The spotlight CAMs generated by the spotlight branch. (c) The compensation CAMs generated by the compensation branch. (d) The weighted CAMs incorporated by two complementary CAMs.

<table>
<thead>
<tr>
<th>Methods</th>
<th>CAM</th>
<th>Pseudo</th>
</tr>
</thead>
<tbody>
<tr>
<td>AffinityNet (Ahn and Kwak 2018)</td>
<td>48.0</td>
<td>59.7</td>
</tr>
<tr>
<td>CONTA (Zhang et al. 2020b)</td>
<td>48.8</td>
<td>67.9</td>
</tr>
<tr>
<td>SEAM (Wang et al. 2020b)</td>
<td>55.4</td>
<td>63.6</td>
</tr>
<tr>
<td>Chang et al. (Chang et al. 2020)</td>
<td>50.9</td>
<td>63.4</td>
</tr>
<tr>
<td>AMR (Ours)</td>
<td>56.8</td>
<td>69.7</td>
</tr>
</tbody>
</table>

Table 2: Quality results (mIoU) of pseudo labels on the VOC2012 train images. The “CAM” column indicates the initial CAM seeds generated by the classification network. The “Pseudo” represents the refined pseudo labels used to supervise segmentation.

<table>
<thead>
<tr>
<th>Baseline</th>
<th>AMM_c</th>
<th>AMM_s</th>
<th>L_cps</th>
<th>mIoU(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>48.3</td>
</tr>
<tr>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>52.9</td>
</tr>
<tr>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>53.5</td>
</tr>
<tr>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>54.9</td>
</tr>
<tr>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>56.8</td>
</tr>
</tbody>
</table>

Table 3: Comparison with different effects of each component of our method. The “Baseline” represents a single classification network. The “AMM_c,” and “AMM_s” denote the proposed channel AMM and spatial AMM respectively. L_cps denotes the semantic regularization.

AMR focuses on the discriminative regions. ii) The compensation CAMs highlight the regions that are essential for targets but easily ignored. It due to the fact that AMM helps to modulates the activation maps to emphasize the minor features. iii) The weighted CAMs contain more complete regions than spotlight CAMs, which is consistent with the essence of the semantic segmentation task.

Ablation Studies

Effectiveness of core components. To verify the effectiveness of core components in our approach, we increase each essential component gradually on the basis of the single classification network (abbreviated as “baseline”) that only contains the spotlight branch. We compare the performance of different components with the variant “baseline” in Tab. 3. As shown in Tab. 3, AMM_c and AMM_s improve the mIoU of CAMs to 52.9% and 53.5% respectively. And the whole AMM achieves 54.9%. Furthermore, the cross pseudo supervision L_cps contributes to achieving 1.9% performance improvement. The whole framework achieves the best performance 56.8%. These ablation experiments demonstrate the effectiveness of each core component in our method.

Effectiveness of modulation functions. In Tab. 4, we compare the results of different modulation functions introduced in Fig. 3. “Threshold” modulates the activation to 1 when exceeding the threshold and sets to 0 when the activation is lower than the threshold, which can obtain 1.8% improvement on the baseline as it remains the most important feature and strengthens some minor activations. The
Figure 5: Qualitative results on the PASCAL VOC2012 validation set. (a) Input images. (b) Ground truth labels. (c) The segmentation results by IRNet (Ahn, Cho, and Kwak 2019). (d) The segmentation results of our approach.

Table 5: Comparison with different recalibration coefficient. mIoU is evaluated on the CAMs of VOC2012 train images.

<table>
<thead>
<tr>
<th>ξ</th>
<th>0.1</th>
<th>0.3</th>
<th>0.5</th>
<th>0.7</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>mIoU (%)</td>
<td>49.2</td>
<td>53.4</td>
<td>56.8</td>
<td>54.5</td>
<td>50.7</td>
</tr>
</tbody>
</table>

Table 6: Generalization results of AMR on IRNet (Ahn, Cho, and Kwak 2019) and SEAM (Wang et al. 2020b).

<table>
<thead>
<tr>
<th>Methods</th>
<th>CAM(mIoU)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRNet+Ours</td>
<td>56.8</td>
</tr>
<tr>
<td>SEAM (Wang et al. 2020b)</td>
<td>55.4</td>
</tr>
<tr>
<td>SEAM+Ours</td>
<td>57.9</td>
</tr>
</tbody>
</table>

Visualize of Segmentation Results

As illustrated in Fig. 5, we compare our method with IRNet (Ahn, Cho, and Kwak 2019) on the segmentation results in the validation set of PASCAL VOC2012 (Everingham et al. 2015). As we can see, the results of IRNet (Ahn, Cho, and Kwak 2019) often fall into misjudgment in some ambiguous regions. On the contrary, our approach successfully dig out more regions belonging to the target objects to achieve superior segmentation performance.

Conclusion

In this paper, we propose a novel activation modulation and recalibration (AMR) scheme for WSSS, which leverages a spotlight branch and a plug-and-play compensation branch to obtain weighted CAMs and provide more semantic segmentation-oriented concepts. An AMM module is designed to rearrange the distribution of feature importance from the channel-spatial sequential perspective, which contributes to highlighting some essential regions for segmentation tasks but are easy to be ignored. Extensive experiments on PASCAL VOC2012 dataset demonstrate that AMR achieves the new state-of-the-art performance of weakly supervised semantic segmentation.
References
Bearman, A.; Russakovsky, O.; Ferrari, V.; and Fei-Fei, L. 2016. What’s the point: Semantic segmentation with point supervision. In European conference on computer vision, 549–565. Springer.


